University of Nevada,
Reno

A Supervised Strain Classifier

A dissertation submitted in partial fulfillment of the requirements for the degree of Masters of Science in Computer Science and Engineering

by

Adrienne E. Breland

Dr. Frederick C. Harris, Jr./Thesis Advisor

May, 2008

University of Nevada, Reno
Statewide • Worldwide

THE GRADUATE SCHOOL

We recommend that the thesis
prepared under our supervision by

ADRIENNE E. BRELAND

entitled

A Supervised Strain Classifier

be accepted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

Frederick C. Harris, Jr., Ph.D., Advisor

Monica Nicolescu, Ph.D., Committee Member

Karen Schlauch, Ph.D., Graduate School Representative

Marsha H. Read, Ph. D., Associate Dean, Graduate School

May, 2008

Acknowledgements

During the last four years I have; had a child, raised a child, gotten my black belt and completed this Masters Degree. Thank God!

Thanks also to my awesome committee members Dr. Karen Schlauch, Dr. Monica Nicolescu and to Dr. Harris for being a truly supportive advisor. Thanks also to Nancy Latourette for being such a good teacher in my first CS classes that I wanted to continue.

Of course, thanks to my parents for helping me in so many ways.

Abstract

Several bacterial and viral species are human pathogens and contain strains exhibiting different degrees of virulence. Nucleic acid sequencing enables strain fingerprinting, which is a term used for identifying bacterial and viral strain species and subtypes based on their DNA. Strain fingerprinting methods are becoming increasingly important in the threat of epidemic outbreaks and the possibility of biothreat agents [3, 4, 17]. This thesis examines the use of oligonucleotide word signatures for strain fingerprinting and related classifications. An investigation into word signature differences exhibited by different strains of the same subtype reveals that words not expressed by individual genomes offer the most potential as differentiating features. Thus, a supervised classifier is built with feature sets derived from absent words. Resulting accuracies are high and are listed for five classifications at different levels of phylogenetic resolution: Mixed Pathogens: 100%, *Influenza A virus/Influenza B virus*: 100%, *Influenza A virus* subtypes (human host): 96%, Avian Influenza A virus H5N1 lineages: 94%, Avian to Human Transmission H5N1 lineages: 100%. While the data set used does not allow complete confirmation of reported accuracies, it is suggested that this method could be a valuable tool in comparative genomics and enable geographic origination determination of *Influenza A virus* and other pathogenic isolates.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Microbes are living species invisible to the naked eye and are ubiquitous in every living system. Several bacterial and viral species are human pathogens and contain strains exhibiting different degrees of virulence. Subsequently, many of these species have been studied extensively and their genomes sequenced. For example, a large database of *Influenza A* and *B virus* genomes has recently been made publicly available by the National Institute of Allergy and Infectious Disease (NIAID) [2]. Influenza epidemics cause an average of 30,000 deaths per year in the U.S. while flu pandemics such as the Spanish flu in 1918 can affect 20-40% of the world's population [22].

Nucleic acid sequencing enables strain fingerprinting, which is a term used for identifying bacterial and viral strain species and subtypes based on their DNA. Strain fingerprinting methods are becoming increasingly important in the threat of epidemic outbreaks and the possibility of biothreat agents [3, 15, 40]. In addition, strain fingerprinting can enable medical diagnosis, insights into microbial strain evolution, as well as geographic distribution and transmission networks [37].

The specific order of nucleotides within and adjacent to coding regions of DNA dictate which proteins will be created under any given condition. In this way, DNA codes how an organism will live in its environment. While the entire genomic sequence is analogous to the main driver in a C program, genes represent embedded functions which are called under specific circumstances and proceed to call on each other in complex patterns. Though genomes can be billions of base pairs (bp) long, they are ultimately composed of only four different nucleotides; Adenine(A), Guanine(G), Tyrosine(T), and Cytosine(C). A base pair is the term used for the bound pair of nucleotides across each DNA strand, with specific binding rules. Unless an error is present, A only binds with T and C only binds with G, and all nucleotides are paired.

As would be expected from a molecule with such a distinct and crucial biological role, the order of base pairs composing DNA is not random. This has well been elucidated by the discovery of the oligonucleotide word signature. A genomic signature is a preferential usage of specific base pair patterns (words) which remains consistent. It is derived by calculating the over- and under- representation of nucleotide words when compared to random expectations. Processes such as replication, transcription, gene coding and defense against invasive DNA are driven by nucleic acid sequences which contribute to nucleotide patterns throughout a genome [14].

This thesis examines the use of oligonucleotide signatures for strain fingerprinting and related classifications. First, an investigation into signature differences exhibited by different strains of the same subtype is described. The results suggest that word sequences simultaneously absent from groups of genomes can indicate lineage relatedness at a resolution finer than subtype. Due to these results, a supervised classifier is built with feature sets derived from absent words. This classifier is tested on its ability to identify genomes from different genera, species within the same genera, subtypes within a species, lineage groups within a subtype, and viral lineage groups across host species (avian to human).

The next sections in this thesis are structured as follows: Chapter 2 presents a background literature review of oligonucleotide word signatures and current strain fingerprinting methods. Chapter 3 describes a preliminary examination of word signature differences between strains of the same subspecies, *Staphylococcus aureous* subs. *aureous*. Chapter 4 presents a supervised classification based on results in Chapter 3, and includes a description of five classification applications. Chapter 5 shows results to all classifications, and Chapter 6 presents a discussion of the strain classifier including its relevance to current events, potential improvements and possible future applications.

Chapter 2

Background

2.1 Rooted Phylogenetic Trees

Rooted phylogenetic trees are equivalent to computational tree data structures with each parent node representing a common ancestor to all of its child nodes. The concept of classifying living organisms with a tree structure was first proposed by Charles Darwin in 1859 [59]. The specific organization of trees representing related groups of organisms depends on both the traits used to differentiate organisms as well as the tree building method. Computational phylogenetics employs computer algorithms in the determination of tree based relationships. Different nucleic acid based traits such as whole genome alignment, 16Srna [59] or the sequence of specific genes can be used to derive any number of tree forms among groups of organisms.

Current taxonomic classifications can be traced back to the grouping of organisms based on physical characteristics. These groupings have since been revised with the Darwinian principle of common descent and genetic sequence information. The root of the tree of life is “life”. Life requires all of a certain group of traits. These include a carbon and water based cellular form, complex organization, heritable genetic information, capacity for metabolism, growth, response to stimuli, reproduction and adaptation of successive generations through natural selection [59]. While bacteria are considered living organisms, viruses are not able to metabolize and are thus generally not considered life forms. Another definition relevant to this thesis is that of “species”. A general definition of species is a group or organisms capable of interbreeding and producing fertile offspring [59]. Two examples of phylogenetic trees are given. Figure 2.1 illustrates the derivation of *Homo sapiens* subsp. *sapiens* (modern man) through current taxonomic groupings. Figure 2.2 shows the path of the H5N1 Influenza A virus subtype (bird flu).

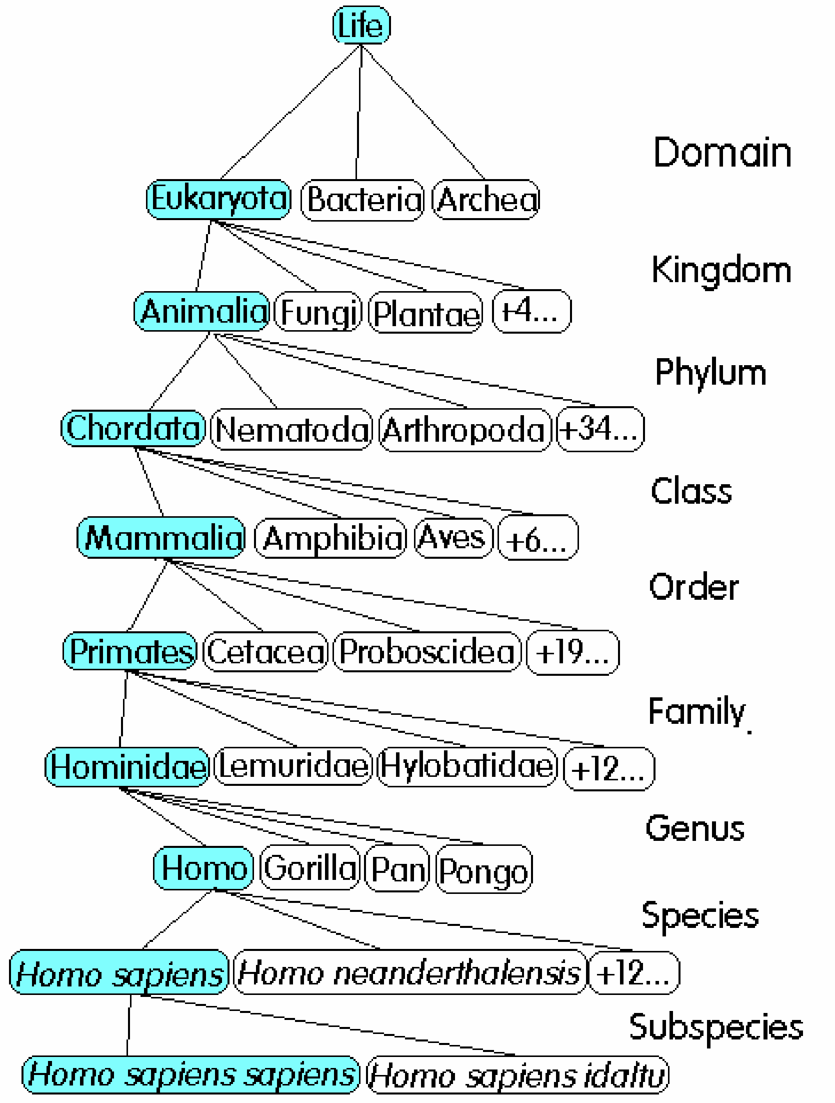

Figure 2.1. Basic phylogenetic tree, modern man trace.

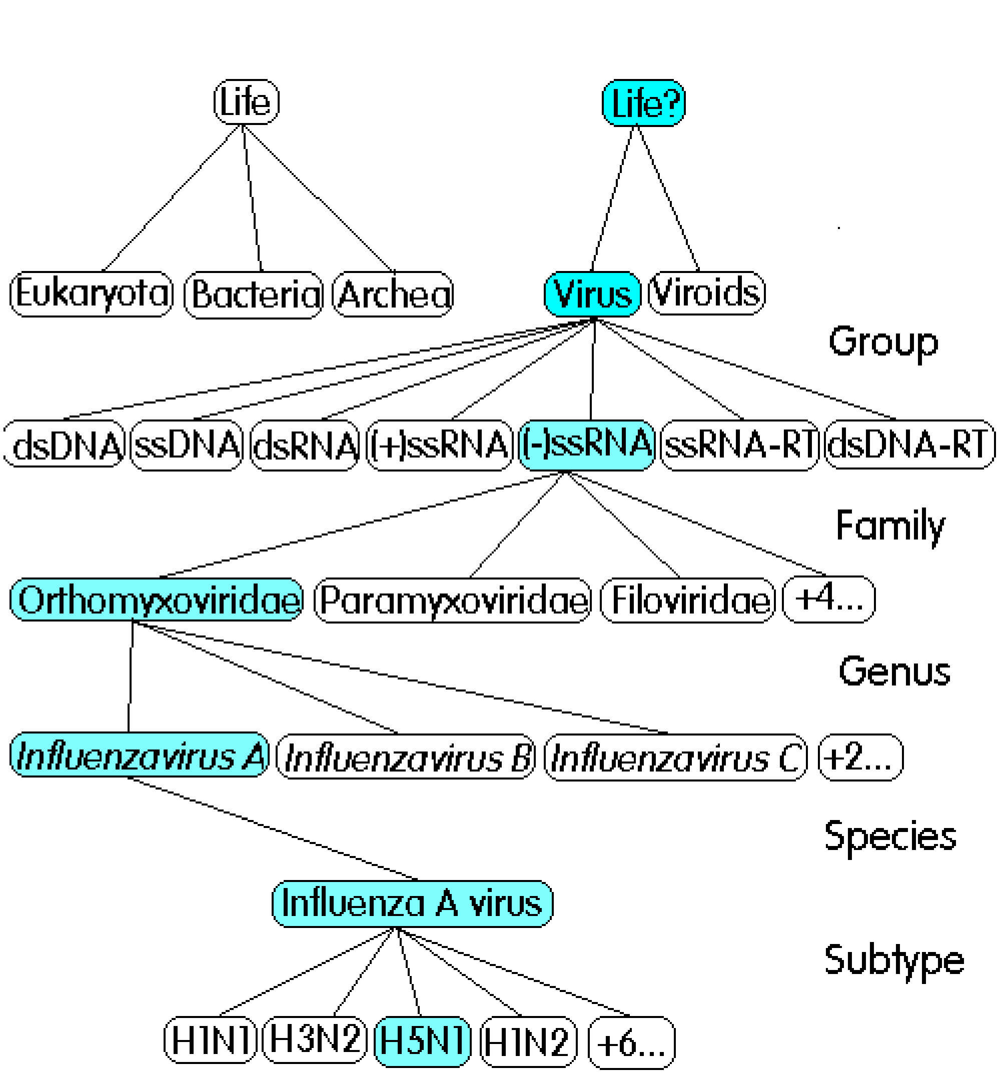

Figure 2.2. Basic phylogenetic tree, *Influenza A virus* H5N1 trace.

2.2 Overview of Human Pathogens

This section presents a brief description of each species included in the data set used for classifications. More specific descriptions of subtypes and strains used are included in the classification results. All information for this section was obtained from the pubmed website [57] and the National Institute of Allergies and Infectious Disease (NIAID) Influenza resource [2].

Bacillis Anthracis**:** This is a bacterial species that causes the human disease known as anthrax. This disease exhibits three forms including cutaneous (skin), pulmonary (lung), and intestinal. The latter two forms can be fatal if not treated. Spores have been used as a terror weapon. The full lineage of *B.anthracis* is as such: (Domain) Bacteria; (Phylum) Firmicutes; (Class) Bacilli; (Order) Bacillales; (Family) Bacillaceae; (Genus) Bacillus; (Species) Bacillus anthracis.

Clostriduim botulinum: This bacterium causes botulism, on often fatal form of paralysis. Botulinum toxin, and *C. botulinum* cells, have been found in many foods, including canned foods. The full lineage of *C.botulinum* is as such: (Domain) Bacteria; (Phylum) Firmicutes; (Class) Clostridia; (Order) Clostridiales; (Family) Clostridiaceae; (Genus) Clostridium; (Species) Clostridium botulinum.

Francisella tularensis: This bacterium is the causative agent of tularemia. This disease can be transmitted to humans by infected ticks, deerflies, carcasses, or by aerosol, and is a potential bioterrorism agent. The full lineage is given: (Domain) Bacteria; (Phylum) Proteobacteria; (Class) Gammaproteobacteria; (Order) Thiotrichales; (Family) Francisellaceae; (Genus) Francisella; (Species) Francisella tularensis.

Mycobacterium tuberculosis: This bacterium causes tuberculosis, a chronic infectious disease of which incidences are increasing worldwide. This species causes more deaths in humans than any other bacteria. The full lineage is: (Domain) Bacteria; (Phylum) Actinobacteria; (Class) Actinobacteridae; (Order) Actinomycetales; (Family) Mycobacteriaceae; (Genus) Mycobacterium; (Species) Mycobacterium tuberculosis.

Staphylococcus aureus: This bacterium is a major source of hospital and community acquired infections. Strains continue to evolve resistance to various antibiotics and infections can sometimes be uncureable. It can often result in mortality as well as superficial infections and more serious forms of infection such as meningitis. The full lineage is as such: (Domain) Bacteria; (Phylum) Firmicutes; (Class) Bacilli; (Order) Bacillales; (Family) Staphylococcaceae; (Genus) Staphylococcus; (Species) Staphylococcus aureus.

Yersinia pestis: This bacterium is the causative agent of plague (bubonic and pulmonary). It can be transmitted from rats to humans through the bite of an infected flea or from human-to-human through the air during widespread infection. This species was the cause of the Black Death in Europe in the 19th century and a current subtype originating in China persists today. Its full lineage is: (Domain) Bacteria; (Phylum) Proteobacteria; (Class) Gammaproteobacteria; (Order) Enterobacteriales; (Family) Enterobacteriaceae; (Genus) Yersinia; (Species) Yersinia pestis.

Influenza A virus: Influenza A virus infects humans and many other species as the flu. Each year in the U.S., an estimated 36,000 deaths are attributed to the flu. Over 100 subtypes of Influenza A virus exists. Specific subtypes have caused all of the world’s major flu pandemics. This species also includes the subtype H5N1 which has been predicted to cause the next worldwide pandemic. All subtypes have been found in wild birds, which are considered reservoirs for the virus as they can transmit the virus to other animals. The full lineage of this virus is: (Domain) Viruses; (Group) ssRNA negative-strand viruses; (Family) Orthomyxoviridae; (Genus) Influenzavirus A; (Species) Influenza A virus.

Influenza B virus: Influenza B virus also contributes to yearly flu epidemics. It's lineage is: (Domain) Viruses; (Group)ssRNA negative-strand viruses; (Family) Orthomyxoviridae; (Genus) Influenzavirus B; (Species) Influenza B virus.

2.3 Strain Fingerprinting

Strain fingerprinting is a term describing the determination of microbial species and or subtype type from its DNA. Subtypes are variations of species which exhibit genetic differences but still remain in the same species group. Sometimes these changes can cause strong characteristic differences. For example, *Francisella tularensis* is a bacterium causing the disease tularemia in mammalian species. Three subtypes of *F. tularensis* have been described: *F. tularensis* subsp. *tularensis*, *F. tularensis* subsp. *holoartica*, and *F. tularensis* subsp. *mediaasiatica.* The first of these subspecies is considered highly virulent while the others are considered less virulent [50]. To prevent dangerous outbreaks, when occurences of *F. tularensis* are found in animal or human cases, it is important to know to which subtype the specific strain belongs.

DNA extracted from non-living samples can be used in strain fingerprinting methods. With rapid advances in genomic sequencing, DNA based identification methods are increasingly being relied upon to identify pathogenic strains and monitor epidemic outbreaks [31]. One advantage to DNA based strain fingerprinting is that live cultures are not required, reducing the risk of accidental work place exposure to pathogenic microbes [26].

Most current fingerprinting techniques incorporate laboratory methods to locate specific nucleotide sequences which are expected to be expressed in different quantities or with different variations between strain genomes. These techniques rely on DNA probes. Probes are short nucleic acid segments which bind to the sequences of interest in unknown DNA samples. These probes can be used as primers in polymerase chain reactions (PCR) whereby if the sequence of interest does exist in the sample, it is copied exponentially and the quantity is measured (see [9] for more detail). Multiple probes can also be used in miccroarrays where the degree of binding between each probe in the array and the DNA sample can be specific to species and subtype samples [11,12]. In these cases, the design of probes plays an important role in the success of such techniques. Algorithmic methods designed to optimize this process often include determining sequences specific to species groups [24, 43, 54].

Ribosomal RNA (rRNA) plays a role in the manufacture of proteins dictated by DNA coding regions. Genes coding for rRNA are the least variable in all cells and subsequently their nucleotide coding sequences are often used to determine the taxonomy of an organism [59]. Designing probes based on these gene sequences is referred to as ribotyping. Ribotyping has been examined as a means of strain differentiation of *Francisella tularensis* samples, but has proven limited in its ability to differentiate subtypes [15, 17].

Strain differentiation through the use of variable number tandem repeats (VNTR) has been applied to discriminating subtypes of *F.tularemia* [17, 26], *B. anthracis* [28], *Y.pestis* [33, 44], *M. tuberculosis* [23, 34], and *C. difficile* [29]. It has also been reported as a relatively fast method, allowing results in 8-12 hours after obtaining a strain isolate

[15]. Short tandem repeats are DNA regions characterized by the repetition of specific nucleotide words, i.e., TATATATA. Variable number tandem repeats are found at the same locus (position) among different genomes of the same species but show different numbers of repetitions between individuals [53]. In strain differentiation, these repeats can be used as probes.

Fragment Length Polymorphism has been another successful tool employed in strain fingerprinting of such pathogens as *F. tularensis* [17, 50] and *C. difficile* [29]. This process uses restriction enzymes which cut DNA samples at any occurrence of sequences specific to each enzyme. After being digested with these enzymes, each strain sequence is left broken into fragments of different lengths. Depending on which restriction enzyme is used, these fragment lengths should differ between subtypes and enable its identification. While this method has also proven successful in strain subtyping, it is slower than PCR amplification methods [59] and requires about twice as much time [15].

2.4 Oligonucleotide Signatures

Oligonucleotide signatures have been studied extensively in nucleic acid sequence analysis. They are derived by calculating the over- and under- representation of nucleotide words when compared to random expectations. In [7], differences in dinucleotide signatures were examined among prokaryote, plasmid, and mitochondrial DNA. Mitochondrial DNA signature differences corresponded with phylogenetic classifications in that mammalian signatures were found to be similar, while animal and fungal signatures were described as moderately grouped and divergent from plants and protists. Plasmids and their hosts were also significantly similar. To quantify signature differences between genomic sequences, the average absolute dinucleotide relative abundance was calculated, given by the formula:

$$\delta^*(f,g) = 1/16\sum|\rho^*_{XY}(f) - \rho^*_{XY}(g)|$$

where f and g are two sequences, and XY represents all 16 possible dinucleotide pairs, and ρ^*_{XY} is the odds ratio of dinucleotide xy and its reverse compliment [7].

Signatures are not restricted to dinucleotide usage patterns, and can include nonrandom utilization of any oligonucleotide. The authors in [20] examined codon pa**irs**, or hexoligonucleotide abundances in *Escherichia coli*. The codon triplet is the basic reading block in transcription, each codon codes for an amino acid, which in turn helps to build a protein. The authors showed that dicodon usage patterns were nonrandom and that the degree of nonrandomness exceeded that which would be contributed by nonrandom dinucleotide abundances alone. The authors also describe that codon signatures vary considerably between organisms. For each of the possible 3,271 codon pairs, a value was calculated:

$$\chi^2_1 = (\text{observed} - \text{expected})^2/\text{expected}$$

where the expected number of each codon pair was the product of the pairs frequency and the total number of pairs. A codon pair's frequency was the product of the frequency of each codon, under the assumption that codons are used randomly.

Tetranucleotide signatures were examined in [45] via the calculation of tetranucleotide usage departures from random (TUD)s for 27 microbial genomes. The authors found that a zero order Markov chain, which measured tetranucleotide bias assuming random nucleotide distributions, gave the best distinction between species signatures. In this approach, the frequency of a word F(W) was a ratio of the observed O(W) and the expected E(W). Through this approach, phylogenetically related species showed similar TUDs.

In [6], word over- representation was determined if the probability calculated for an oligonucleotide and its reverse complement were greater than one ($\rho_{XY} >1$ and $\rho_{I(XY)} >1$). Similarly, both values less than one indicated under- representation. In the calculation of all probabilities, adjustments were made for the complimentary antiparallel nature of DNA. For example, the adjusted frequency of the mononucleotide A would be $f^*_A = f^*_T = \frac{1}{2}(f_A + f_T)$. Similarly, $f^*_G = f^*_C = \frac{1}{2}(f_G + f_C)$. For dinucleotides, an example for GT is given: $\rho^*_{GT} = f^*_{GT}/f^*_G f^*_T = 2(f_{GT} + f_{AC})/(f_G + f_C)(f_T + f_A)$. Trinucleotides were calculated as : $\gamma^*_{XYZ} = f^*_{XYZ} f^*_X f^*_Y f^*_Z / f^*_{XY} f^*_{YZ} f^*_{XNZ}$, where $f^*_{XYZ} = \frac{1}{2}(f_{XYZ} + f_{I(XYZ)})$, I(XYZ) is the inverted complement of trinucleotide XYZ, and N is some nucleotide.

The specific mechanisms which maintain short oligonucleotide frequencies throughout a genome are not clearly understood. Some of these mechanisms may have opposing results creating a mixed composition [14]. In a study of *E.coli*, it was suggested that a Very Short Patch (VSP) repair mechanism may be responsible for the under-representation of T containing oligonucleotides (i.e. CTAG) and the over-representation of C containing oligonucleotides (i.e. CCAG) [4]. The VSP repair in *E.coli* targets T:G mismatches in the sequences N**T**WGG/N'**G**W'CC and C**T**WGN/G**G**W'CN' (W = A or T, N = A,C,T,or G, primes ' indicate complementary bases and slashes / indicates double stranded pairs) and converts them to C:G base pairs. The authors used first, second and third order Markov chains to test for over- and under- representation of penta, tetra and tri nucleotides expected to be affected by VSP repair. Results supported the VSP repair hypothesis, short oligonucleotides favored C:G over T:A bases in all expected sequences.

Restriction-Modification (R-M) systems have also been suggested as a mechanism contributing to short oligonucleotide bias throughout a genome. Type II R-M systems recognize specific nucleotide sequences and cleave within or very close to those sequences [30]. In [18], a direct connection was found between type II R-M target sites and under-representation of those palindromic sequences in several bacterial genomes. Palindromes are the combination of nucleotide sequences followed by their reverse complements, i.e. TGGCCA. Markov models were used in the derivation of positive or negative representation of 4, 5, and 6 base pair(bp) palindromes in *Haemophilus influenzae*, *E.coli*, *Methanococcus jannashii*, *Bacillus subtilis*, *Mycoplasma genitalium*, *Synechocystis* sp., *Marchantia polymorpha* mitochondrion and chloroplast genomes. Palindromes representing R-M target sites within the same genome were highly avoided while palindromes representing R-M target sites from foreign genomes were mildly avoided. It was suggested that bacterial genomes in environmental populations may be exposed to R-M systems from other strains and species through lateral gene transfer. This was proposed to explain the avoidance of palindromes not specifically targeted by indigenous R-M systems. The researchers also showed that in mitochondrial and chloroplast DNA which do not contain R-M systems, palindrome avoidance was not detected. R-M systems were compared between two *Helicobacter Pylori* strains in [32]. Though both strains shared 90% of their R-M systems, it was found that only strain specific R-M genes were expressed. The shared genes were inactive. This suggests that different strains may differ in palindrome avoidance patterns as a result of R-M systems.

Chapter 3

Strain Signature Comparisons

A preliminary comparison of strain word signature differences is presented in this section. The method of DNA sequence comparison used hinged on two parameters. The first was the method for determining the degree of over- or under- representation of any given word. The second was the length of word to use. Detailed descriptions and reasons for selection of these parameters are described in sections 3.1 and 3.2.

Oligonucleotide signatures of short word lengths (2-5) have been shown to be similar for organisms within kingdom groupings [7] as well as species specific [27, 45, 49]. They are also consistent enough to be used in the regrouping of mixed fragments from multiple species genomes [1, 36, 49].

While genome signatures show consistencies within phylogenetic groups, examining differences at the strain level may reveal significant differences that could enable strain classification and give insight into strain evolution. In this work, genomic signature differences between strains within the same subspecies group were examined as preliminary research in the potential building of a strain classifier. To extract strain differences, over- and under- represented words were examined as potential feature sets. The most significant results stemming from this research concerned the under-represented subset which allowed the highest level of strain differentiation and even seemed to cluster along lineage groupings.

Staphylococcus aureous is an opportunistic pathogen in humans and animals characterized by high genetic plasticity. Many drug resistant and virulent strains have evolved in hospital and clinical settings [21], and deaths caused by *S. aureous* have reached epidemic proportions. Ten strains of bacterium S*taphylococcus aureous* subs. *aureous* (Mu50, NCTC8325, USA300, N315, Mu3, MSSA476, MRSA252, RF122, COL, NEWMAN) are compared against each other in terms of their oligonucleotide signatures. These strains have been collected from Japan, United Kingdom and the United States.

3.1 Markov Chain Selection

In genomic signature analysis, Markov chains have often been utilized to calculate nucleotide word bias. In Markov chains, the current state of a system is predicted by its previous states. In signature analysis, this translates to predicting a word frequency based on the observed frequencies of its subwords. Depending on the degree of the Markov model, bias contributing to a word of length *m* from occurrences of subwords of length 1..m-1 can be removed. For example, in a sequence dominated by TA and AG, unless specifically selected against, TAA and AAG will naturally show high frequencies due to the abundance of their sub words. With the ultimate goal of matching DNA's internal word selection mechanisms, the optimal degree of Markov model to use remains undetermined. The expected count of a nucleotide word (W) of length m with a maximal order Markov model is written as follows:

$$\mathbf{E(W) = N(w_1w_2w_3..w_{m-1})\ N(w_2w_3...w_m)\ /\ N(w_2w_3..w_{m-1})}$$

$w_1 \ldots w_m$ represent characters forming word (W) of length m. While a zero order Markov model would predict:

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

where A,C,T,G represent nucleotide frequencies in a given sequence of length N and a,c,t,g are the number of each nucleotide in word W. Findings in [45] and [46] suggest that minimal order Markov models allow the most differentiation between species genomic signatures. Based on this past research, minimal order models were used to calculate E(W) in this work. The degree of over or under representation of each word was then derived by the ratio of O(W)/E(W) counts (see [45] for more detail).

3.2 Signature Subset Selection

Oliognucleotide word length selection was based on the maximal length which enabled random representation of all possible words in a signature. Signatures of long word lengths may include more words than a microbial genome can contain. *S.aureous* strain sequence lengths ranged from 2,799,802 to 2,902,619 (bp). With a word length of eight, this allowed for an average of 43 occurrences of each possible word (~2,850,000/65,536).

Histograms were created by binning word counts into their over- or under-representation calculations (Figure 3.1). The complete octanucleotide signature showed similar histograms across all strains, illustrating the high degree of coincidence in strain signatures. Two signature subsets were then examined as potential differentiating feature sets for strain identification. These included highly over- represented words and under-represented words, determined by words absent from at least one strain.

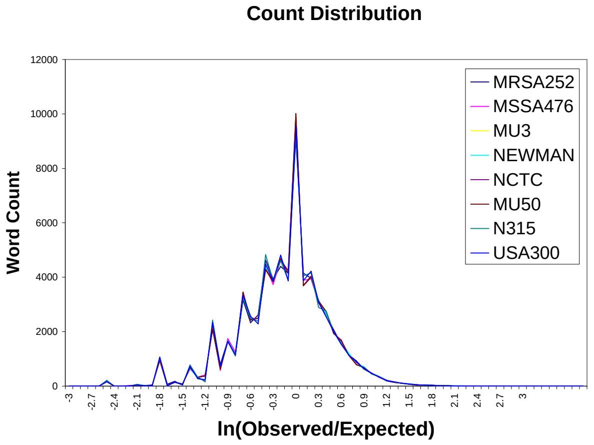

Figure 3.1. Word counts for each frequency bin. Bin size 0.1 increments of ln(O(W)/E(W).

To examine how the expression of each word in the set varied across strains, the normalized interstrain variance was also derived for each word in the signature set and was calculated as:

$$V(W) = \sum_{i=0}^{n} \frac{(x_i - \mu)}{(n-1)}$$

Where x_i = O(W)/E(W) for word (W) in strain (i) , μ = mean (x_i) across strains, n = number of strains compared. Before variance was derived, O(W)/E(W) values for each word were normalized across strains by reducing the range of values to a common scale between zero and one. The formula used to achieve this was:

$$x_{i\ norm} = (x_i - x_{min})/(x_{max} - x_{min})$$

Where x_i = O(W)/E(W) for word (W) in strain (i), $x_{i\ norm}$ is the normalized value, x_{max} is the maximum O(W)/E(W) for word (W) across all strains in a group, and x_{min} is the minimum O(W)/E(W) for word (W) across all strains in a group. Without normalization, highly over- represented words resulted in the highest interstrain variance measures due to the larger magnitude of their O(W)/E(W) values, regardless of their relative interstrain variance. Variance calculations are referred to in section 3.4.

3.3 Over- Representation Subset

Words over- represented in any genome by at least 10x their expected values were examined for this subset. It is not predicted that this group would contribute to strain differentiation. This is illustrated by figure 3.2 in which O(W)/E(W) values across all strains converge.

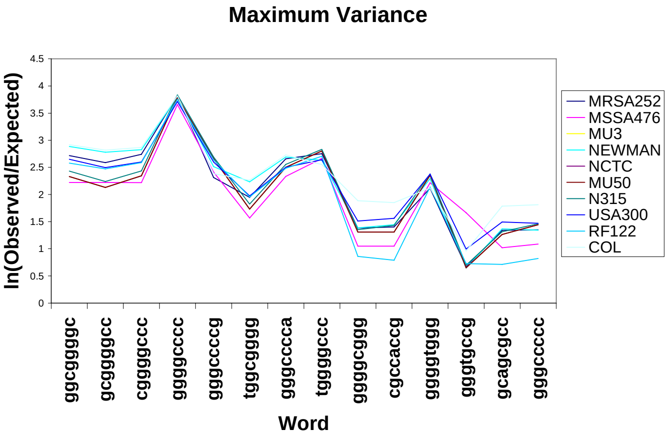

Figure 3.2. ln (O(W)/E(W)) of over represented words.

3.4 Absence Subset

The absence subset was created by selecting all words absent from at least one strain but not absent from all. This subset enabled groupings based on lineage and offers the most potential as a feature set for strain classification and differentiation. This group included 1136 words.

Table 3.1 Strain avoidance pattern for large absent word clusters.

Cluster size	MSSA476	RF122	MRSA252	COL	USA300	NEWMAN	NCTC	MU3	MU50	N315
125		Absent								
114	Absent									
74		Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent
69	Absent		Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent
52	Absent	Absent								
49			Absent							
49			Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent
46							Absent	Absent	Absent	Absent
24			Absent	Absent	Absent	Absent				
18				Absent	Absent	Absent	Absent	Absent	Absent	Absent
16	Absent	Absent	Absent	Absent	Absent	Absent				
14	Absent	Absent	Absent							
14		Absent					Absent	Absent	Absent	Absent
14	Absent	Absent					Absent	Absent	Absent	Absent

Table 3.1 shows word groups, or clusters absent from at least one strain. The largest word cluster of 125 was avoided solely by RF122, a strain obtained from cattle and assumed to be methicillin sensitive. The second largest word cluster of 114 was avoided by the methicillin sensitive strain MSSA476. The next two largest word clusters contained words uniquely expressed by RF122 and MSSA476 respectively. The distinction of these two strains may represent earlier lineages which have not evolved to the state of antibiotic resistance. Table 3.1 also shows identical words avoidance patterns for the COL, NEWMAN, USA300 group as well as the NCTC, MU3, MU50, N315. Strains MU3, MU50, and N315 were all collected in Japan, while the NCTC strain is

considered an international specimen. If these strains are derived from a similar lineage, it may explain their grouping. The origin of the COL and NEWMAN strains is unreported, but they may share common ancestry with the USA300 strain based on these results. The methicillin resistant strain MRSA252, originating from the U.K. was more similar to the other methicillin resistant strains in avoidance patterns than to its U.K. counterpart MSSA476.

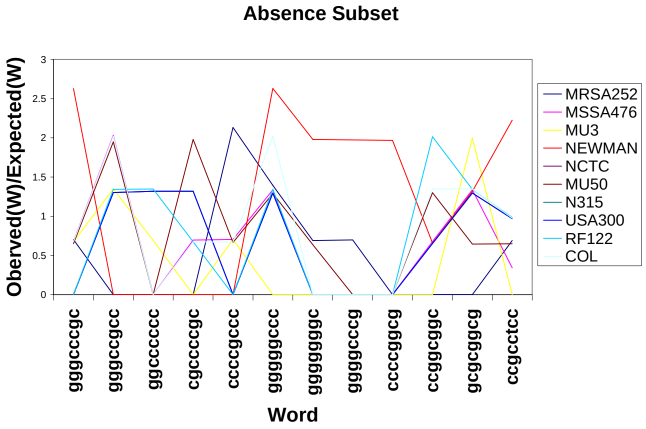

Figure 3.3. O(W)/E(W) for some words in the absence subset.

A graph of some of the most varied words within the absence subset shows visible signature differentiation between strains (Figure 3.3). In addition, words from the absence subset resulted in the highest normalized interstrain variance. The 60 words exhibiting the highest variance were solely from this subset. It is suggested that the absent word subset could contribute significantly to strain differentiation with standard classification methods and may also enable insight into lineage histories.

Chapter 4

A Supervised Strain Classifier

While current strain fingerprinting methods are able to resolve species subtypes, they are not sensitive enough to differentiate individual strain lineages. Furthermore, most comparative genomics approaches involve alignment to compare related genomes. This research builds on results in Chapter 3 and proposes a strain classification method based on the absence subset of oligonucleotide signatures. The goal of this approach is to increase strain classification sensitivity as well as to avoid the computationally expensive process of sequence alignment. The advantage of this process is that it extracts differences between strain sequences while masking their high degrees of similarity. A standard supervised classification approach, described in Section 4.1, will be used with the Manhattan distance as the differencing factor between training and test genomes. Feature vectors will be derived at classification run time and will be based on training data. Specifically, feature vectors will contain signature values for words of a specified length which are absent from at least one training genome (but not absent from all). Signatures of word lengths up to 10bp will be tested to determine the shortest word length allowing the highest degree of accuracy for each classification.

The discriminatory power of this approach will be tested at increasing levels of phylogenetic resolution including classifying genomes from different genera, species within the same genus, subtypes within a species, and strain lineages within a subtype. The final classification will be used as a test to determine geographic region of *Influenza*

A virus H5N1 acquisition from bird host to human. Each classification schema is described in more detail in Section 4.4. All viral and bacterial species used for classification purposes are pathogenic to humans or contain pathogenic subtypes. This characteristic along with data availability dictated the choice of genomes used for ensuing classifications.

4.1 Supervised Classification

In supervised classifications, test data is comprised of "unknown" data from samples which must be classified. Training data denotes "known" data which will be used as class identifiers. This supervised algorithm compares each test data sample to each class identifier, and assigns it to which ever class it is closest to. Methods of test and training data comparisons vary widely and have engendered many different classification algorithms. For strain classification, a set of strain genomes will be selected as training data and each will represent a unique class. The algorithm will assign all other test genomes to the class each is closest to, using the Manhattan distance metric as a distance measure. Before classification, all genomes will be reduced to equal length feature vectors. The feature vector associated to each genome will be the absence subset of training genome oligonucleotide signatures. This is the observed to expected ratio (O(W)/E(W)) for each member in the set of words for which at least one is absent and at least one is present in the group of training genomes. This enables the derivation of a unique feature set for each classification run which has been tailored to the differences between the particular training genomes being used. E(W) for each word will be calculated using a zero order Markov model as described in Section 3.1. The difference between two genomes will then be the sum of differences between their respective feature vectors, the Manhattan distance. If feature vectors contain K features and a comparison is made between two genomes p and q, then the difference between them is calculated as:

$$\text{dif}[p,q] = \sum_{k=1}^{K} |p(k) - q(k)|,$$

the number of words in the absence subset with different expected frequencies between genomes P and Q.

4.2 Feature Set Selection

The finite nature and small alphabet of DNA sequences lends itself well to representation by formal language definitions. Knowing which words are avoided by a specific genome makes its set of all subwords decidable and theoretically enables their representation by finite automata. While this topic is not explored in this thesis, the feature set used in this classification is written as a formal language. The feature set used to represent each genome in comparisons is O(W)/E(W) for all words in the oligonucleotide signature which at least one training genome does not express and at least one training genome does express. This word set may be written as a formal language as such:

Let $\sum$ = **{a,c,t,g},**

Let the set of all possible nucleotide words on $\sum$ with length n be written as:

$$\mathbf{L = \{w:|w| = n\}}$$

Let training genomes 1,..,N be represented as strings $G_{1,..,}G_N$.

Let $L_{1..N}$ represent the set of all words of length n found in training genomes $G_1..G_N$ such that:

$$\mathbf{L_1 = L \cap \{w: w \text{ is a substring of } G_1\}}$$

$$\mathbf{L_2 = L \cap \{w: w \text{ is a substring of } G_2\}}$$

…..

$$\mathbf{L_N = L \cap \{w: w \text{ is a substring of } G_N\}}$$

Let $L_{PRESENT}$ represent the set of all words in $L_1 \ldots L_N$ such that:

$$\mathbf{L_{PRESENT} = L_1 \cup L_2 \ldots . \cup L_N}$$

Let L_{ABSENT} represent all words not found in $L_1 \ldots L_N$ such that:

$$\mathbf{L_{ABSENT} = L_1{}' \cup L_2{}' \ldots \cup L_N{}'}$$

Then the classification feature set $L_{FEATURE}$ is:

$$\mathbf{L_{FEATURE} = L_{PRESENT} \cap L_{ABSENT}}$$

4.3 The Classification Algorithm

All of the code and algorithmic design described in this thesis were developed independently by the author. Oligonucleotide signature sets grow exponentially with increasing word length and reliance on dynamic memory as opposed to array-based data structures soon becomes crucial. The central data structure for this application is a linked list of linked lists, created from the C++ Standard Template Library (STL). Each word in the signature set can be considered a node in the primary linked list. From each node is a list of the O(W)/E(W) calculation for each genome in a test or training data set. Thus, after all genomes are processed, the primary list can be examined for usable feature set words. In this case, we select those that contain at least one zero value and one value greater than zero (figure 4.1). These feature set word values are then inserted into training and test vector lists and the classification is performed.

To compare two genomes with brute force would require determining the maximum alignment existing between them. Reducing each genome to an oligonucleotide feature set removes the need for alignment to compare genomes. At this point, each genome is represented by a single vector containing values for each word in the signature. Further reduction of this signature set to the absent word subset allows even faster comparison between genomes. The actual feature set size for each classification is given in Chapter 5.

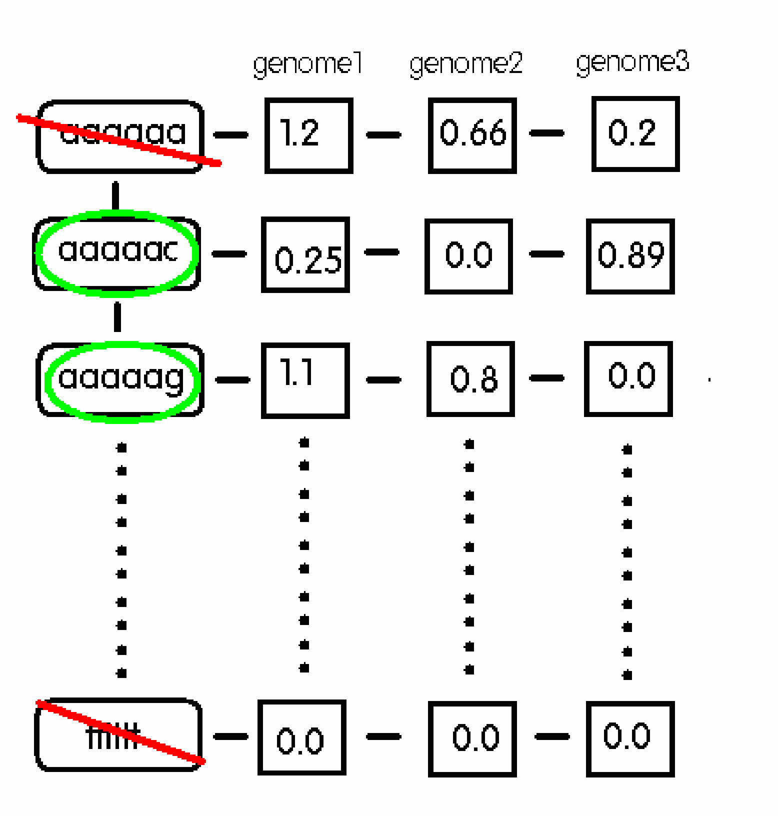

Figure 4.1. Linked List Arrangement and Word Selection.

The code written for this thesis maintains the flexibility to calculate signature sets of words with a user specified length. Each genome is processed in the same manner and results are either inserted into test or training data lists. The main processing goals per genome are achieved in the following order:

1. Calculate the observed count of each word O(W) in the oligonucleotide word set of word length n.
2. Calculate the expected count of each word, E(W) using a zero order Markov model.
3. Insert values into training or test data lists.
4. Perform feature set extraction and classification.

Step 1 is described in most detail in this section. Obtaining word counts for each genome involves reading in each word of the specified length and incrementing the count for that specific sequence. It is important to note that words were overlapping and not adjacent, see figure 4.2.

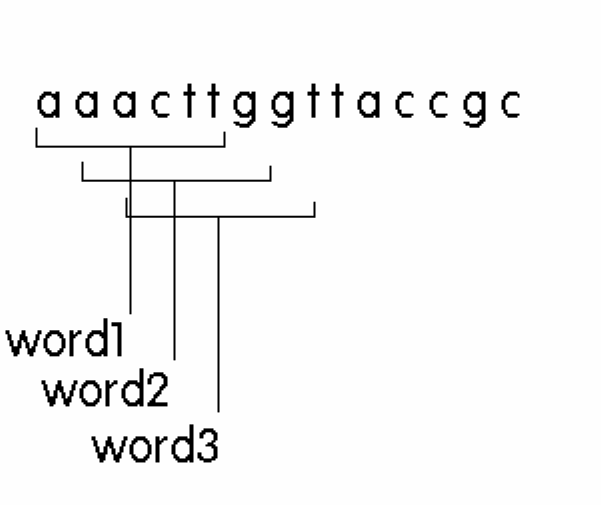

Figure 4.2. Overlapping Words

Counts for each word are stored in a numeric array because each word can be converted to its corresponding integer index in the array. Let a word of length n be represented by the character array word[n]. The indexing formula for converting this nucleotide string to its integer value index is given as:

$$\text{indx} = \sum_{i=0}^{n} \text{num}(\text{word}[i]) * 4^{(n-i-1)}$$

where num(word[i]) is a function which returns the integer value designated for each nucleotide in a word, a = 0, c = 1, g = 2, t = 3. Thus, aaaaaa would be indexed at 0 and tttttt would be indexed at 4095. While this indexing formula was derived independently for this thesis, it has also been used for other word counting programs [16, 25].

It should be noted that many genomic sequences contain non-nucleotide entries to represent unknown values, usually denoted by N, n, X or x. Genomic algorithms should account for this by checking each read character or word to ensure that it is fully composed of real nucleotide values. The pseudocode for counting all occurrences of words of length n in a genome is given in figure 4.3.

When nucleotide words are represented numerically so that a = 0, c = 1, g = 2, t = 3, the signature word list becomes a tertiary counting system. Iterating through all words in the set can then be accomplished by adding 1 to the previous with tertiary addition rules adhered to. Character word arrays must first be converted to integer arrays of the same length following the described substitution rules. To iterate through all words, the initial integer word is set to all zeros, representing $a_1a_2\ldots a_n$. To create the next word, a one is added and the rules of tertiary addition are implemented. Specifically, if any single integer digit is greater than 3, reduce the 3 to zero and carry a one to the higher order digit.

```
* A function "getindex()" is used in this pseudocode. This is based on the index
calculation description described above.

start Count(input = n)          //input desired word length

        declare character word[n]  //input word container
        declare double wordcount //total number of words found in genome
        declare double list[n4] //array list of counts for each word in signature set

        for( i = 0 to n-2)          //obtain n-1 characters in first word
                readin word[i]
        end for

        while(readin word[n-1]) //while readin character from file insert as last char in
                                        word
                if(word) //if word contains all nucleotide values – no errors
                   wordcount++
                   list[getindex(word)] //increment value in list at word index location
                end if

                for(i = 0 to n -2) //shift all letters to left to prepare for next char input at
                                        end
                  word[i] = word[i+1]
                end for
        end while
end Count()
```

Figure 4.3. Word counting pseudocode.

4.4 Data

A large database of *Influenza* A and B *virus* genomes has recently been made publicly available by the National Institute of Allergy and Infectious Disease [2]. This database contains multiple isolates of different strains of the Influenza virus from diverse geographic regions, times and host species. Due to the availability of data, all classifications except the first focused on the Influenza A and B virus species. The remaining genomes were acquired from the pubmed website [57].

4.5 Classification Applications

This section describes five classification applications using the algorithm presented above. Details regarding exact subtype, lineage and collection dates are provided in the results for each classification.

Mixed Pathogens: Bacterial and viral species groups which are generally regarded as threats to public health through pandemic or bioterrorism events available through pubmed were selected for the first classification. Genomes from organisms from different genera are used to test the proposed method at a low differencing resolution. Multiple strains representing seven species from different genera are used. Species include *B. anthracis, C. botulinum, F. tularensis*, *S. aureus, M. tuberculosis, Y. pestis* and *Influenza A virus.*

Influenzavirus A/B: Both Influenza A and B virus are different species within the same genus, Influenzavirus. This classification is used to determine whether the proposed method can discriminate between these different species within the same family. *Influenza A virus* has many subtypes and most available are included in this datum group. *Influenza B virus* is not divided into subtypes so this was not a determining factor when selecting sample genomes. Ten *Influenza A virus* and ten *Influenza B virus* genomes are used in this classification. All genomes used are from human hosts.

***Influenza A virus* subtypes:** This classification was designed to test whether the proposed method can discriminate between different subtypes of *Influenza A virus* genomes from the same host species. Genomes used for this classification represent five human host subtypes with six genomes per subtype.

Avian Influenzavirus A, H5N1 Origin: The Avian *influenza A virus* H5N1 subtype is a highly pathogenic avian influenza (HPAI) which can cause high mortality in poultry and kill 90-100% of infected chickens [48]. Domestic poultry have historically been key in the harboring of flu strains which can cross over into humans and have caused major pandemics in the 20th century [42]. This subtype, H5N1, has infected humans and subsequently has become a focus as a potential source of the next major flu pandemic [39, 41, 42].

This classification tests the proposed method in its ability to discriminate between individual avian H5N1 strain lineages and thereby determine their geographic origins. Genomes from domestic bird (chicken, duck, turkey, goose) outbreaks in China (2006), Africa (2006), Thailand (2006), and Vietnam (2005) are used for this classification. Regions represented in the China dataset include Guanxi province, Hunan province, Guandong province, and HongKong. Countries represented in African dataset include Afganistan, Nigeria and Sudan. Specific regions of genomes collected in Thailand and Vietnam were not indicated on the NIAID website and are thus referred to by their country name.

Training genomes were selected in attempts to represent major avian H5N1 outbreaks in each region of interest during specified time periods and included 17 genomes from domestic birds. Test genomes include 61 domestic avian H5N1 genomes from time periods corresponding with training genomes in the same regions and included multiple, concurrent samples from all areas.

Avian to Human Transmission of *Influenza A virus* (H5N1): As of April, 2007, there have been 292 cases of Avian H5N1, 192 of which proved fatal [58]. The first case was in Hong Kong in 1997 [5]. This classification attempts a cross host species matching of individual strain lineages between avian and human cases. Human acquired *Influenza* A *virus* H5N1 genomes from Indonesia in 2005, Thailand in 2004, and Vietnam in 2004 are used as test data. The training data set included all H5N1 strains from domestic bird hosts in Indonesia, Thailand and Vietnam during the same years of 2005, 2004 and 2004 respectively. Thus, human cases are assigned to their closest bird counterparts to determine whether the proposed method can point to the location of viral crossover from bird to human.

Chapter 5

Classification Results and Analysis

5.1 Mixed Pathogens Results

Table 5.1. Results summary for mixed pathogens.

accuracy:	**100%**
word length:	**6**
feature set size:	**643**

The application of the proposed method resulted in 100% accuracy with all genomes being assigned to their correct species groups. This classification requires the least amount of discriminatory power as all genomes are from different genera. The genome dataset includes entire sequences of four *B. anthracis* strains, three *C. botulinum* strains, six *F. tularensis* strains, eleven *S. aureous* strains, four *M.tuberculosis* strains, six *Y. pestis* and nine *Influenza A virus* strains. *B. anthracis* is a very genetically homogenous species and is not classified into subtypes. Thus, this group contains the four different strains available on Genbank. The *C.botulinum* group is composed of four different strains from subtypes A and F. The *F.tularensis* group includes three subtypes; holarctica, novicida, and tularensis. The *S.aureous* dataset contains ten strains of the human pathogen subtype aureous, nine of which are methicillin resistant while one is methicillin sensitive. This group also contains one genome of subtype RF122 which infects cattle. The four strains in the *M.tuberculosis* group are from the same subtype. This species, like *B. anthraci*s, is also highly genetically homogeneous and is not divided into subspecies [29]. The *Y.pestis* group contains seven strains which represent all four subtypes; Antiqua, Mediaevalis, Orientalis and Microtus. The *Influenza A virus* set contains ten samples collected from human hosts including subtypes H1N1, H1N2, H2N2, H3N2, H5N1.Table 5.2 presents the individual genomes used as training data. Table 5.3 presents strains used as test data and the classes to which they are assigned.

Table 5.2. Mixed pathogen training data.

Training Data: Mixed Pathogens

Species/Class	subtype/strain	Common disease name
Bacillus anthracis	Ames	Anthrax
Clostridium botulinum	A /ATCC 19397	Botulism
Francisella tularensis	holarctica	Tularemia
Staphylococcus aureus	aureus/MRSA252	Staph
Mycobacterium tuberculosis	CDC1551	Tuberculosis
Yersinia pestis	Angola	Plague
Influenzavirus A	H1N1/Brazil/11/1978	Flu

Table 5.3. Mixed Pathogen, test data and classification results.

Test Data: Mixed Pathogens

Species	subtype/strain	Common disease name	Classified as
Bacillus anthracis	A2012	Anthrax	Anthrax
Bacillus anthracis	'Ames Ancestor'	Anthrax	Anthrax
Clostridium botulinum A	ATCC 3502	Botulism	Botulism
Clostridium botulinum A	Hall	Botulism	Botulism
Clostridium botulinum F	Langeland	Botulism	Botulism
Francisella tularensis	holarctica FTA	Tularemia	Tularemia
Francisella tularensis	holarctica OSU18	Tularemia	Tularemia
Francisella tularensis	novicida U112	Tularemia	Tularemia
Francisella tularensis	tularensis FSC198	Tularemia	Tularemia
Francisella tularensis	tularensis SCHU S4	Tularemia	Tularemia
Francisella tularensis	tularensis WY96-3418	Tularemia	Tularemia
Staphylococcus aureus	aureus MSSA476	Staph	Staph
Staphylococcus aureus	aureus Mu3	Staph	Staph

Staphylococcus aureus	aureus Mu50	Staph	Staph
Staphylococcus aureus	aureus N315	Staph	Staph
Staphylococcus aureus	aureus NCTC 8325	Staph	Staph
Staphylococcus aureus	aureus str. Newman	Staph	Staph
Staphylococcus aureus	RF122	Staph	Staph
Mycobacterium tuberculosis	F11	Tuberculosis	Tuberculosis
Mycobacterium tuberculosis	H37Ra	Tuberculosis	Tuberculosis
Mycobacterium tuberculosis	H37Rv	Tuberculosis	Tuberculosis
Yersinia pestis	Antiqua	Plague	Plague
Yersinia pestis	CO92	Plague	Plague
Yersinia pestis	KIM	Plague	Plague
Yersinia pestis	biovar Microtus str. 91001	Plague	Plague
Yersinia pestis	Nepal516	Plague	Plague
Yersinia pestis	Pestoides F	Plague	Plague
Influenza A virus	H1N1/Cam/1946	Flu	Flu
Influenza A virus	H1N2/New York/78/2002	Flu	Flu
Influenza A virus	H1N2/New York/226/2003	Flu	Flu
Influenza A virus	H2N2/Albany/1/1968	Flu	Flu
Influenza A virus	H2N2/Canada/720/2005	Flu	Flu
Influenza A virus	H3N2/AuklandNZ/602/2001	Flu	Flu
Influenza A virus	H3N2/England/1972	Flu	Flu
Influenza A virus	H5N1/Hong Kong/212/2003	Flu	Flu
Influenza A virus	H5N1/Vietnam/CL26/2004	Flu	Flu

5.2 *Influenza A virus /Influenza B virus* Results

Table 5.4. Results summary for *Influenza* A/B *virus*.

accuracy:	**100%**
word length:	**6**
feature set size:	**783**

Classification accuracy for this dataset was also 100%. Nine genomes of *Influenza A virus* and nine genomes of *Influenza B virus* were accurately assigned to their correct species class. Training genomes include one additional randomly selected genome representing each class. These results were somewhat surprising considering that the

Influenza A and *B virus* datasets are highly diverse in both geographic origins and acquisition dates. The *Influenza A virus* test data contains genomes from Cambodia in 1946, UK in 1933, New York in 2003, Albany, New York in 2005, Canada in 2005, England in 1972, Indonesia in 2005, Canada in 2004, and Hong Kong in 1999. These genomes also represent seven distinct subtypes which included H1N1, H1N2, H2N2, H3N2, H5N1, H7N3, and H3N2. The *Influenza A* training genome is a H3N2 subtype collected in Auckland, New Zealand in 2005. The *Influenza B virus* test data is composed of genomes from China, Guangzhou province in 1972, China, Nanchang province in 2002, Cordoba Spain in 2007, the Czech Republic in 1994, St. Petersburg, Russia in 1979, Victoria, Australia in 1990 and 2000, and Vienna, Austria in 2006 and 1999. The training genome for this group was collected in Chile in 2000. Table 5.5 lists training genomes, table 5.6 contains test genomes and classification results.

Table 5.5. *Influenza* A/B *virus* training data.

Training Data: *Influenza A & B virus*	
Species/Class	**subtype/strain**
Influenza A virus	H3N2/AuklandNZ/602/2001
Influenza B virus	Chile/16188/2000

Table 5.6. *Influenza* A/B *virus* test data and classification results

Test Data: *Influenza A & B virus*		
Species/Class	**subtype/strain**	**Classified As**
Influenza A virus	H1N1/Cam/1946	InfA/H3N2/AuklandNZ/602/2001
Influenza A virus	H1N2/UK/WilsonSmith/1933	InfA/H3N2/AuklandNZ/602/2001
Influenza A virus	H1N2/New York/226/2003	InfA/H3N2/AuklandNZ/602/2001
Influenza A virus	H2N2/Albany/1/1968	InfA/H3N2/AuklandNZ/602/2001
Influenza A virus	H2N2/Canada/720/2005	InfA/H3N2/AuklandNZ/602/2001
Influenza A virus	H3N2/England/1972	InfA/H3N2/AuklandNZ/602/2001
Influenza A virus	H5N1/Indonesia/CDC184/11/08/2005	InfA/H3N2/AuklandNZ/602/2001
Influenza A virus	H7N3/Canada/rv504/2004	InfA/H3N2/AuklandNZ/602/2001

Influenza A virus	H3N2/HongKong/1073/1999	InfA/H3N2/AuklandNZ/602/2001
Influenza B virus	China/Guangzhou/5/1972	InfB/Chile/16188/2000
Influenza B virus	China/Nanchang/3162/2002	InfB/Chile/16188/2000
Influenza B virus	Spain/Cordoba/01/2007	InfB/Chile/16188/2000
Influenza B virus	Czech Republic//630/1994	InfB/Chile/16188/2000
Influenza B virus	Russia/St. Petersburg/1979/	InfB/Chile/16188/2000
Influenza B virus	Australia/Victoria/69/1990	InfB/Chile/16188/2000
Influenza B virus	Austria/Vienna/14/2006	InfB/Chile/16188/2000
Influenza B virus	Australia/Victoria/504/2000	InfB/Chile/16188/2000
Influenza B virus	Austria/Vienna/1/1999	InfB/Chile/16188/2000

5.3 *Influenza A virus* Subtypes, Results

Table 5.7. Results summary for *Influenza* A *virus* subtypes.

accuracy:	**96%**
word length:	**5**
feature set size:	**43**

The accuracy in classifying 25 *Influenza A virus* subtypes to their correct subtype classes was 96%. Five genomes of each subtype including H1N1, H1N2, H2N2, H2N3, and H5N1 are used as test data. Training data contains one additional randomly selected genome from each of the five subtypes. It is important to note that all subtype groups were accurately classified except for the H3N2 group, from which one genome was assigned to the H2N2 group. This genome is an older strain with a collection date in 1960. Being assigned to the H2N2 class which is also represented by an older strain, "H2N2/Japan/305/1957" may suggest an influence of relative circulation time overriding subtype groupings. The procedure for subtyping Influenza A genomes only examines two genes out of the ~120,000 bp sequences, see [59] for a description. The classification method described in this thesis compares global statistics relating to the entire genome. While this is an inaccurate classification, it may also indicate a particular relationship between the H3N2 and H2N2 influenza subtypes that emerged in some portion of the 20th century that is not reflected in subytpe grouping indicators. The H3N2 subtype has been found to contain a broad range of diversity in [39] and it has been suggested that current I*nfluenza A virus* subtyping methods are not comprehensive enough to represent strain groupings [19, 22]. A larger dataset with a more detailed attention to time periods as well as subtypes may allow better accuracy and enable insight into the evolution of individual lineages within and across subtype groupings. Tables 5.5 and 5.6 present training data, test data, and classification results. The misclassified genome in table 5.6 is denoted with bold lettering and an asterisk.

Table 5.8. ***Influenza A virus*** **subtypes test data.**

Training Data: *Influenza A virus* Subtypes		
Species/Class	**Subtype/Strain**	**host**
*Influenza A virus/*H1N1	H1N1/UK/Wilson-Smith/1933	human
*Influenza A virus/*H1N2	H1N2/New York/296/2003	human
*Influenza A virus/*H2N2	H2N2/Japan/305/1957	human
*Influenza A virus/*H3N2	H3N2/New York/197/2003	human
*Influenza A virus/*H5N1	H5N1/China/Beijing/01/2003	human

Table 5.9. Influenza A virus subtypes, training data and classification results

Test Data: *Influenza A virus* Subtypes			
Species/Class	**subtype/strain**	**host**	**Classified as**
Influenza A virus /H1N1	H1N1/Brazil /11/1978	human	H1N1/UK/Wilson-Smith/1933
Influenza A virus /H1N1	H1N1/Cambodia /1946	human	H1N1/UK/Wilson-Smith/1933
Influenza A virus /H1N1	H1N1/China/Nanchang /8/1996	human	H1N1/UK/Wilson-Smith/1933
Influenza A virus /H1N1	H1N1/Henry/1936	human	H1N1/UK/Wilson-Smith/1933
Influenza A virus /H1N1	H1N1/South Australia /64/2000	human	H1N1/UK/Wilson-Smith/1933
Influenza A virus /H1N2	H1N2/New York /78/2002	human	H1N2/New York/296/2003
Influenza A virus /H1N2	H1N2/New York /226/2003	human	H1N2/New York/296/2003
Influenza A virus	H1N2/New York	human	H1N2/New York/296/2003

/H1N2	/400/2003		
Influenza A virus /H1N2	H1N2/New York /491/2003	human	H1N2/New York/296/2003
Influenza A virus /H1N2	H1N2/New York /C1/2003	human	H1N2/New York/296/2003
Influenza A virus /H2N2	H2N2/Albany /1/1968	human	H2N2/Japan/305/1957
Influenza A virus /H2N2	H2N2/Canada /720/2005	human	H2N2/Japan/305/1957
Influenza A virus /H2N2	H2N2/Sinagpore /1/1957	human	H2N2/Japan/305/1957
Influenza A virus /H2N2	H2N2/South Korea /426/1968	human	H2N2/Japan/305/1957
Influenza A virus /H2N2	H2N2/Taiwan /1964	human	H2N2/Japan/305/1957
Influenza A virus /H3N2	H3N2/NewZealand /Auckland/602/2001	human	H3N2/New York/197/2003
Influenza A virus /H3N2	H3N2/England /1972	human	H3N2/New York/197/2003
Influenza A virus /H3N2	H3N2/New York /786/1993	human	H3N2/New York/197/2003
Influenza A virus /H3N2	H3N2/Australia /Northern Territory /60/1968	human	**H2N2/Japan/305/1957** *
Influenza A virus /H3N2	H3N2/Washington /UR060252/2007	human	H3N2/New York/197/2003
Influenza A virus /H5N1	H5N1/Hong Kong /156/1997	human	H5N1/China/Beijing/01/2003
Influenza A virus /H5N1	H5N1/China /Shenzen/406H/2006	human	H5N1/China/Beijing/01/2003
Influenza A virus /H5N1	H5N1/Indonesia /175H/2005	human	H5N1/China/Beijing/01/2003
Influenza A virus /H5N1	H5N1/Thailand /16/2004	human	H5N1/China/Beijing/01/2003
Influenza A virus /H5N1	H5N1/Vietnam /1194/2004	human	H5N1/China/Beijing/01/2003

5.4 Avian *Influenza A virus* H5N1 Geographic Origin, Results

Table 5.10. Results summary for *Influenza* A *virus* H5N1 origins.

accuracy:	**94%**
word length:	**7**
feature set size:	**3244**

This classification achieved 94% accuracy in assigning all genomes to their correct place of origin. Out of 61 genomes, 57 were accurately assigned to their exact collection region. Instead of selecting training genomes randomly, an attempt was made to represent all outbreak groups in each region resulting in 17 genomes being used as training data. There are no standardized naming conventions for data on the NIAID flu database so an assumption was made that genomes with similar strain identification labels represented samples from related outbreaks, i.e. H5N1/Africa/Afghanistan/1573-92 and H5N1/Africa/Afghanistan/1573-65 were assumed to be from the same group.

The misclassifications in this application present an interesting question regarding this method. All misclassified genomes were assigned to regions directly adjacent to their true collection origins. One genome from Vietnam was incorrectly assigned to China's Guanxi province. The Guanxi province lies north of the Vietnamese border. One genome from Thailand was matched to genomes originating in Vietnam. These two countries are also directly adjacent to each other. Within the Chinese genomes, one genome from the Hunan province was assigned to the Guanxi province while one genome from Shantou was assigned to the Hunan province. Hunan is next to Guanxi, and the Guandong province which contains Shantou is next to the Hunan province. While these errors may represent a fault with the proposed method or an incomplete training genome set, it is also possible that the geographic proximity of true and mistaken regions may play an important role. Some of these samples may have been collected near regional borders and the disease may have been acquired from birds in adjacent region populations. More interestingly, these misclassifications could potentially follow commerce or trade routes for domestic poultry and follow border crossing transmission routes. Tables 5.8 & 5.9

show training genomes, test genomes and classification results. Misclassified genomes in table 5.9 are denoted with bold lettering and an asterisk.

Table 5.11. Avian Influenza A virus H5N1, training data.

Training Data : Avian Influenza A virus H5N1		
Species/Class	**subtype/strain**	**host**
Influenza A virus	H5N1/Africa/Afghanistan/1573-7/2006	chicken
Influenza A virus	H5N1/China/Guangxi/150/2006	duck
Influenza A virus	H5N1/China/Guangxi/1633/2006	goose
Influenza A virus	H5N1/Hong Kong/282/2006	chicken
Influenza A virus	H5N1/China/Hunan/856/2006	duck
Influenza A virus	H5N1/China/Shantou/1233/2006	chicken
Influenza A virus	H5N1/China/Shantou/2086/2006	goose
Influenza A virus	H5N1/Indonesia//CDC25/2005	chicken
Influenza A virus	H5N1/Africa/Ivory Coast/4372-2/2006	turkey
Influenza A virus	H5N1/Africa/Nigeria/1047-30/2006	chicken
Influenza A virus	H5N1/Africa/Nigeria/SO300/2006	chicken
Influenza A virus	H5N1/Africa/Sudan/1784-10/2006	chicken
Influenza A virus	H5N1/Africa/Sudan/2115-10/2006	chicken
Influenza A virus	H5N1/Thailand//39692/2004	chicken
Influenza A virus	H5N1/Thailand/Nontaburi/CK-162/2005	chicken
Influenza A virus	H5N1/Vietnam//10/2004	chicken
Influenza A virus	H5N1/Vietnam//10/2005	chicken

Table 5.12. Avian Influenza A virus H5N1, test data and classification results.

Test Data: Avian *Influenza A virus* H5N1			
Species	**subtype/strain**	**host**	**Classified as**
Influenza A virus	H5N1/Africa/Afghanistan /1573-92/2006	chicken	../Afghanistan/1573-7/2006/chicken
Influenza A virus	H5N1/Africa/Afghanistan /1573-65/2006	chicken	../Afghanistan/1573-7/2006/chicken
Influenza A virus	H5N1/Africa/Afghanistan /1573-47/2006	chicken	../Afghanistan/1573-7/2006/chicken
Influenza A virus	H5N1/China/Guangxi /1458/2006	goose	.../China/Guangxi/150/2006/goose
Influenza A virus	H5N1/China/Guangxi /1898/2006	goose	.../China/Guangxi/150/2006/goose
Influenza A virus	H5N1/China/Guangxi /224/2006	goose	.../China/Guangxi/150/2006/goose
Influenza A virus	**H5N1/China/Guangxi /532/2006**	**goose**	**../China/Hunan/856/2006/duck ***
Influenza A virus	H5N1/China/Guangxi /582/2006	goose	../China/Guangxi/150/2006/duck
Influenza A virus	H5N1/China/Guangxi /288/2006	duck	.../China/Guangxi/150/2006/goose

Influenza A virus	H5N1/China/Guangxi /1830/2006	duck	.../China/Guangxi/150/2006/goose
Influenza A virus	H5N1/China/Guangxi /2143/2006	duck	.../China/Guangxi/150/2006/goose
Influenza A virus	H5N1/China/Guangxi /392/2006	duck	.../China/Guangxi/150/2006/goose
Influenza A virus	H5N1/China/Guangxi /744/2006	duck	.../China/Guangxi/150/2006/goose
Influenza A virus	H5N1/China/Guangxi /804/2006	duck	../China/Guangxi/150/2006/duck
Influenza A virus	**H5N1/China/Guangxi /89/2006**	**duck**	**../Vietnam//10/2005 ***
Influenza A virus	H5N1/Hong Kong /947/2006	chicken	../Hong Kong/282/2006/chicken
Influenza A virus	H5N1/China/Hunan /988/2006	duck	../China/Hunan/856/2006/duck
Influenza A virus	H5N1/China/Hunan /324/2006	duck	../China/Hunan/856/2006/duck
Influenza A virus	H5N1/China/Hunan /344/2006	duck	../China/Hunan/856/2006/duck
Influenza A virus	H5N1/China/Shantou /3295/2006	goose	../China/Shantou/1233/2006/chicken
Influenza A virus	**H5N1/China/Shantou /3265/2006**	**goose**	**../China/Hunan/856/2006/duck ***
Influenza A virus	H5N1/China/Shantou /3840/2006	chicken	../China/Shantou/1233/2006/chicken
Influenza A virus	H5N1/China/Shantou /3923/2006	chicken	../China/Shantou/1233/2006/chicken
Influenza A virus	H5N1/Indonesia /175H/2005	chicken	../Indonesia//CDC25/2005/chicken
Influenza A virus	H5N1/Indonesia /PA/2005	chicken	../Indonesia//CDC25/2005/chicken
Influenza A virus	H5N1/Indonesia /Dairi/BPPVI/2005	chicken	../Indonesia//CDC25/2005/chicken
Influenza A virus	H5N1/Indonesia /Deli Serdang/BPPVI/2005	chicken	../Indonesia//CDC25/2005/chicken
Influenza A virus	H5N1/Indonesia /Gunung Kidal/BPPW/2005	chicken	../Indonesia//CDC25/2005/chicken
Influenza A virus	H5N1/Indonesia /Magetan/BPPW/2005	chicken	../Indonesia//CDC25/2005/chicken
Influenza A virus	H5N1/Indonesia /Parepare/BPPVM/2005	chicken	../Indonesia//CDC25/2005/chicken
Influenza A virus	H5N1/Indonesia /Purworejo/BPPW/2005	chicken	../Indonesia//CDC25/2005/chicken
Influenza A virus	H5N1/Indonesia /Simalanggang/BPPVI/2005	chicken	../Indonesia//CDC25/2005/chicken
Influenza A virus	H5N1/Indonesia /Tarutung/BPPVI/2005	chicken	../Indonesia//CDC25/2005/chicken
Influenza A virus	H5N1/Indonesia /Tebing Tinggi/BPPVI/2005	chicken	../Indonesia//CDC25/2005/chicken
Influenza A virus	H5N1/Africa /Ivory Coast/4372-3/2006	turkey	../Ivory Coast/4372-2/2006/turkey
Influenza A virus	H5N1/Africa /Ivory Coast/4372-4/2006	turkey	../Ivory Coast/4372-2/2006/turkey
Influenza A virus	H5N1/Africa /Nigeria/1047-34/2006	chicken	../Nigeria/1047-30/2006/chicken

Influenza A virus	H5N1/Africa /Nigeria/1047-54/2006	chicken	../Nigeria/1047-30/2006/chicken
Influenza A virus	H5N1/Africa /Nigeria/1047-62/2006	chicken	../Nigeria/1047-30/2006/chicken
Influenza A virus	H5N1/Africa /Nigeria/1047-8/2006	chicken	../Nigeria/1047-30/2006/chicken
Influenza A virus	H5N1/Africa /Nigeria/SO452/2006	chicken	../Nigeria/SO300/2006/chicken
Influenza A virus	H5N1/Africa /Nigeria/SO493/2006	chicken	../Nigeria/SO300/2006/chicken
Influenza A virus	H5N1/Africa /Nigeria/SO494/2006	chicken	../Nigeria/SO300/2006/chicken
Influenza A virus	H5N1/Africa /Sudan/1784-7/2006	chicken	../Sudan/2115-10/2006/chicken
Influenza A virus	H5N1/Africa /Sudan/1784-8/2006	chicken	../Sudan/2115-10/2006/chicken
Influenza A virus	H5N1/Africa /Sudan/2115-12/2006	chicken	../Sudan/2115-10/2006/chicken
Influenza A virus	H5N1/Africa /Sudan/2115-9/2006	chicken	../Sudan/1784-10/2006/chicken
Influenza A virus	H5N1/Thailand /Ayutthaya/CU23/2004	chicken	../Thailand//39692/2004
Influenza A virus	H5N1/Thailand /Kanchanburi/CK-160/2005	chicken	../Thailand/Nontaburi/CK-162/2005
Influenza A virus	H5N1/Thailand /Nakom Patom/CUK2/2004	chicken	../Thailand//39692/2004
Influenza A virus	H5N1/Vietnam /35/2004	chicken	../Vietnam//10/2004/chicken
Influenza A virus	H5N1/Vietnam /36/2004	chicken	../Vietnam//10/2004/chicken
Influenza A virus	H5N1/Vietnam /38/2004	chicken	../Vietnam//10/2004/chicken
Influenza A virus	**H5N1/Vietnam /C57/2004**	**chicken**	**../Thailand//39692/2004 ***
Influenza A virus	H5N1/Vietnam /LD080/2004	chicken	../Vietnam//10/2004/chicken
Influenza A virus	H5N1/Vietnam /TG023/2004	chicken	../Vietnam//10/2004/chicken
Influenza A virus	H5N1/Vietnam /TN025/2004	chicken	../Vietnam//10/2004/chicken
Influenza A virus	H5N1/Vietnam /2/2005	chicken	../Vietnam//10/2004/chicken
Influenza A virus	H5N1/Vietnam /5/2005	chicken	../Vietnam//10/2004/chicken
Influenza A virus	H5N1/Vietnam /8/2005	chicken	../Vietnam//10/2004/chicken
Influenza A virus	H5N1/Vietnam /9/2005	chicken	../Vietnam//10/2004/chicken

5.5 Avian to Human Transmission (H5N1) Region, Results

Table 5.13. Results summary for avian to human transmission (H5N1).

accuracy:	**100%**
word length:	**8**
feature set size:	**9131**

This classification was 100% accurate in matching all human cases of H5N1 to bird cases in their correct countries. Regional accuracy could not be assessed due to a lack of data information. Complete regional information was only available for avian strains in Indonesia and some avian strains in Thailand. All other genomes were only labeled by country. While results can not be verified, this classification indicates that all human H5N1 cases in Indonesia 2005 listed in the NIAID website were acquired in the areas of Magetan, Pare Pare and Gunung Kidal. The training data set contains genomes from all domestic bird H5N1 cases from Indonesia in 2005, Thailand in 2004 and Vietnam in 2004. The test data includes all human H5N1 cases from Indonesia in 2005, Thailand in 2004 and Vietnam in 2004. Training data, test data and classification results are presented in tables 5.11 and 5.12.

Table 5.14. ***Influenza A virus*** **H5N1 avian to human transmission, training data.**

Training Data: *Influenza A virus* H5N1 Avian to Human Transmission		
Species/Class	**subtype/strain**	**host**
Influenza A virus /Indonesia	H5N1/Indonesia/CDC25/2005	chicken
Influenza A virus /Indonesia	H5N1/Indonesia/Dairi/BPPVI/2005	chicken
Influenza A virus /Indonesia	H5N1/Indonesia/Deli Serdang/BPPVI/2005	chicken
Influenza A virus /Indonesia	H5N1/Indonesia/Gunung Kidal/BPPW/2005	chicken
Influenza A virus /Indonesia	H5N1/Indonesia/Magetan/BPPW/2005	chicken
Influenza A virus /Indonesia	H5N1/Indonesia/Parepare/BPPVM/2005	chicken
Influenza A virus /Indonesia	H5N1/Indonesia/Purworejo/BPPW/2005	chicken
Influenza A virus /Indonesia	H5N1/Indonesia/Simalanggang/BPPVI/2005	chicken
Influenza A virus /Indonesia	H5N1/Indonesia/Tarutung/BPPVI/2005	chicken

Influenza A virus /Indonesia	H5N1/Indonesia/Tebing Tinggi/BPPVI/2005	chicken
Influenza A virus /Thailand	H5N1/Thailand//9/2004	chicken
Influenza A virus /Thailand	H5N1/Thailand/Ayutthaya/CU23/2004	chicken
Influenza A virus /Thailand	H5N1/Thailand/Nakom Patom/CUK2/2004	chicken
Influenza A virus /Vietnam	H5N1/Vietnam//10/2004	chicken
Influenza A virus /Vietnam	H5N1/Vietnam//35/2004	chicken
Influenza A virus /Vietnam	H5N1/Vietnam//37/2004	chicken
Influenza A virus /Vietnam	H5N1/Vietnam//38/2004	chicken
Influenza A virus /Vietnam	H5N1/Vietnam//C57/2004	chicken
Influenza A virus /Vietnam	H5N1/Vietnam//LD080/2004	chicken
Influenza A virus /Vietnam	H5N1/Vietnam//TG023/2004	chicken
Influenza A virus /Vietnam	H5N1/Vietnam//TN025/2004	chicken

Table 5.15. Influenza A virus H5N1 avian to human transmission, test data and classification.

Test Data: Avian to Human Transmission			
Species/Class	**subtype/strain**	**host**	**Classified as**
Influenza A virus /Indonesia	H5N1/Indonesia/5/2005	human	../Indonesia/Magetan/BPPW/2005
Influenza A virus /Indonesia	H5N1/Indonesia/7/2005	human	../Indonesia/Magetan/BPPW/2005
Influenza A virus /Indonesia	H5N1/Indonesia/175H/2005	human	../Indonesia/Gunung Kidal/BPPW/2005
Influenza A virus /Indonesia	H5N1/Indonesia/239H/2005	human	../Indonesia/Parepare/BPPVM/2005
Influenza A virus /Indonesia	H5N1/Indonesia/245H/2005	human	../Indonesia/Parepare/BPPVM/2005
Influenza A virus /Indonesia	H5N1/Indonesia/CDC7/2005	human	../Indonesia/Magetan/BPPW/2005
Influenza A virus /Indonesia	H5N1/Indonesia/CDC184/2005	human	../Indonesia/Magetan/BPPW/2005
Influenza A virus /Indonesia	H5N1/Indonesia/CDC287/2005	human	../Indonesia/Parepare/BPPVM/2005
Influenza A virus /Indonesia	H5N1/Indonesia/CDC292T/2005	human	../Indonesia/Parepare/BPPVM/2005
Influenza A virus /Thailand	H5N1/Thailand/1(KAN-1)/2004	human	../Thailand//9/2004
Influenza A virus /Thailand	H5N1/Thailand/2(SP-33)/2004	human	../Thailand//9/2004
Influenza A virus /Thailand	H5N1/Thailand/5(KK-494)/2004	human	../Thailand//9/2004
Influenza A virus /Thailand	H5N1/Thailand/16/2004	human	../Thailand//9/2004
Influenza A virus /Thailand	H5N1/Thailand/SP83/2004	human	../Thailand//9/2004
Influenza A virus /Vietnam	H5N1/Vietnam/1194/2004	human	../Vietnam//TN025/2004
Influenza A virus /Vietnam	H5N1/Vietnam/1203/2004	human	../Vietnam//35/2004
Influenza A virus /Vietnam	H5N1/Vietnam/3062/2004	human	../Vietnam//35/2004
Influenza A virus /Vietnam	H5N1/Vietnam/CL26/2004	human	../Vietnam//35/2004

5.6 Word length vs. Accuracy

To examine how increasing word length in classifications influences accuracy, accuracy vs. word length plots are presented for word length from five to ten for each classification.

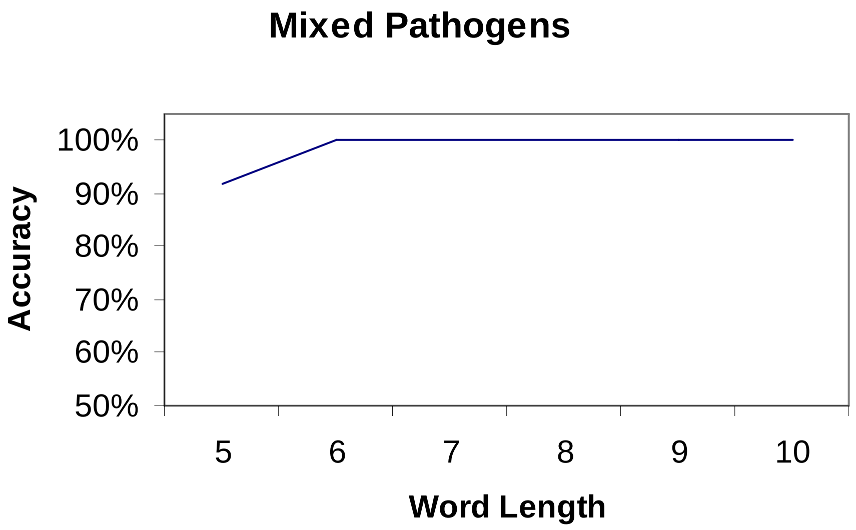

Figure 5.2. Classification Accuracy vs. Word Length for Mixed Pathogens.

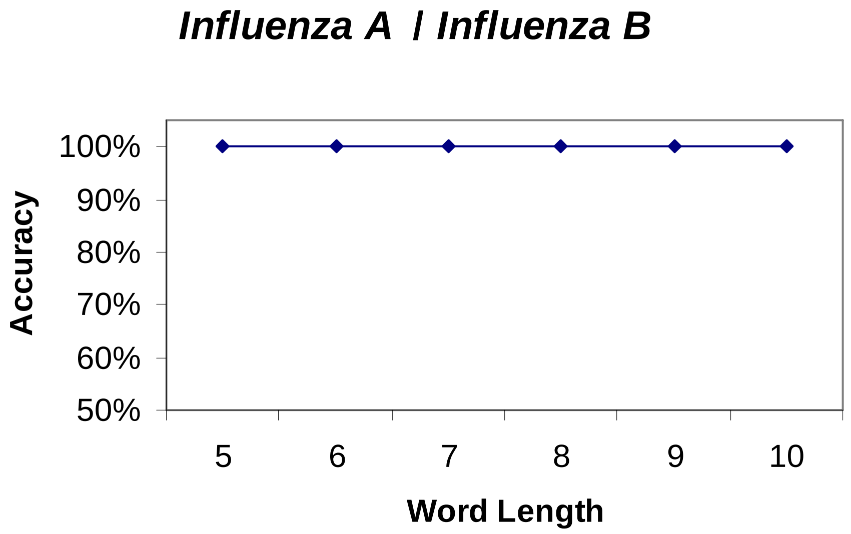

Figure 5.2 Classification Accuracy vs. Word Length for *Influenza A/ Influenza B*.

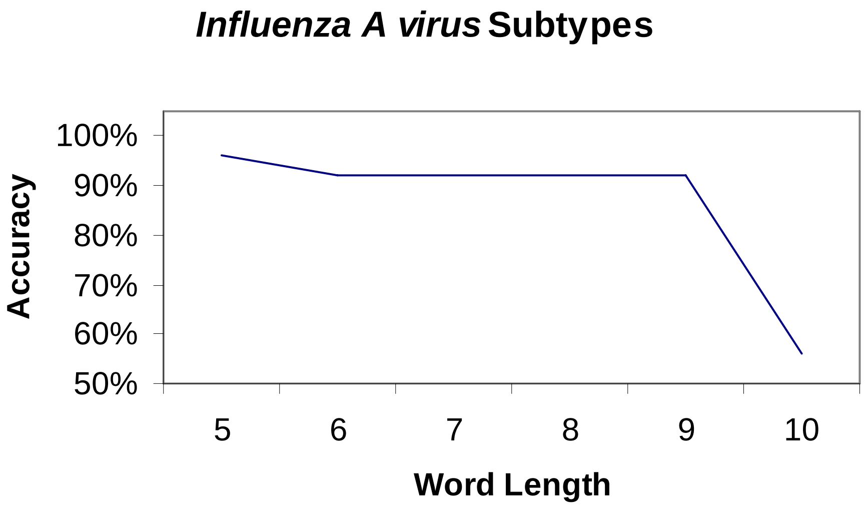

Figure 5.3 Classification Accuracy vs. Word Length for *Influenza A* subtypes.

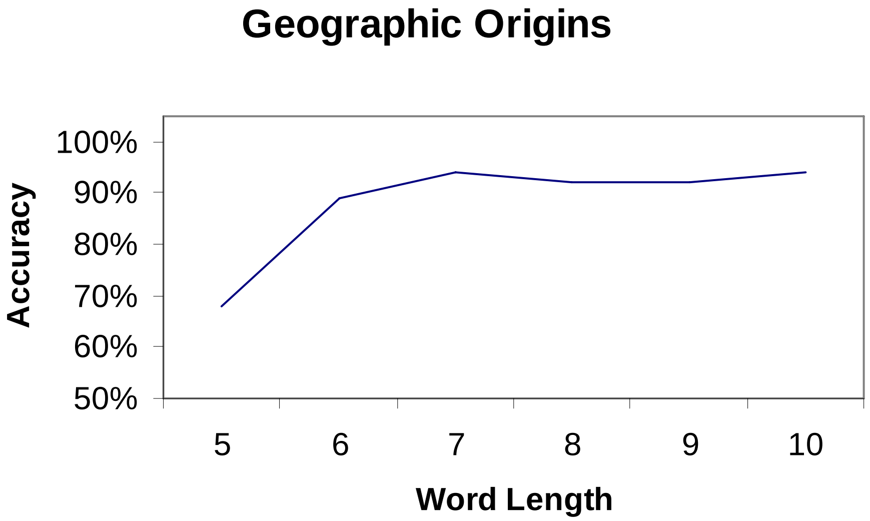

Figure 5.4 Classification Accuracy vs. Word Length for Geographic Origins.

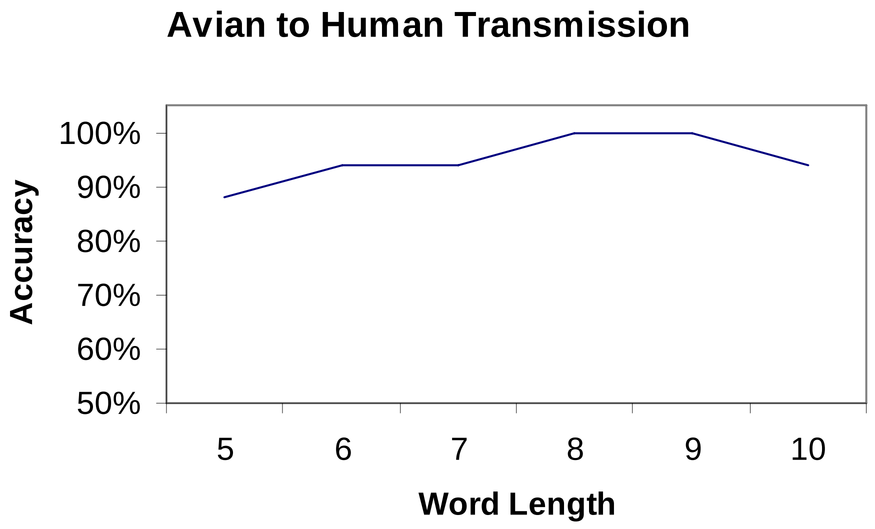

Figure 5.5 Classification Accuracy vs. Word Length for Avian to Human Transmission.

These graphs illustrate that each classification application shows a distinct response to increases in word length. The mixed pathogen classification attains 100% accuracy at word lengths of six or more (Figure 5.1), while the Influenza A/Influenza B data set is 100% accurate at all word lengths (Figure 5.2). The *Influenza A virus* subtypes classification exhibits the highest degree of accuracy, 96% at the shortest word length of five (Figure 5.3) and the lowest accuracy, 56% at the longest word length of ten. The geographic origins data set shows accuracy peaks of 94% at word lengths seven and ten (Figure 5.4). The avian to human transmission classification obtains a peak 100% accuracy at word lengths of eight and nine (Figure 5.5).

Presenting an explanation for these results would be only speculative, but these results illustrate the significance of the length of word used to compare related genomes. They are also suggestive of a relationship between word length and type of relatedness between genomes. For example, the *Influenza A virus* subtypes showed the most marked difference at a word of length five, while the matching of avian to human host viruses (Figure 5.5) required longer word lengths of eight to achieve maximal accuracy.

Chapter 6

Discussion and Future Work

This research suggests that genome classification based on absent words can be accurate at many levels of phylogenetic resolution. Determining the geographic origins of emergent strains may also be enabled. One of the most important applications of a sensitive strain lineage classifier would be in monitoring pandemics and bioterror events. The H5N1 flu subtype is of particular current interest. Many efforts have been made to investigate the transmission of avian H5N1 in Southeast Asia [8, 13, 38, 55, 56], India [47], Europe [5, 51], Africa [49], and worldwide [10, 52]. Genome comparisons in these applications have included whole genome alignments and coding region comparisons. The method described in this thesis may be another approach towards the same goal, but with a different comparison technique which may allow a geographic trace of lineage spreads. Obtaining a detailed history of all current human influenza outbreaks and transmission, particularly of those subtypes presently in circulation may also enable predictions to aid in vaccine design [22] and allow a better understanding of epidemic behavior at the genomic level. Other similar applications of a fine resolution strain classifier could be determining the source of public food contamination such as *E.coli* in grocery produce or forensics and human DNA matching.

Being able to map the evolution of any disease on both spatial and temporal scales could enable unforeseeable insights. To enable such projects, a refinement of the proposed method would most likely be required. This may include a fuzzy instead of discrete classifier with the potential use of word trees to enable multilevel comparisons among genomes. A fuzzy classifier could give an extra measure of how closely test strains are related to the training strains they are assigned to. A DNA word tree could be created as a tertiary tree structure, with each node pointing to an “a”child, “c” child, “g” child, and “t” child. Each overlapping word of length *n* found in a sequence could be inserted into the tree along with its subwords. This would result in a dynamic, linear-time data structure to hold counts and statistics for all words up to length *n*. This would also enable a more precise record of which specific words are avoided in sequences. For example, if a genome maintains a strict bias against the word CTAG, then all words containing CTAG will also be missing, but the bias is still only against CTAG. Experimentation with the degree of word absence across all training strains is also a subject for future work. The current algorithm uses all words that are absent from at least one genome, but this cutoff value could be raised so that words selected must be absent from any given percentage of the training data set. This has the potential of either decreasing classification accuracy, or selecting more biologically significant words and thereby increasing accuracy. Future work will be applied towards improving the sensitivity and accuracy of the described classification method, and towards its ability to highlight significant nucleotide sequences across closely related genomes. Improvements will most likely allow a more precise and informative method of classifying strain lineages, mapping transmission routes and potential probe design.

Bibliography

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

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

[3] Beckstrom-Sternberg SM, Auerbach RK, Godbole S, Pearson JV, Beckstrom-Sternberg JS, Deng Z, Munk C, Kubota K, Zhou Y, Bruce D, Noronha J, Scheuermann RH, Wang A, Wei X, Wang J, Hao J, Wagner DM, Brettin TS, Brown N, Gilna P, Keim PS. (2007). Complete Genomic Characterization of Pathogenic A.II Strain of Francisella tularensis Subspecies tularensis. *PLoS ONE.*:**2**:e947.

[4] Bhagwat, Ashok S. and Michael McClelland. (1992). DNA mismatch correction by Very Short Patch repair may have altered the abundance of oligonucleotides in the E.coli genome. *Nucleic Acids Research*, **20 ;** 1663-1668.

[5] Bragstad, Karoline, Poul H. Jorgensen, Kurt Handberg, Anne S. Hammer, Susanne Kabell and Anders Fomsgaard. (2007). First introduction of highly pathogenic H5N1 avian influenza A viruses in wild and domestic birds in Denmark, Northern Europe. *Virology Journal,* **4:**43.

[6] Burge, Chris, Allan Campbell, Samuel Karlin. (1991). Over- and Under-Representation of Short Oligonucleotides in DNA Sequences. *Proceedings of the National Academy of Sciences of the United States of America*, **89 ;** 1358-1362.

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

[8] Cauthen, Angela N., David E. Swayne, Stacey Schultz-Cherry, Michael l. Perdue and David L. Suarez. (2000). Continued Circulation in China of Highly Pathogenic Avian Influenza Viruses Encoding the Hemagglutinin Gene Associated with the 1997 H5N1 Outbreak in Poultry and Humans. *Journal of Virology*, **74:**6592-6599.

[9] Chain, Patrick, Stefan Kurtz, Enno Ohlebusch, Tom Slezak. (2003).An applications-focused review of comparative genomics tools: Capabilities, limitations and future challenges. *Briefings in Bioinformatics*, **4:**105-128.

[10] Colizza, Vittoria, Alain Barrat, Marc Barthelemy, Alain-Jacques Valleron and Alessandro Vespignani. (2007). Modeling the Worldwide Spread of Pandemic Influenza: Baseline Case and Containment Interventions. *PLos Medicine*, **41**:e(13).

[11] Darrell P. Chandler, Oleg Alferov, Boris Chernov, Don S. Daly, Julia Golova, Alexander Perov, Miroslava Protic, Richard Robison, Matthew Schipma, Amanda White, and Alan Willse. (2006). Diagnostic Oligonucleotide Microarray Fingerprinting of Bacillus Isolates *J Clin Microbiol*, **44:** 244–250.

[12] Doran M, Raicu DS, Furst JD, Settimi R, Schipma M, Chandler DP. (2007). Oligonucleotide microarray identification of Bacillus anthracis strains using support vector machines. *Bioinformatics,* **23**:487-92.

[13] Duan, L., L. Campitelli, X. H. Fan, et. al. (2007). Characterization of Low-Pathogenic H5 Subtype Influenza Viruses from Eurasia: Implications for the Origin of Highly Pathogenic H5N1 Viruses. *Journal of Virology*, **81:**7529-7539.

[14] Eduardo, P., C. Rocha, Alain Viari, and Antoine Danchin. (1998). Oligonucleotide bias in Bacillus Subtilis: general trends and taxonomic comparisons. *Nulceic Acids Research*, **26:** 2971-2980.

[15] Fey, P.D., Major M. P. Dempsey, PhD, M.E. Olson, M.S. Chrustowski, J. L. Engle, MS, J. J. Jay, M. E. Dobson, PhD, K.S. Kalasinsky, PhD, A. A. Shea, P.C. Iwen, PhD, R.C. Wicket, MS, S.C. Francesconi, PhD, R.M. Crawford, PhD, and S.H.Hinrichs, MD. (2007) Molecular Analysis of *Francisella tularensis* Subspecies *tularensis* and *holarctica*. *American Journal of Clinical Pathology*. **128:** 926-935.

[16] Friburg, Markus. (2003). Virus Classification using k-nucleotide Frequencies. Electronic Publication.

[17] Garcia Del Blanco, N., M. E. Dobson, A. I. Vela, V. A. De La Puenta, C. B. Gutierrez, T. L. Hadfield, P. Kuhnert, J. Frey, L. Dominguez, and E. F. Rodriguez Ferri. (2002) Genotyping of Francisella tularensis Strains by Pulsed-Field Gel Electrophoresis, Amplified Fragment Length Polymorphism Fingerprinting, and 16S rRNA Gene Sequencing. *Journal of Clinical Microbiology*. **40:** 2964-2972.

[18] Gelfand, Mikhail and Eugene Koonin. (1997). Avoidance of palindromic words in bacterial and archeal genomes:a close connection with Restriction enzymes. *Nucleic Acids Research*, **25:**2430-39.

[19] Ghedin E, Sengamalay NA, Shumway M, Zaborsky J, Feldblyum T, et al.(2005).Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution. *Nature.* **37:**1162–1166.

[20] Gutman, George A and G. Wesley Hatfield. (1989) Nonrandom utilization of codon pairs in Esherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, **86;** 3699-3703.

[21] Holden MT, et al. (2004). Complete genomes of two clinical Staphylococcus aureus strains: evidence for the rapid evolution of virulence and drug resistance. *Proc Natl Acad Sci* U S A. **101:**9786- 91.

[22] Holmes EC, Ghedin E, Miller N, Taylor J, Bao Y, et al. (2005).Whole-genome analysis of human influenza A virus reveals multiple persistent lineages and reassortment among recent H3N2 viruses. *PLoS Biology*, **3:**e300.

[23] Huard, Richard C., Luiz Claudio de Oliveira Lazzarini, W. Ray Butler, Dick van Soolingen, and John L. Ho.(2003) PCR-Based Method to Differentiate the Subspecies of the Mycobacterium tuberulosis Complex on the Basis of Genomic Deletions. *Journal of Clinical Microbiology*.**41:** 1637-1650.

[24] Huyghe A, Francois P, Charbonnier Y, Tangomo-Bento M, Bonetti EJ, Paster BJ, Bolivar I, Baratti-Mayer D, Pittet D, Schrenzel J; the Geneva Study Group on Noma (GESNOMA). (2008). A Novel Microarray Design Strategy to Study Complex Bacterial Communities. *Applied and environmental microbiology*, [Epub ahead of print]

[25] Jianming, Xie, Sun Xiao, Lu Zhiyan, Xue Weiyan, Dong Xianjun, Lu Zuhong. (2007). gWord: A tool for genome-wide word search and count. In Proceedings of The 1st International Conference on Bioinformatics and Biomedical Engineering, July 6-8, 2007, Wuhan China. http://ieeexplore.ieee.org/iel5/4272484/4272485/04272568.pdf.

[26] Johansson, Anders, Jason Farlow, Par Larson, Meghan Dukerich, Elias Chambers, Mona Bystrom, James Fox, May Chu, Mats Forsman, Anders, sjostedt, and Paul Keim. (2004). Woldwide Genetic Relationships among *Francisella tularensis* Isolates Determined by Multiple_locus Variable-Number Tandem Repeat Analysis. Jourmal of Bacteriology. **186:**5808-5818.

[27] Karlin S, Burge C.(1995) Dinucleotide relative abundance extremes:a genomic signature. *Trends in Genetics* **11:** 283-90.

[28] Keim, P., L.B. Price, A.M. Klevytska, K. L. Smith, J. M. Schupp, R. Okinaka, P. J. Jackson and M.E. Hugh-Jones. (2000). Multiple-Locus Variable_number Tandem Repeat Analysis Reveals Genetic Relationships within *Bacillus anthracis*. **182**:2928-2936.

[29] Killgore G, Thompson A, Johnson S, Brazier J, Kuijper E, Pepin J, Frost EH, Savelkoul P, Nicholson B, van den Berg RJ, Kato H, Sambol SP, Zukowski W, Woods C, Limbago B, Gerding DN, McDonald LC.(2008). Comparison of Seven Techniques for Typing International Epidemic Strains of Clostridium difficile: Restriction Endonuclease Analysis, Pulsed-Field Gel Electrophoresis, PCR-Ribotyping, Multilocus Sequence Typing, Multilocus Variable-Number Tandem-Repeat Analysis, Amplified Fragment Length Polymorphism, and Surface Layer Protein A Gene Sequence Typing *.Journal of Clinical Microbiology*. [Epub ahead of print].

[30] Kong, Huimin, Lee-Fong Lin, Nicole Porter, Shawn Stickel, Devon Byrd, Janos Postfai amd Richard Roberts. (2000). Functional analysis of putative restriction-modification system genes in the *Helicobacter pylori* J99 genome. *Nucleic Acids Research*, **28:** 3215-3223.

[31] Lauer, Kim, I. Llorente, E. Blair, J. Seto, V. Krasnov, A. Purkayastha, S. E. Ditty, T. L. Hadfield, C. Buck, C. Tibbetts and D. Seto. (2004). Natural variation among human adenovirus: genome sequence and annotation of human adenovirus serotype 1. *Journal of Gnereal Virology*. **85:**2615-2625.

[32] Lin Lee-Fong, Janos Postfai, Richard Roberts and Huimin Kong. (2001). Comparative genomics of the Restriction-modification systems in Helicobacter-pylori. *Proceedings of the National Academy of Sciences of the United States of America,* **98:** 2740-2745.

[33] Lowell, Jennnifer L., David M. Wagner, Bakyt Atshabar, Michael F. Antolin, Amy J. Vogler, Paul Keim, May C.Chu, and Kenneth L. Gage. (2005). Identifying Sources of Human Exposure to Plague. *Journal of Clinical Microbiology,* **43:**650-656.

[34] Mazar, Edith, Sarah Lesjean, Anne-Laure Banuis, Michele Gilbert, Veronique Vincent, Bridgitte Gicquel, Michel Tibayrenc, Camille Locht, and Philip Supply. (2001). High-resolution minisatellite-based typing as a portable approach to global analysis of Mycobacterium tuberculosis molecular epidemiology. *Proceedings of the National Academy of Sciences*, **98:** 1901-1906.

[35] McClelland, Michael, Robert Jones, Yogesh Patel and Michael Nelson. (1987). Restriction endonucleases for pulsed field mapping of bacterial genomes. *Nucleic Acids Research*, **15:** 5985-6005.

[36] McHardy, Alice, Hector Martin, Aristotelis Tsirigos, Philip Hugenholtz and Isidore Rigoutsos. (2007). Accurate phylogenetic classification of variable-length DNA fragments. *Nature methods* **4:** 63-72.

[37] Ng, Lai-King and Irene E. Martin BSc. The Laboratory Diagnosis of Neisseria gonorrhoeae (2005). *Canadian Journal of Infectious Diseases and Medical Microbiology***. 16:**15-25.

[38] Nguyen, Doan C., Timothy M. Uyeki, Samdhan Jadhao et al. (2004). Isolation and Characterization of Avian Influenza Viruses, Including Highly Pathogenic H5N1, from Poultry in Live Bird Markets in Hanoi, Vietnam, in 2001. *Jounal of Virology*, **79:** 4201-4212.

[39] Nicoll, A. (2008). Children; Avian Influenza H5N1 and Pandemics. *Archives of disease in childhood* . Jan 11, [Epub ahead of print].

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

[41] Peiris JS, de Jong MD, Guan Y.(2007) Avian influenza virus (H5N1). a threat to human health. *Clinical microbiology reviews*, **20:**243-67.

[42] Perez DR, Sorrell EM, Donis RO.(2005). Avian influenza: an omnipresent pandemic threat. *The Pediatric infectious disease journal*. 24(11 Suppl):S208-16, discussion S215.

[43] Phillippy , AM, Mason JA, Ayanbule K, Sommer DD, Taviani E, et al. (2007). Comprehensive DNA Signature Discovery and Validation. *PLoS Biology*, **3:**e98.

[44] Pourcel, C., F. Andre-Mazeaud, H. Neubauer, F. Ramisse and G. Vergnaud. (2004). Tandem repeats analysis for the high resolution phylogenetic analysis of Yersinia pestis. *BMC Microbiology*, **4:**1471-2180.

[45] Pride, David, Richard Meinersmann, Trudy Wassenaar and Martin Blaser. (2003). Evolutionary Implications of Microbial Genome Tetranucleotide Frequency Biases. *Genome Research*, **13:**145-58

[46] Rainer Merkl, Manfred Kroger, Peter Rice, and Hans-Joachim Fritz. (1992). 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,* **20:**1657- 1662.

[47] Ray K, Potdar VA, Cherian SS, Pawar SD, Jadhav SM, Waregaonkar SR, Joshi AA, Mishra AC. (2008). Characterization of the complete genome of influenza A (H5N1) virus isolated during the 2006 outbreak in poultry in India. *Virus Genes*, Jan 24 [Epub ahead of print].

[48] Rendell, Edward G. (2006). PA Pandemic Influenza Preparedness . Planning Summit 2006.

[49] Teeling, Hanno, Anke Meyerdierks, Maragrete Bauer, Rudolf Amann and Frank Glockner. 2004. Application of tetranucleotide frequencies for the assignment of genomic fragments. *EnvironmentalMicrobiology,* **6:** 938-947.

[50] Thomas, Rebecca, Anders Johansson, Brendan Neeson, Karen Isherwood, Anders Sjostedt, Jill Ellis and Richard Titball. (2002). Discrimination of Human Pathogen Subspecies of *Francisella tularensis* by Using Restriction Fragment Length Polymorphism. *Journal of Clinical Microbiology*. **41:**50-57.

[51] Salzberg, Steven L., Carl Kingsford, et al. (2007). Genome Analysis Linking Recent European and African Influenza(H5N1) Viruses. *Emerging Infectious Diseases*, **13:**713-718.

[52] Sampath R, Russell KL, Massire C, Eshoo MW, HAirpin V, et al. (2007) Global Surveillance of Emerging Influenza Virus Genotypes by Mass Spectrometry. *PLoS ONE* **2:**e489.

[53] van Belkum, A., S. Scherer, L. van Alphen, H. Verbrugh.(1998). Short-sequence DNA repeats in rokaryotic genomes. *Microbiology and molecular biology reviews,* **62:**275-93.

[54] Wang D, Coscoy L, Zylberberg M, Avila PC, Boushey HA, Ganem D, DeRisi JL.(2002). Microarray-based detection and genotyping of viral pathogens. *Proc Natl Acad Sci U S A.*, **99:**15687-92.

[55] Wang J, Vijaykrishna D, Duan L, Bahl J, Zhang JX, Webster RG, Peiris JS, Chen H, Smith GJ, Guan Y.(2008). Identification of the progenitors of Indonesia and Vietnam avian influenza A (H5N1) viruses from southern China. *Journal of Virology*, Jan 23 [Epub ahead of print].

[56] Webster, Robert G., Yi Guan, Malik Peiris et. al. .(2002). Characterization of H5N1 Influenza Viruses That Continue To Circulate in Geese in Southeastern China. Journal of Virology, 76:118-126.

[57] www.pubmed.gov

[58] www.who.int

[59] www.wikipedia.org