INFLUENZA A *VIRUS* (H3N2) GENOMIC SEQUENCE DIFFERENCE MEASURES BASED on WORD ABSENCE and EXPRESSION LEVELS

Adrienne Breland[1], Sara Nasser[1], Karen Schlauch[2], Frederick C. Harris Jr.[1]

[1]Department of Computer Science and Engineering
University of Nevada, Reno
Reno, Nevada 89512

[2]Department of Biochemistry and Molecular Biology
University of Nevada, Reno
Reno, Nevada 89512

Abstract

In a genomic sequence, the oligonucleotide signature represents the ratio of the observed to expected number of occurrences of all possible nucleotide words of a specific length. Word absence is also found in genomic sequences whereby specific words are not observed despite their expected presence in a random nucleotide distribution. This research uses a combination of word absence and oligonucleotide signatures to quantify differences between inter-epidemic and intra-epidemic *Influenza* A *virus* (H3N2) genomic sequences. In addition, word absence/presence patterns are examined for their discrimination of sequences from distinct epidemics or geographic origins within the same epidemic. Inter-epidemic sequences are well delineated by word absences and difference measures into epidemic specific groups. Intra-epidemic sequences are not consistently well separable in terms of their geographic origins, but show similarities across geographic regions.

Keywords: Oligonucleotides, word absence, alignment-free, nullomers, k-mers, Minimal Markov Model.

1. INTRODUCTION

Finding patterns in genomic sequences lends to understanding their internal mechanisms for conservation of order and functionality. Processes such as replication, gene expression level, gene coding and defense against invasive DNA are driven by embedded sequences which contribute to nucleotide patterns throughout a genome [7]. A genome describes the complete DNA or RNA sequence which "encodes" an organism. Genomic patterns refer to patterns in the nucleotide ((A)denine, (C)ytosine, (G)uanine, (T)yrosine) makeup of whole genomes or subset sequences. These enable the comparison of sequences from different classes of organisms. These classes can be derived to compare phylogeny, subspecies strains, or even subspecies characteristics, i.e. virulence. A genomic signature derives frequency patterns by calculating the over- and under- representation of specific base pair sequences (words) when compared to random expectations. If genomic sequences were randomly organized, most short nucleotide words would have an equal probability of being found within any given sequence of sufficient length. The study of short word frequencies has shown a biased distribution of words which deviates from random, leaving some words over- and some under- represented to differing levels [5, 6, 10]. Genomic signatures of short word lengths are similar for organisms within kingdom groupings [6] and are sometimes consistent enough to be used in the regrouping of mixed fragments from multiple species genomes [1, 13, 15].

Some words have been found to be commonly absent from species groups [9] and have been referred to as nullomers and primes. While the reason as to why certain words are absent and others present in particular genomes is most likely complex, the inheritance of absent words has been examined on a broad evolutionary scale. It has been proposed that word absence is an inherited characteristic through the observation that human and chimp DNA contain 28 common absent words of length 11 and 14 absent words which differ by only one nucleotide [2]. It could also be expected that word absence is inherited by the immediate progeny of microbial samples in a micro-evolutionary sense. Absent words are an integral part of any genomic sequence as much as present words, and by inheriting a nucleotide sequence, or a close derivative of it, a microbial offspring should also inherit many of the words absent from that sequence as well. This may offer the delineation of closely related microbes, including viral pathogens. Researchers in [8] found word absence/presence to show more correlation between genomes within the same species than between genomes of different species. Even so, less correlation was found between same species genomes than was statistically expected, and it was suggested that word absences may offer delineation within species groups as well. How detailed the distinction of subspecies groups through word absence can be has yet to be examined.

While genomic word signatures show consistencies within higher level phylogenetic groups, examining differences at subspecies levels may reveal differences that could enable pathogen transmission mapping and give insight into strain evolution. In addition, efficient methods for strain identification and differentiation are becoming increasingly important in the threat of epidemic outbreaks and the possibility of biothreat agents [4, 16]. Most current methods for genotypic subspecies comparisons focus on computationally expensive sequence alignment, although global pattern based methods have reported. In [3], Average Mutual Information (AMI) profiles enabled the clustering of subtypes of the HIV-1 virus. In [12] a triplet Markov model was used to derive a phylogenetic tree which grouped subspecies samples of human pathogens *Escherichia coli*, *Staphylococcus aureous*, and *Yersinia pestis* into appropriate branches. Oligonucleotide word signatures present another global method for comparing whole genomes without the need for sequence alignment. In addition, it is possible to select aspects of word signature sets with the intent of exposing subtle subspecies differences which may otherwise be masked by a high degree of sequence similarity. Word absence offers such a selection criteria through which the expression level of any word present in one genome but absent in another may be considered in the calculation of a sequence difference measure. This research examines the potential use of word absence in conjunction with oligonucleotide signatures for determining *Influenza* A *virus* (H3N2) isolate relatedness. Section 2.2 describes the derivation of word over- and under- representation using Minimal Markov Models. In section 3.1, word absences are first examined for pattern similarities across inter- and intra-epidemic sequences. In section 3.2, a sequence difference measure is described and tested on both groups of sequences to quantify relationships among them.

2. METHODOLOGY

2.1 Data

All sequences were acquired through the publicly available National Institute of Allergies and Infectious Disease (NIAID) Influenza resource [2].

Influenza virus genomes naturally exist in eight disjoint segments. Thus data from each segment (1-8) was kept disjoint during processing and sequences were not concatenated. In addition, Influenza viruses are single stranded, and so reverse complementation of data sets was not required.

Inter-epidemic data included eight strains of *Influenza* A *virus* (H3N2) representing three distinct epidemics. Two strains were from Hong Kong in 1980, three strains from Managua, Nicaragua in 2007, and three from New South Wales in 1999. Identifier strings are listed:

A/Hong Kong/46/1980(H3N2),
A/Hong Kong/45/1980(H3N2),
A/Managua/14/2007(H3N2),
A/Managua/15/2007 (H3N2),
A/Managua/16/2007 (H3N2),
A/New South Wales/20/1999(H3N2),
A/New South Wales/21/1999(H3N2),
A/New South Wales/22/1999 (H3N2).

Intra-epidemic data included nine *Influenza* A *virus* (H3N2) isolates collected in the United States within a three month period during the 2007 flu season. Three are from New York collected between 3/5-3/6. Three isolates were collected in Colorado all on 1/8, and three are from Vermont collected between 2/27-3/1. Identifier strings are listed:

A/New York/UR06-0510/2007(H3N2),
A/New York/UR06-0515/2007(H3N2),
A/New York/UR06-0529/2007(H3N2),
A/Vermont/UR06-0469/2007(H3N2),
A/Vermont/UR06-0470/2007(H3N2),
A/Vermont/UR06-0471/2007(H3N2),
A/Colorado/UR06-022/2007(H3N2),
A/Colorado/UR06-023/2007(H3N2),
A/Colorado/UR06-024/2007(H3N2).

2.2 Markov Chain Selection

In genomic word expression analysis, Markov Models are often used as a means of calculating the expected count of each word (E(w)) in a signature set [11, 14, 17]. In Markov chains, the current state of a system is predicted by its previous states. In word signature analysis, this translates to predicting a word frequency based on the observed frequencies of its sub words. Depending on the order of the Markov model, bias contributed to a word of length *m* from sub words of lengths 1,…,*m-1* can be removed. For example, assume a sequence is dominated by di-nucleotides TA and AG. Unless specifically selected against, TAA and AAG, which are the concatenations of the dominant sub words will naturally show high frequencies. Separating the degree of selection for or against exactly these tri-nucleotides from the contributions of their sub words may require the removal of their sub word frequency bias. With the ultimate goal to match genomic internal word selection mechanisms, the optimal degree of Markov Model to use remains undetermined. Consistent with findings in [14] and [18], minimal order Markov Models allowed the most differentiation between genomic signatures of different prokaryotic species and were thus used to calculate expected values for signature calculations in this research.

A minimal order Markov model does not remove bias from sub words longer than one character. The expected count of a word E(w), in a genomic sequence of length N is expressed as:

$$\mathbf{E(w) = [(A^a * C^c * G^g * T^t) * N]}$$

A, C, T and **G** represent specific nucleotide frequencies in the total sequence **N** and **a, c, t, g** are the number of each nucleotide in a word **w**. As described in [14], the ratio of the observed count over its expected count, O(w)/E(w) was then used to derive the degree of over- or under- representation of each word in a signature set.

2.3 Oligonucleotide Word Length Selection

The *k*-nucleotide signature of a genomic sequence contains the degree of over- or under- representation of all 4^k possible k-length nucleotide words composed of a subset of {A,C,G,T}. To ensure that all words in a signature set can be contained in a sequence of length L, it is required that *k* be small enough so that $4^k <= L-(k+1)$. It should be noted that in previous word absence studies [8, 9] this upper bound on the length of *k* was not deemed necessary and longer lengths of *k* were examined. In the case of word expression in genomic sequences which do not adhere to random expectations, the examination of word absences at longer *k* values offers valuable information. For this study, word absences were noted in relatively high amounts at a word length of six, and so the upper bound proportion is adhered to. Influenza sequences are ~12,000 base pairs (bp) in length while a hexa-nucleotide word set contains 4096 possible nucleotide words. Because nucleotide words were read in an overlapping fashion (Figure 1), each hexa-nucleotide word had the probability of occurring approximately 2.9x under the assumption of a random nucleotide distribution.

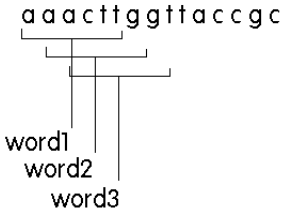

Figure 1. Overlapping words.

3. RESULTS and DISCUSSION

3.1 Absence/Presence Word Patterns

Word absence vs. presence was examined for the potential to discriminate between related groups of sequences. Inter-epidemic sequences had the potential to be differentiated based on the epidemic outbreak instance from which they were collected; Hong Kong 1980, Nicaragua 2007 or New South Wales 1999. Intra-epidemic sequences could potentially be differentiated based on the specific state from which they were obtained; New York, Colorado or Vermont.

Signature sets for each genome were reclassified in a binary format so that presence was indicated by a one and absence by a zero. Within inter-epidemic sequences, 379 words were commonly absent from all sequences, 3173 were present in all sequences, and 544 word exhibited absence/presence variation across samples. Intra-epidemic sequences had 507 commonly absent words, 3299 commonly present words, and 290 words exhibiting absence/presence variation across samples. Inter-epidemic sequences had more words exhibiting absence/presence variation and less commonly absent from all sequences. In contrast, intra-epidemic sequences had a higher proportion of commonly absent words than varied words. This is to be expected under the assumption that intra-epidemic sequences will show more similarity. In both data groups (inter- and intra-epidemic), groups of words exhibiting identical absence/presence patterns across sequences were found. Tables 1 and 2 show absent vs. present word pattern clusters. Commonly absent words and commonly present words are not included.

Table 1 shows a visible distinction of samples from the three epidemics through absence and presence. The first word cluster contains 123 words that are distinctly absent from the Hong Kong 1980 epidemic sequences (s1, s2) while being present in all other epidemic sequences (s3-s8). Similarly, the next largest word cluster of 109 words contains those only present in the Hong Kong 1980 sequences while absent from all others. The six largest word pattern clusters in this table are uniquely absent or present for one of the three epidemics, and all three epidemics are delineated through absence and present words. For this group, such epidemic discriminating words account for 82% of all varied words.

Discriminating word patterns with respect to location are not as evident in Table 2 for the intra-epidemic data set. The two largest word clusters show distinct absence or presence for one sequence from the New York group (s2). The third largest word cluster shows distinct absence for two sequences from New York (s1,s2) and one from Colorado (s8).

Interpreting results in Table 2 are not intuitive except that they may offer insight into intra-epidemic isolate relatedness Similarly, the remaining 18% or words in Table 1 which do not discriminate clearly between epidemics may also reflect relationships across epidemics. While results in Table 2 are not clear due to lack of knowledge regarding specific samples, Table 1 strongly supports the notion that absence/presence can delineate sample relatedness in a meaningful manner.

Table 1. Word Presence/Absence patterns and word clusters from inter-epidemic samples. Dotted line indicates epidemic discriminating word groups.

Inter-epidemic								
Hong Kong 1980		Nicaragua 2007			New South Wales 1999			
s1	s2	s3	s4	s5	s6	s7	s8	# words
0	0	1	1	1	1	1	1	123
1	1	0	0	0	0	0	0	109
1	1	0	0	0	1	1	1	93
0	0	1	1	1	0	0	0	71
1	1	1	1	1	0	0	0	32
0	0	0	0	0	1	1	1	19
1	1	1	1	1	0	0	1	14
0	0	0	0	0	0	0	1	13
0	0	0	0	0	1	1	0	11
0	0	1	1	1	0	0	1	11
0	0	1	1	1	1	1	0	10
1	1	0	0	0	1	1	0	9
0	0	0	1	0	0	0	0	5
1	1	0	0	0	0	0	1	5
1	1	0	0	1	1	1	1	3
0	0	0	0	1	0	0	0	2
0	0	1	1	0	1	1	1	2
1	1	1	0	1	1	1	1	2
1	1	1	1	0	1	1	1	2
0	1	0	0	0	0	0	0	1
0	1	1	1	1	0	0	0	1
1	0	1	1	1	0	0	0	1
1	1	0	0	1	0	0	0	1
1	1	0	1	0	0	0	0	1
1	1	1	0	0	1	1	1	1
1	1	1	1	0	0	0	0	1
1	1	1	1	1	1	0	0	1
						total		544

3.2 Sequence Difference Measure

A measure for comparing two sequences based on word expression levels constrained by word absence was derived. To compare two sequences, s1 and s2, let AP be the set of all words present in only one sequence so that for all $w \in AP$, $O(w_{s1})/E(w_{s1})>0$ and $O(w_{s2}) = 0$, or $O(w_{s2})/E(w_{s2})>0$ and $O(w_{s1}) = 0$. AA is the set of all words absent from both sequences so that for all $w \in AA$, $O(w_{s1}) = O(w_{s2}) = 0$. $|AP|$ denotes the total number of words in AP and $|AA|$ denotes the total number of words in AA. The difference between s1 and s2 is calculated as:

$$\frac{\sum_{w}^{AP} | O(w_{s1})/E(w_{s1}) - O(w_{s2})/E(w_{s2}) |}{|AP| + |AA|}$$

Table 2. Word Presence/Absence patterns and word clusters from intra-epidemic samples.

Intra-epidemic									
New York 2007			Vermont 2007			Colorado 2007			
s1	s2	s3	s4	s5	s6	s7	s8	s9	# words
1	0	1	1	1	1	1	1	1	54
0	1	0	0	0	0	0	0	0	49
1	1	0	0	0	0	0	1	0	22
0	1	1	1	1	1	1	1	1	22
0	1	1	1	1	1	1	0	1	22
1	1	1	1	1	1	1	0	1	15
0	0	1	1	1	1	1	0	1	15
1	0	0	0	0	0	0	0	0	14
1	0	0	0	0	0	0	1	0	10
1	1	1	0	0	1	1	1	1	6
0	0	0	1	1	0	0	0	0	6
1	1	1	1	1	1	0	1	0	5
1	1	1	1	1	0	1	1	1	5
1	1	0	0	0	0	0	0	0	5
0	0	0	0	0	1	0	0	0	5
1	1	0	1	1	1	1	1	1	4
1	0	1	1	1	1	1	0	1	3
1	0	1	1	0	1	1	1	1	3
1	0	0	1	1	0	0	1	0	3
0	0	1	0	0	0	0	0	0	3
1	1	1	1	0	1	1	1	1	2
1	0	0	1	1	0	0	0	0	2
0	0	1	0	0	1	0	0	0	2
0	0	1	0	0	0	0	1	0	2
0	0	0	0	0	0	1	0	1	2
1	1	1	1	0	0	1	0	1	1
1	1	0	1	1	0	0	1	0	1
1	1	0	0	0	1	0	1	0	1
1	0	1	1	1	1	0	1	0	1
1	0	1	0	0	1	1	1	1	1
0	1	0	1	1	0	0	0	0	1
0	1	0	0	0	1	0	0	0	1
0	1	0	0	0	0	0	1	0	1
0	0	0	1	0	0	0	0	0	1
						total			290

Thus the difference between two sequences is the sum of the observed to expected ratios for words which are absent from *exactly* one sequence divided by the total number of words absent from *at least* one sequence. This allows the comparison of only words which exhibit some degree of absence.

In contrast to comparing only two sequences, if comparing *relative* similarities between a group of sequences, removing $|AA|$ from the equation allows a higher degree of distinction between all pairs. This is because words absent from all sequences offer no inter-

sequence differentiation and do not contribute to the derivation of relative differences. This adjusted measure was used to derive difference matrices for the inter- and intra-epidemic sequences (Tables 3,4).

Table 3. Inter-epidemic difference matrix (full precision is not shown).

Inter-epidemic								
	Hong Kong 1980		Nicaragua 2007			New South Wales 1999		
	s1	s2	s3	s4	s5	s6	s7	s8
s1	0.00	0.01	0.51	0.52	0.51	0.43	0.43	0.44
s2	0.01	0.00	0.51	0.52	0.51	0.42	0.43	0.44
s3	0.51	0.51	0.00	0.01	0.02	0.34	0.35	0.34
s4	0.52	0.52	0.01	0.00	0.03	0.35	0.35	0.34
s5	0.51	0.51	0.02	0.03	0.00	0.35	0.35	0.34
s6	0.43	0.42	0.34	0.35	0.35	0.00	0.00	0.13
s7	0.43	0.43	0.35	0.35	0.35	0.00	0.00	0.13
s8	0.44	0.44	0.34	0.34	0.34	0.13	0.13	0.00

Table 4. Intra-epidemic difference matrix (full precision not shown).

	Intra-epidemic								
	New York 2007			Vermont 2007			Colorado 2007		
	s1	s2	s3	s4	s5	s6	s7	s8	s9
s1	0.00	0.42	0.31	0.30	0.30	0.31	0.31	0.22	0.31
s2	0.42	0.00	0.38	0.37	0.37	0.38	0.36	0.38	0.36
s3	0.31	0.38	0.00	0.10	0.11	0.06	0.07	0.27	0.07
s4	0.30	0.37	0.10	0.00	0.02	0.10	0.09	0.26	0.09
s5	0.30	0.37	0.11	0.02	0.00	0.11	0.10	0.27	0.10
s6	0.31	0.38	0.06	0.10	0.11	0.00	0.07	0.27	0.07
s7	0.31	0.36	0.07	0.09	0.10	0.07	0.00	0.26	0.00
s8	0.22	0.38	0.27	0.26	0.27	0.27	0.26	0.00	0.26
s9	0.31	0.36	0.07	0.09	0.10	0.07	0.00	0.26	0.00

Table 3 shows all sequences having minimal difference measures with sequences within their respective epidemic groups. These values are highlighted in yellow. For example, s1 is least different from s2, and these two sequences are both members of the Hong Kong 1980 epidemic. This table suggests that samples from distantly related epidemics can be accurately delineated using the proposed measure. The average difference between samples within epidemics is 0.037 while the average distance between samples from different epidemics is 0.42, an order of magnitude larger. It is also notable that while s8 from New South Wales is closest to a sequence from the same epidemic, it appears relatively distant within that epidemic. All other difference measures between same epidemic sequences are less than 0.03, while s8 differs from other NSW sequences by at least 0.13.

Table 4 mirrors Table 2 in that samples are not unanimously discriminated based on their geographic location. In this case, only four out of nine samples s4, s5, s7 and s9 show the lowest difference values from same state samples. In addition, s4 and s5 from Vermont are closest to each other, while s7 and s9 from Colorado are as well. These indicate two strongly related isolates found in close geographic proximity. All sequences from New York are closest to Colorado and Vermont sequences than to those collected within the same state. One sample from Colorado, s8 show a minimal difference from a New York sample. Similarly, while two samples from Vermont point to each other, s6 is closest to a sample from Colorado. The average difference between sequences from the same state is 0.21 while the average distance between different state sequences is almost identically valued at 0.22.

Tables 3 and 4 show the differentiability of distantly related epidemic samples, as well as the inter relatedness of same epidemic samples. It is to be expected that intra-epidemic sequences, particularly within a well traveled country such the United States, be highly related. Similarly, it is not surprising that sequences from geographically and temporally distant epidemics show more differences. The clear distinction between sequences from distant epidemics is slightly more surprising and encouraging. In addition, the observation that some sequence pairs are highly similar, i.e. New South Wales s6 and s7, and Colorado s8 and s9, with difference measures of less than 0.001, suggest the ability of this method to detect very closely related isolates given an adequate data set.

4. CONCLUSION

A comparison method for closely related sequences of the *Influenza* A *virus* (H3N2) subtype was proposed. Hexa-nucleotide word signatures using Minimal Markov Models were derived for inter- and intra-epidemic sequences. These signature sets contained representation values of 4096 words per sequence. Out of these word sets, only values for words which exhibited some degree of absence were used in a proposed difference measure. This was in attempts to isolate and utilize significant lineage differences. This measure was successful in delineating samples from distinct epidemics while showing more complex relationships across samples from different U.S. states during the same epidemic year, 2007.

Although difference measures have only been derived for a small number of samples, they are suggestive a highly detailed and quantifiable network among Influenza viral isolates. Furthermore, the described method does not rely on sequence alignment which is computationally expensive. Instead, the differences between closely related genomes are extracted in a

relatively inexpensive manner. This could prove useful for epidemiological studies, particularly with regards to understanding global transmission networks involving large numbers of sequences.

The accuracy of the proposed method must be tested on larger datasets and compared with existing Influenza phylogenies. In addition, cutoff values for difference measures will be examined in terms of what degree of phylogenetic relatedness they represent between genomes.

REFERENCES

[1] Abe, Takashi, Shigehiko Kanaya, Makato Kinouchi, Yuta Ichiba, Tokio Kozuki and Toshimichi Ikemura, "A Novel Bioinformatic Strategy for Unveiling Hidden Genome Signatures of Eukaryotes: Self-Organizing Map of Oligonucleotide Frequency," *Genome Informatics* , vol. 12, pp. 12-20, 2002.

[2] Bao Y., P. Bolotov, D. Dernovoy, B. Kiryutin, L. Zaslavsky, T. Tatusova, J. Ostell, and D. Lipman, "The Influenza Virus Resource at the National Center for Biotechnology Information," *J. Virol,* vol. 82, pp. 596-601, 2008.

[3] Bauer, Mark, Sheldon Shuster and Khalid Saywood, " The Average Mutual Information Profile as a Genomic Signature," BMCBioinformatics, vol. 9, pp. 48-, 2008.

[4] Beckstrom-Sternberg SM,et. al. , "Complete Genomic Characterization of Pathogenic A.II Strain of Francisella tularensis Subspecies tularensis, " PLoS ONE, vol..2, pp. e947, 2007.

[5] Burge, Chris, Allan Campbell, Samuel Karlin, "Over- and Under-Representation of Short Oligonucleotides in DNA Sequences," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 89, pp.1358-136.2, 1991.

[6] Campbell, A., J. Mrazek and S. Karlin, "Genome signature comparisons among prokaryote, plasmid and mitochondrial DNA," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, pp. 9184-9, 1999.

[7] Eduardo, P., C. Rocha, Alain Viari, and Antoine Danchin, "Oligonucleotide bias in Bacillus Subtilis: general trends and taxonomic comparisons," *Nucleic Acids Research*, vol. 26, pp. 2971-2980, 1998.

[8] Fofanov, Yuriy et al.., "How independent are the appearances of n-*mer*s in different genomes?," *Bioinformatics* , vol. 20, pp. 2421-2428, 2004.

[9] Hampikian, Greg, Tim Anderson, "Absent Sequences: Nullomers and Primes," Pacific Symposium on Biocomputing, vol. 12, pp 355-366, 2007.

[10] Karlin S, Burge C. ,"Dinucleotide relative abundance extremes:a genomic signature," *Trends in Genetics,* vol. 11, pp. 283-90, 1995.

[11] Leung M. –Y, Marsh, G. M., and Speed, T. P, "Under- and overrepresentation of short DNA words in Herpesvirus Genomes", *Journal of Computational Biology*, vol. 3, pp. 345-360, 1996.

[12] Li, Jian, Khalid Saywood, "A Genome Signature Based on Markov Modelling," Proceedings of the 2005 IEEE, Engineering in Medicine and Biology 27th Annual Conference, 2005.

[13] McHardy, Alice, Hector Martin, Aristotelis Tsirigos, Philip Hugenholtz and Isidore Rigoutsos, " Accurate phylogenetic classification of variable-length DNA fragments," *Nature methods,* vol. 4, pp. 63-72, 2007.

[14] Pride, David, Richard Meinersmann, Trudy Wassenaar and Martin Blaser, "Evolutionary Implications of Microbial Genome Tetranucleotide Frequency Biases," *Genome Research* , vol. 13, pp. 145-58, 2003.

[15] Teeling, Hanno, Anke Meyerdierks, Maragrete Bauer, Rudolf Amann and Frank Glockner, "Application of tetranucleotide frequencies for the assignment of genomic fragments," *Environmental Microbiology,* vol. 6, pp. 938-947, 2004.

[16] Pannucci J, Cai H, Pardington PE, Williams E, Okinaka RT, Kuske CR, Cary RB, "Virulence signatures: microarray-based approaches to discovery and analysis," *Biosens Bioelectron*. vol. 20, pp. 706-18, 2004.

[17] Schbath, S. , Prum, B., and Turckheim, E. DE, "Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences", *Journal of Computational Biology*, vol. 2, pp. 417-437, 1995.

[18] Rainer Merkl, Manfred Kroger, Peter Rice, and Hans-Joachim Fritz, "Statistical Evaluation and biological interpretation of non-random abundances in the E.coli K-12 genome of tetra- and pentanucleotide sequences related to VSP DNA mismatch repair," *Nucleic Acids Research.* , vol. 20, pp. 1657- 1662, 1992.