COMPUTATIONAL METHODS FOR BACTERIAL CHARACTERIZATION AND BACTERIA-HOST/ENVIRONMENT INTERACTION ANALYSES

A Dissertation

Presented to

The Faculty of the Graduate School

At the University of Missouri

In Partial Fulfillment

Of the Requirements for the Degree

Doctor of Philosophy

By

CHAO ZHANG

Dr. Dong Xu, Dissertation Supervisor

JULY 2012

The undersigned, appointed by the dean of the Graduate School,

have examined the Dissertation entitled

COMPUTATIONAL METHODS FOR BACTERIAL CHARACTERIZATION AND BACTERIA-HOST/ENVIRONMENT INTERACTION ANALYSES

Presented by Chao Zhang

A candidate for the degree of

Doctor of Philosophy

And hereby certify that, in their opinion, it is worthy of acceptance.

Dr. Dong Xu

Dr. Jianlin Cheng

Dr. Dmitry Korkin

Dr. Jianguo Sun

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr. Dong Xu for giving me an opportunity to work with him, who is an experienced researcher and has always been supportive through the years of my study. With his guidance, I have finished the project successfully and learned how to do research and write papers. Dr. Dong Xu is a great mentor in life as well.

I would like to thank Dr. Jianlin Cheng, Dr. Dmitry Korkin, and Dr. Jianguo Sun who are willing to serve on my committee. I acknowledge Dr. Jianlin Cheng for his valuable discussions on solving machine learning issue. I acknowledge Dr. Dmitry Korkin for the advices on bacteria-host interactions. I acknowledge Dr. Jianguo Sun for his suggestions on the statistic model design.

I would like to thank the members in Digital Biology Laboratory, University of Missouri: Dr. Nick Lin, Dr. Jianjiong Gao, Dr. Jingfen Zhang, Zhiquan He and Qiuming Yao for their effective work and support in the collaboration. I also appreciate the great friendship we have developed through these years.

I would express my heartfelt thanks to my parents for their encouragement and support during my entire course of study.

The last but not the least, this dissertation is dedicated to my wife, Huiheng Wen, who is not only my soul mate in life, but also a great help for my doctoral studies.

Table of Contents

LIST OF FIGURES

Figure Page

LIST OF TABLES

1. INTRODUCTION

1.1 Motivation

As the largest domain of all living organisms on earth, bacteria are estimated to have more than five nonillion(10^{30}) individuals worldwide [1], which are far more than previous estimations of the total number of bacteria [2]. These single-cell organisms can be found everywhere, e.g., deep sea, hot springs, human gut, and even in radioactive waste [3]. Due to close connections between bacteria and human life, we cannot live without them and actually benefit from the microorganisms in many cases, e.g. food production, human health [4], environmental sciences [5], and chemical industry [6, 7]. On the other hand, pathogenic bacteria are one of the most serious threats to human life. For example, tuberculosis, the most common fatal bacterial disease, kills about 2 million people every year [8]. Since 1676, when Antonie van Leeuwenhoek first observed bacteria, scientists have never stopped exploring the micro-world. The task of identification and classification of bacteria remains challenging because bacteria are invisible to naked eyes and cannot be easily differentiated morphologically. During the past two decades, DNA sequencing technologies have become a powerful tool for scientists to take up the challenge.

In 1995, when John Craig Venter just started to sequence the first bacterial genome – *Haemophilus influenza* [9], DNA sequencing was extremely difficult and time consuming. The common thought at the time was that it would be sufficient to build a gene pool of the whole microbial community from just a few dozen representative genomes. Today, thanks to new sequencing technologies, more than 1600 microbial whole genome sequences have been released and many more bacterial genome-sequencing projects are ongoing [10]. With the accumulation of bacterial genomic data, the focus of microbial genomics (study of genomes of microorganisms including archaea, bacteria and fungi) is shifting from single genome to pan-genome (gene pool of a particular species) and meta-genome (environmental gene/species pool). However, the explosion of data has not answered all the questions of researchers in this field. It becomes evident that these data just revealed a tip of the iceberg for the bacterial world. In-depth analysis of these data is needed to help better understand the genome diversity and dynamics of bacteria, interactions between bacteria and their hosts/environments, and the pathogenicity of pathogens. Meanwhile, the unprecedented amount of genome data also poses major challenges for computational analysis, which is an essential tool for microbial genomics. In fact, computational methods for massive genomic sequence analysis have become a bottleneck of microbial genomics.

1.2 About this dissertation

In this dissertation, we will focus on computational methods for discovering the interactions between bacteria and hosts/environments and bacterial characterization (i.e. identification and classification), based on sequencing data with consideration of bacteria's hosts and environments. While this topic has been brought up in recent publications [11-16], no in-depth review has been presented. Bacterial identification through detecting variations of genome sequences across different species/genus is a very important and essential step of analyzing genomic data, especially for metagenomic data. Thus, in chapter 2, we first review existing computational tools and their limitations for bacterial identification. As bacteria evolve rapidly in response to the environments, bacterial adaptations to different environments/hosts will reflect in their genome sequences. Many bacteria, even belonging to the same species, still show extensive genomic plasticity and diverse pathogenicity. For example, three different *E. coli* strains, laboratory strains *E. coli* MG1655, enterohemorrhagic *E. coli* EDL933, and an uropathogenic strain *E. coli* CFT073-), share only 39.2% common genes [17].Thus, chapter 3 of this dissertation, we will assess the practical computational methods for detecting the sequence variations of bacteria in different environments for a given species. In chapter 4, we will dissect the evolutionary dynamics of bacterial virulence and review the methods for identification of genetic markers in bacterial DNA sequences that are associated with a disease or host. In chapter 5, based on our observations and works in chapter 4, we predict some novel effectors for those known pathogens. The last chapter is the summary of this dissertation.

2. REVIEW OF GENERAL MICROBIAL IDENTIFICATION

2.1 Microbial identification

In the past, analysis of microbial communities was a complicated task due to their high diversity and inaccessibility via culturing. The emerging next-generation sequencing technologies provide a potential way for doing this analysis on a routine basis [12]. The Human Microbiome Project [18], which began in May 2007, aimed to survey the microbial communities that colonize the human body. Currently, over 100 similar metagenomic projects are ongoing, covering microbial communities of skin and several tracts, including gastrointestine, genitourinary tract, oral cavity, nasopharynx, and respiratory tract [12]. These studies will undoubtedly provide new insight into many aspects of complex microbial communities, such as metabolic capabilities of microorganisms, co-evolution of bacteria and host, communication of microbial cells and so on [16]. Although metagenomics is still in its early stage, this emerging field has already discovered many surprises in microbial genomics and microbiology [16]. Among the extensive genomic sequencing data of microbial communities generated by various metagenomic projects, approximately 62% of the bacteria that can be identified from the human intestine were previously unknown and 80% of them are not cultivatable [19]. Due to the explosion of metagenomic data, DNA sequence-based identification and classification are becoming more and more important in exploring microbial diversity.

In the 1970s, DNA-DNA hybridization was introduced to differentiate bacterial species. Any two bacterial strains with more than 70% DNA-DNA hybridization were considered to be the same species. Later, with the development of new sequencing techniques, Carl Woese pioneered other criteria for bacterial identification[20]. For example, the 16S ribosomal RNA (rRNA) gene is highly conserved in bacteria and archaea, and was used for identification and discovery of pathogens starting from 1990 [21]. Nowadays, 16S rRNA gene is also widely used for phylogenetic studies [22]. However, due to the limitations of 16S rRNA gene, other genetic markers have been employed for bacterial identification, e.g. multilocus sequence typing (MLST) [23]. Recently, whole genome-based methods have been developed for bacterial identification. Despite these advances, complete genome sequence is not easily obtained [24, 25].

The 16S rRNA gene, a molecular clock, has a relatively slow evolutionary rate of 1% sequence divergence per 50 million years. It is around 1500 nucleotides in length and contains 9 hypervariable regions [26] (Figure 1) as well as conserved regions interspersed with the variable ones. In terms of similarity of 16S rRNA gene sequences, bacteria within the same genus and species usually share about 95% and 97% pairwise sequence identities, respectively [27]. Because of the consistency of sequences in bacteria, 16S rRNA gene sequencing has become the gold standard for characterization of bacterial communities.

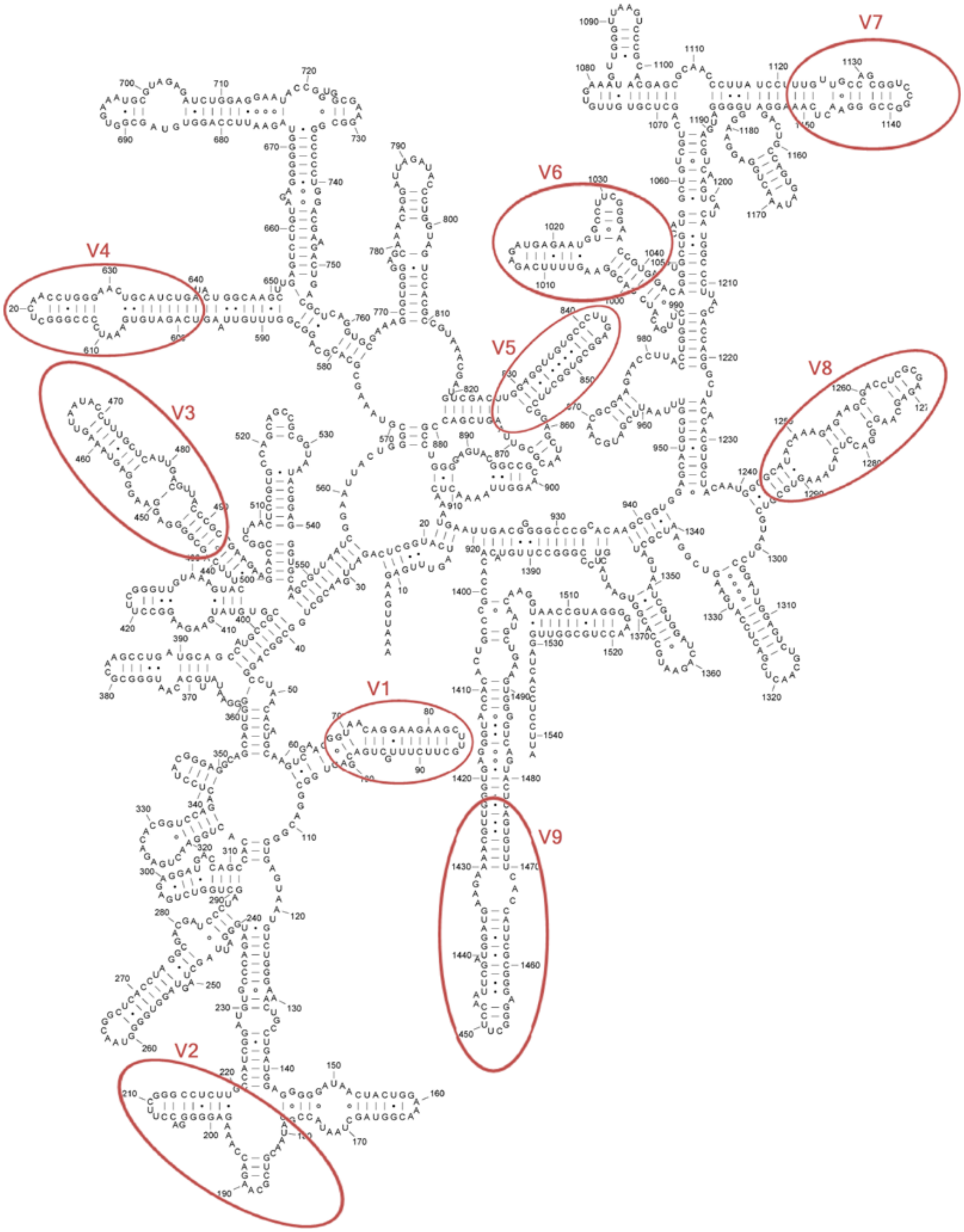

Figure 1. The secondary structures of 16S rRNA gene of *Escherichia coli*. Generated by using XRNA (**http://rna.ucsc.edu/rnacenter/xrna/xrna.html**) with 9 hypervariable regions circled.

2.2 Common factors affecting bacterial identification and classification

Bacterial identification is based on a specific taxonomic scheme. There are several taxonomic schemes proposed by independent curators, e.g., the Ribosomal Database Project (RDP) (Bergey's) [28], Norman Pace [29], Wolfgang Ludwig [30], Phil Hugenholtz [31], and the National Center for Biotechnology Information (NCBI). All major rRNA sequence databases, such as RDP (http://rdp.cme.msu.edu/) [32], Greengenes (http://greengenes.lbl.gov) [33], and ARB-SILVA (http://www.arb-silva.de) [34] were designed based on different taxonomic schemes. The variations among different taxonomic schemes have a direct impact on the identification results. For example, there are 31 phyla in the RDP database, 50 in Ludwig's taxonomy, 68 in NCBI and 88 in the system proposed by Pace and Hugenholtz. Within each phylum, the number of sub-groupings also varies. After 2005, the oldest and most traditional bacterial classification system - Bergey's taxonomy – started to build taxonomy based on analyses of nucleotide sequences of ribosomal small subunit RNA rather than on phenotypic data [28]. Nevertheless, most classification systems are still based on structural and functional attributes of bacteria. Thus, 16S rRNA gene-based identification results may never match those taxonomies exactly.

Sequence alignment is a necessary step in 16S rRNA gene-based identification. Besides the multiple sequence alignment programs such as ClustalW [35], MEGA [36], NAST [37] and MUSCLE [38], some databases also include alignment programs, such as RDP II, Greengenes, and ARB-SILVA. It is shown that alignment quality has a significant impact on sequence classification [39]. Incorporating the well-determined secondary structures of 16S rRNA gene with the pairwise or multiple sequence alignment will improve alignment quality [32], but the extra information will also significantly increase the computational complexity. Another recent research reveals that the longest totally conserved segment in 16S rRNA gene across all bacteria is only 11 bps and in most regions the longest absolutely conserved stretches are only 4 bps [12, 40]. This stark reality is a challenge for developing effective and accurate alignment algorithms, especially for those 16S rRNA gene fragments with less than half of their full lengths.

Different hypervariable (V) regions show different efficacies in identifying species, and no single hypervariable region can differentiate bacterial species among all bacteria. At the genus level, using 2-region set for identifications has become a standard approach, and about 90% of bacterial strains successfully identified by this approach cannot be identified through biochemical (phenotype) methods [41]. Chakravorty *et al.* proposed that the V2 and V3 regions were most suitable for universal genus identification of pathogenic bacteria [42]. The V5-V6 region set was reported to be the most useful in study of human oral microbiome [43, 44]. It is suggested that analyzing three different 2-region sets (V2–V3, V4–V5, and V6–V8) in parallel was effective in determining the bacterial consortia in maize rhizospheres [45]. Some studies also revealed that the V6-V9 set [46], especially the V6 region [47, 48], represented an outlier and might not be suitable to use directly for taxonomic assignment. Therefore, the choice of hypervariable regions is critical for bacterial identification [41]. There is room for further computational algorithm development in designing an optimal hypervariable set for bacterial identification.

Due to their highly conserved nature, 16S rRNA gene sequences might not be a good genetic marker to distinguish the sub-populations within a species. Even different species within the same genus, such as *Bacillus cereus*, *B. thuringiensis* and *B. anthracis* [49], have only a few bases different in their 16S rRNA gene sequences. No matter what computational methods are used, there will be a theoretical upper limit of the average accuracy for species identification across all species.

2.3 Major computational methods and their limitations

Generally speaking, computational methods for bacterial identification can be divided into two major categories: homology- and composition-based [50, 51] as summarized in Figure 2. Homology-based approaches use traditional sequence alignment algorithms to compare sequences similarity. According to the techniques of alignment, it can be further divided into two subgroups, i.e., sequence search (especially using BLAST) and phylogeny. Composition-based methods build models based on the different features extracted from sequences, e.g., GC content [52], codon usage, and frequencies of motifs. The typical classifiers used in composition-based methods are naïve Bayes classifier, Markov model, and support vector machine (SVM).

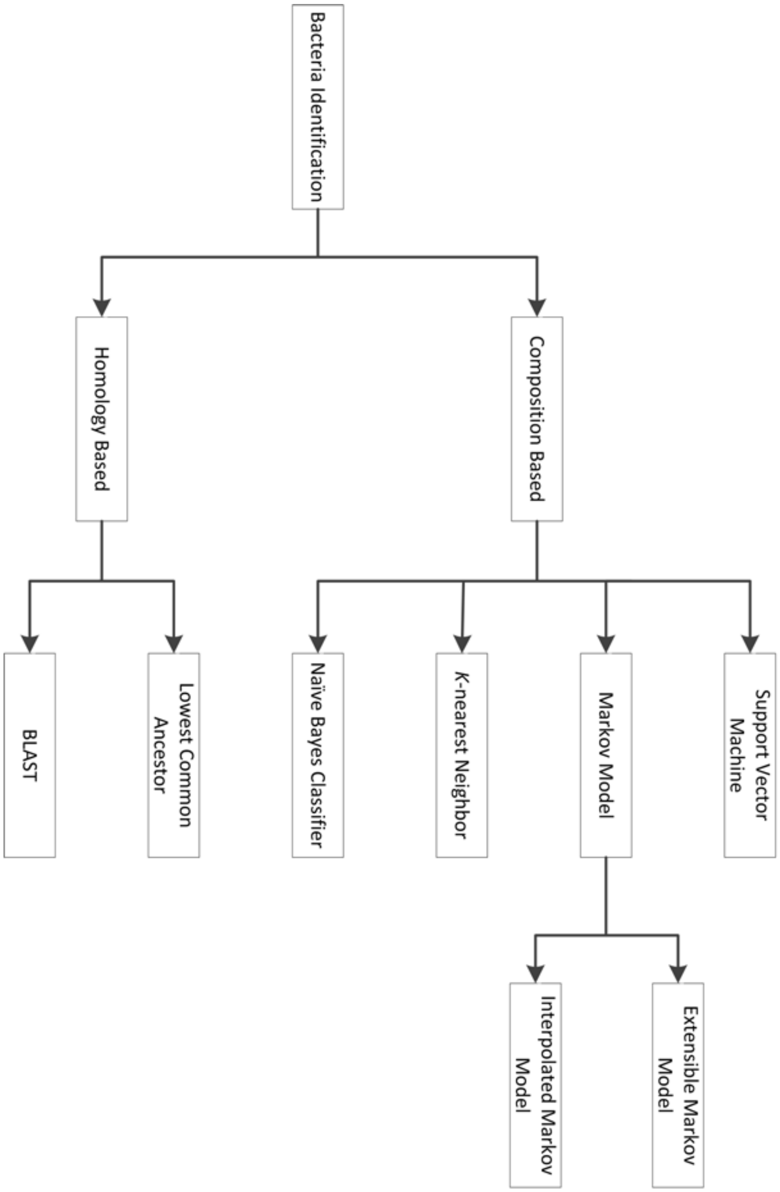

Figure 2. Major algorithms used in bacterial identification.

BLAST: Basic Local Alignment Search Tool (BLAST) [53] is one of the most popular bioinformatics programs. It is most often used for comparing biological sequences, such as searching a query sequence against a sequence library. Thus, it naturally became the first choice of metagenomic studies in the early stage and has been shown to be effective in many studies [54-56]. Due to the limitation of its algorithm, the closest BLAST hit may not be the nearest neighbor [57] and this approach can reach a high-accuracy level only when the query reads have significant similarities to the matches in the sequence library [13]. Since the lengths of reads generated by next-generation sequencing technologies are still not long enough, short-reads are generally not unique and often cause ambiguous identification results. Recently, some researchers started to evaluate the performance of BLAST for analyzing metagenomic data [58, 59]. For some metagenomics datasets, the significant BLAST hits only accounted for 35% of the reads in the sample [54]. With the improvement of sequencing techniques, the length of reads are getting longer, and the reference genome libraries are becoming more comprehensive [18]. Extremely expensive computational complexity is another common drawback of alignment-based identification techniques. While BLAST is an efficient software tool, its capacity in handling of metagenomic data can barely satisfy the needs of current analyses. With the explosive increase of metagenomic data, further reducing the computational complexity becomes an important challenge of alignment-based identification methods.

Phylogeny: Because a significant proportion of short query reads hit more than one species with significant E-value in the BLAST, a simple algorithm, the Lowest Common Ancestor (LCA), has been employed to assign the ambiguous reads to the right taxa [47, 60]. Instead of choosing the nearest neighbor, LCA assigns each reads to the ancestor taxa by computing means of the least common taxonomic ancestor of a suitable set of hits, and it can also reflect the level of conservation of the sequence. While this approach is more sophisticated than BLAST, it has two drawbacks [50, 61]: first, LCA has a relatively low coverage, because for some reads with very few numbers of hits on the reference taxonomy, the least common taxon cannot be computed; second, many reads have been assigned to non-informative high taxonomic ranks. The first issue has been addressed by a modified method – multiple taxonomic ranks (MTR) [50]. Traditionally, LCA only uses local taxonomic information for matching reference sequences and treats each read independently. MTR proposes a two-step method to use global type of information: 1) clustering reads with the same taxon; and 2) selecting the 'best' subset of each cluster with a combinatorial optimization algorithm for LCA. The results of MTR experiments show a significant increase in coverage compared to the traditional LCA. The second drawback of LCA has been tackled by Clemente *et al.*[62]. By evaluating the number of mismatches between the read and the reference taxonomy to balance the relevance of precision and recall in the assignment, Clemente's method assigns each read to the inner nodes (a rank lower or equal) of the taxon selected by the standard LCA.

Naïve Bayes classifier: In order to avoid the heavy computational expense, some composition-based methods have been proposed as alternatives to classic alignment for sequence comparison [63]. A typical method is naïve Bayes classifier (NBC) [51, 64]. In 1997, Wang *et al.* [64] developed a NBC (RDP Classifier) with 8-mers (8 consecutive nucleotides) for using 16s rRNA gene sequences to classify bacteria into new taxonomy which has become one of the most popular classifiers in microbiology. As an extremely conserved gene, 16S rRNA gene has a much slower evolutional rate than other genes, and partial 16s rRNA gene sequence has a different *k*-mer distribution to full-length 16s rRNA gene sequence. With incomplete 16s rRNA gene sequence, the accuracy of bacterial identification may drop dramatically. Due to the limitations of the method, RDP classifier only can provide taxonomic assignments from domain to the genus level, and it also needs users to provide full length of 16s rRNA gene sequence to obtain high classification accuracy. It does not work at either species level or sub-species level. One study [51] suggests that NBC works better on whole genome sequences than 16S rRNA gene sequence. The same study also tried to increase the length *k* to 15 to achieve better performance on short reads. When *k* equals 15, there are about 1 billion possible words and the longest bacterial genome is only around 26 million nucleotides; so an increase of k to 15 might cause the counting statistics insignificant. Furthermore, computational and storage expenses can be a concern.

Other models: Signal processing and machine-learning approaches are widely used to solve problems with the background noise, clutter, and jamming signals, and they also have been applied for bacterial identification. Phymm [65], a classifier based on interpolated Markov model (IMM), has been trained on 539 curated genomes. It constructs probability distributions representing observed patterns of nucleotides on chromosomes or plasmids. Phymm shows good performance at ranks Class and Phylum levels on metagenomic datasets with relatively long reads (800 bp and 1000 bp), but low accuracy for short reads (100 bp) [50]. Recently, an extensible Markov model (EMM) [66] was proposed to use a time-varying Markov chain model for bacterial identification. The sequence data can be considered as states representing clusters of similar sequence segments and inter-state transition probabilities representing the implicit order within the sequences. This model outperformed the RDP classifier, but still did not show satisfactory accuracy at rank Species. PhyloPythia [67], a multiclass SVM based approach, examines oligonucleotide composition to characterize taxonomic groups. This method is effective for genomic fragments of 3000 bp and longer, but for 1000 bp sequences, its sensitivity drops drastically.

2.4 Challenges and future work

As an indispensable step, most bacterial identification tools have been integrated into the metagenomic analysis systems. The drawbacks of the current metagenomic analysis systems are also the drawbacks of bacterial identification tools. Near half of the current metagenomic analysis software tools (Table 1) uses a 'pipeline' approach. Within a pipeline, a set of applications is connected in a sequential order and the output of one application becomes the input of the next application. As a double-edged sword, pipeline methods can significantly reduce the cost of time and labor of development process by using existing, stable and well-established applications. However, the pipeline approach usually does not have an efficient structure for a system to handle large data sets, which is the case for metagenomic data. Furthermore, at each step of the pipeline, some analysis results and resources are subject to re-computation and re-allocation. This redundancy definitely affects the efficacy of using computational resources, hence decreasing the performance of the system. Another common problem of the pipeline approach is that the input/output within the pipeline could be time-consuming and error-prone. Thus, a cohesive public open-source development platform, such as Cytoscape [68], is in dire need of construction. Such a platform will not only significantly save the development time of individual researchers, but also speed up the potentially revolutionary improvement of this field. A similar open-source framework is necessary for bacterial identification and it will help this research area to rapidly improve.

A rapid growth in high-performance computing power is timely for analyzing dramatic rise in data volume. Different models of parallel computing, such as distributed computing, general-purpose computing on graphics processing units (GPUs), and cloud computing can be applied as bioinformatics tools to analyze these data. Open-source new bioinformatics software tools are being developed by exploiting web-based services to increase computing power provided by academic and commercial "cloud computing networks." Some resources are already available, e.g. Science Clouds (http://workspace.globus.org/clouds), which allows researchers to have full control over using a leasing model. MG-RAST-CLOUD (http://metagenomics.anl.gov) is a metagenomics analysis server with capability of handling data from Gigabytes to Terabytes. CloudBurst [80] is a highly sensitive genome sequence mapping tool by using cloud computing. Soon, access to the Internet plus a pad or smartphone will be the only requirements for large-scale bioinformatics analysis. High-performance computing also makes it possible to implement algorithms with high computational complexities. Due to the size of large data, current bacterial identification systems tend to use simple algorithms with low computational complexity. Some of the computationally expensive

Table 1. Metagenomic analysis software

Name and reference	Type	Open source	Algorithm	URL	Last update
OTUbase [69]	R package	Yes	BLAST	http://www.bioconductor.org/packages/release/bioc/html/OTUbase.html	2011
CAMERA [70]	Webserver	No	BLAST	http://camera.calit2.net	2011
MG-RAST [71]	Pipeline/web	Yes	BLAST	http://metagenomics.anl.gov/	2011
WebCARMA [72]	Pipeline/web	Yes	BLAST	http://webcarma.cebitec.uni-bielefeld.de	2011
PANGEA [73]	Pipeline	Yes	BLAST	http://pangea-16s.sourceforge.net	2011
MARTA [74]	Pipeline	Yes	BLAST	http://bergelson.uchicago.edu/software/marta	2010
BIBI [75]	Webserver	No	BLAST	http://umr5558-sud-str1.univ-lyon1.fr/lebibi/lebibi.cgi	2010
QIIME [76]	Pipeline	Yes	BLAST/NBC	http://qiime.sourceforge.net/	2010
STAP [77]	Pipeline	Yes	BLAST	http://bobcat.genomecenter.ucdavis.edu/STAP/	2008
MEGAN [60]	Pipeline	No	LCA	http://ab.inf.uni-tuebingen.de/software/megan/	2011
Galaxy [78]	Pipeline	No	LCA	http://galaxy.psu.edu/	2011
MTR [50]	Executables	Yes	LCA	http://www.cs.ru.nl/~gori/software/MTR.tar.gz	2010
TANGO [61]	Perl script	Yes	LCA	http://www.lsi.upc.edu/~valiente/tango/	2010
NBC [79]	Webserver	No	NBC	http://nbc.ece.drexel.edu/	2011
RDP [32]	Pipeline	No	NBC	http://rdp.cme.msu.edu/	2011
Phymm [65]	Executables	Yes	Markov	http://www.cbcb.umd.edu/software/phymm/	2011
EMM [66]	Executables	Yes	Markov	http://lyle.smu.edu/IDA/EMM/	2010
PhyloPythia [67]	Webserver	No	SVM	http://cbcsrv.watson.ibm.com/phylopythia.html	2007

algorithms are explored with high-performance computing. Supervised and unsupervised learning methods, e.g. language models, linear classifiers, and advanced Bayesian techniques are promising for bacterial identification with high accuracies. Another promising approach to improving identification accuracy is using mixed models or a meta-analysis technique to combine the identification results from different methods. For example, PhymmBL [65], a hybrid classifier, outperforms both BLAST and Phymm on the same dataset.

Although it is still in the early stage, metagenomics analysis has already been used in many research areas, e.g. clinical microbiology [81-83], bacteria-environment symbioses [84, 85], and host-microbial interactions [55, 86, 87]. Most of those applications are still using the two most traditional identification approaches - BLAST and the RDP classifier, since the newly developed methods still have some limitations and cannot significantly outperform them. Although all new algorithms are trying to overcome the common drawbacks, some issues remain unsolved. Generally, homology-based approaches work well for long reads (>800 bp), while composition-based approaches can handle relatively short reads and partial gene sequence (down to 100 bp for some datasets). No single algorithm can dominate the identification results across both cases and the performance will significantly drop with the decrease of the read length. Therefore, improving performance on bacterial identification with short reads (less than 400 bp) is still an open problem.

Because of its unique characteristics, 16s rRNA gene remains as the most commonly used genetic marker [88]. However, using partial 16S rRNA gene sequence for bacterial identification is more difficult than using whole genome sequence or some other genetic markers, since the correlation of the sequence patterns between different hypervariable regions of 16S rRNA gene is relatively low and the variations of different hypervariable regions are species-specific. Thus, selection of hypervariable regions with a specific underlying database is tricky in bacterial identification and classification as it can significantly affect the identification results. To date, no matter what computational method is used, highly confident bacterial identification can only be achieved at rank of Genus, but many microbiology issues require higher resolution approaches to differentiate bacteria at the species level or even at sub-species level. For such differentiation, whole genome sequence-based and MLST [23] methods are the two approaches currently available. MLST is based on the partial sequences of seven housekeeping genes with around 450 bp each, but its resolution power is still limited by the little sequence variation among some bacterial species. Another possible approach is the use of single-nucleotide polymorphisms (SNPs) as genetic markers. This approach was originally developed for diagnosis of human genetic diseases, and now it has been used for the analyses of bacterial genomes [89, 90]. When multiple potential markers are available, selecting sequence markers for a classifier is even more challenging than developing a general classifier with a given marker. Until now there was no universal protocol for solving this problem.

3. HOST TRACKING BY USING INTERVENING SEQUENCES OF FAECALIBACTERIUM 16S RDNA

3.1 Introduction

Bacteria can mutate and adapt to the changing environments. Studies on bacteria-host/environment interactions not only provide an opportunity to dissect the genetic basis of adaptive evolution, but also can be very useful on infectious disease prevention and environment-quality monitoring. Host- or host group-specific bacterial identification is an important step in studying bacteria-host interactions [91]. Unlike the general bacterial identification methods that we discussed in chapter 2, high identification accuracy at the species or sub-species level is necessary for this type of identification. Here, we use identification of fecal source in aquatic environments as an example to introduce a practical application of computational methods in identification of host-specific bacteria.

Microbiological quality of water poses a risk to human health. During 2005-2006, 78 waterborne-disease outbreaks were reported, which caused the sickness of 4,412 people, 116 hospitalizations, and five deaths in the United States [92]. These recent outbreaks have highlighted the importance of microbiological water quality. Animal manures are the major cause for the impaired water quality. Animal gastrointestinal (G.I.) tract maintains a rich microbial community with specific mutualistic associations with different hosts [93]. Thus, the bacterial community in the G.I. system not only is used to model the evolutionary relationships between hosts and bacteria, but also provides a reliable indicator in identification of the fecal pollution source in aquatic environments. Current regulations created to assure the microbiological quality of water are based on the numbers of fecal indicator bacteria (FIB), *Escherichia coli* (*E. coli*) or *enterococci*. The presence of FIB indicates fecal contamination of water, but it does not identify the source of pollution. To overcome this limitation, microbial source tracking (MST) methods have been developed [94-96]. It is essential to identify the sources of fecal pollution before best management practices can be applied to eliminate or mitigate the pollution sources. A variety of alternative fecal indicator microorganisms have been proposed and used in MST with varying degrees of success. Bacteria such as *Bacteroides-Prevotella* spp. [97], *Bifidobacterium* spp. [98], *Clostridium perfringens* [99], *Lactobacillus* spp. [100], *Methanogens* spp. [101], and *Faecalibacterium* [102] have been proposed and used for MST. If the pollution source is identified and located, the source can then be eliminated or mitigated to improve quality of water.

Methods applied to MST can be divided into two types. Library-dependent methods, such as antibiotic-resistance profiling [103], ribotyping [104], and rep-PCR [105], require construction of a library or database composed of phenotypic or genotypic fingerprints of FIB isolated from known fecal sources, that is a database or library of patterns constructed using isolates from known faecal sources. The other MST type is host-specific and library-independent. Both MST methods are based on detection of intestinal microorganisms or their biomolecules exclusively found in a particular host species. Library-independent methods are popular as they are more rapid and less costly than library-dependent MST methods [106]. Both types of MST methods would not be indicative of current fecal pollution if the organisms they target can persist or grow in the environment; this is the case for the standard FIB, *E. coli*, and *enterococci* [107]. Most of the reported host-specific MST have been done using alternative fecal indicator microorganisms and has been thoroughly reviewed [106, 108, 109]. Currently there is no single MST that is able to identify the source of fecal pollution with absolute certainty. Using a combination of methods offers more reliability and validity to determine the fecal source [110-112].

Human fecal pollution poses the highest risk to human health, but livestock waste can contribute to spread of zoonotic pathogens causing environmental contamination. It is estimated that over 99% of animal manure production is from three agriculturally important animals**:** swine, cattle, and poultry. Poultry litter, often used as fertilizer, can contain important water-/food-borne pathogens including *Campylobacter jejuni, Listeria monocytogenes, and Salmonella enterica,* presenting a significant fecal pollution source to surface waters [113]. Poultry fecal pollution in water can be caused by improper manure application on cropland, intentional pumping of manure onto the ground, malfunction or overflow of manure storage, uncontrolled runoff from feedlots or operations and intentional breeches of storage lagoons.

Recently, genetic markers associated with poultry feces or litter have been identified and are potentially useful for tracking poultry fecal pollution in environment. [91] used a metagenomic approach to identify chicken-specific fecal microbial DNA sequences, which mainly resulted in *Bacteroidetes* genes. [114] identified a genetic marker specific for the 16S rRNA gene of a *Brevibacterium* spp. for poultry litter. [115]

used mitochondrial DNA of epithelial cells of poultry intestinal tracts to track the poultry fecal pollution in water. The prevalence of these genetic markers in poultry feces varies from 6% to 75%, except for that of the poultry mitochondrial DNA marker, which is 100% in theory. However, the amount of poultry mitochondrial DNA is significantly lower than bacterial DNA in poultry feces.

We and other researchers found *Faecalibacterium* to be among the dominant bacteria in the intestinal tract of major animals that are often found to be the sources of fecal pollution in water, which makes this bacterium a candidate as an alternative fecal indicator. *Faecalibacterium* is the newly established genus [116], composed of a single species *Faecalibacterium prausnitzii* [28] with the type strain being *F. prausnitzii* ATCC27768. *F. prausnitzii*, previously *Fusobacterium prausnitzii*, is phylogenetically distinct from known *Fusobacterium* species, based on the 16S rDNA sequence and G+C content [116]. *Faecalibacterium* is the dominating fecal bacterium in humans [102, 117], cattle [118], swine [119], and poultry [120].

3.2 Methods and materials

3.2.1 Data sources

A collection of 7,470 sequences of Faecalibacterium 16S rDNA associated with intestinal and fecal samples from different animal species was obtained from the public Ribosomal Database Project 10 (RDP 10, http://rdp.cme.msu.edu/). The host species include human (6,419 sequences), cattle (811 sequences), turkey (132 sequences), chicken (88 sequences), pig (16 sequences), dog (3 sequence) and sheep (1 sequence).

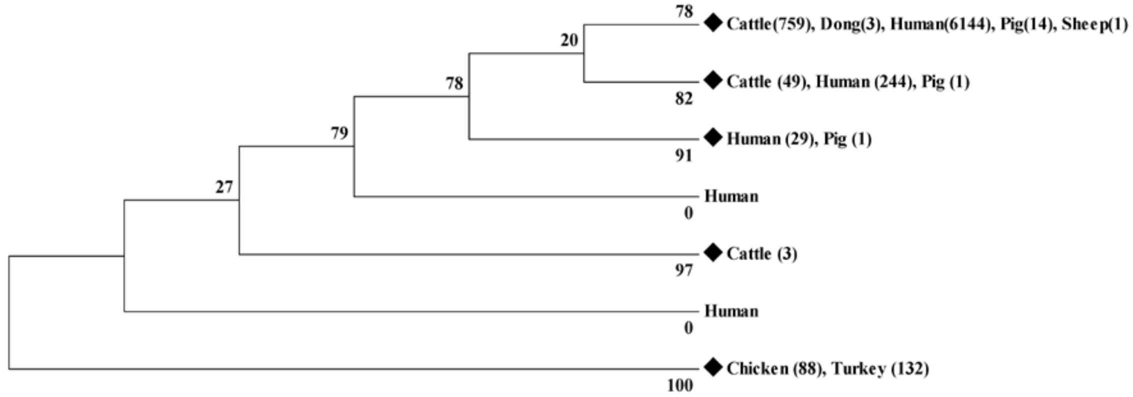
78
Cattle(759), Dong(3), Human(6144), Pig(14), Sheep(1)
20
Cattle (49), Human (244), Pig (1)
82
78
Human (29), Pig (1)
91
79
Human
0
27
Cattle (3)
97
Human
0
Chicken (88), Turkey (132)
100

Figure 3. Sequences were aligned and an approximately maximum likelihood tree was created using FastTree2. The tree was visualized via MEGA 5. The values along branches indicate the evolution distance. The numbers in parentheses indicate numbers of sequences.

3.2.2 Phylogenetic analyses

Multiple sequence alignment (MSA) was applied to all collected sequences by using the MUSCLE [38]. Based on the multiple alignments, a phylogenetic tree was derived from an approximately maximum likelihood analysis by using FastTree 2 [121] and the tree is visualized using the Mega 5 [36], as shown in Figure 3. The topology of the poultry host branch is highly conserved, and the sequences from poultry host have higher similarity than those sequences from any other animal species.

3.2.3 Entropy analyses

Based on the results of Phylogenetic analyses, we speculated that the 16S rDNA sequences from poultry host might carry some unique segments which can be used to distinguish stains from poultry and other hosts. In order to detect the exact locations of sequence markers, the aligned sequences were then divided into poultry and non-poultry groups according to their hosts. For the sequences of two host groups, both combinatorial entropy (Eq. 1) [122] and background entropy (Eq. 2) were calculated for each site of the sequences and described as follows:

$$C_i = \sum_k \ln \frac{N_k!}{\prod_{\alpha=1...20} N_{\alpha,i,k}!} \tag{1}$$

where N_k represents the number of sequences in group *k*; $N_{\alpha,i,k}$ denotes the number of nucleotides of type α in the column *i* of group *k*; $N_{\alpha,i}$ is the number of nucleotides of type α in the column *i*; N represents the total number of sequences in alignment.

$$B_i = \sum_k \ln \frac{N_k!}{\prod_{\alpha=1...20} \tilde{N}_{\alpha,i,k}!} \tag{2}$$

where $\tilde{N}_{\alpha,i,k} = N_k N_{\alpha,i} / N$.

Then the entropy difference of any two host-group sequences was measured as previously reported [123]. Three extreme cases are defined as in Figure 4. In case P1, the nucleotides are 'randomly and uniformly distributed' over all groups and there is no significantly conserved pattern for this position. Case P2 represents a 'globally conserved' pattern and all the nucleotides are the same across both groups. In case P3, some specific nucleotides are only conserved in particular groups, and different groups have different nucleotides. We call this case ‘locally conserved’. According to the calculation results of the entropy difference for the three cases, the entropy difference is 0, 0 and the minimum value for the ‘randomly and uniformly distributed’ case, ‘globally conserved’ case, and ‘locally conserved’ case, respectively. Hence, the entropy difference is a proper measurement for detecting a ‘locally conserved’ sequence pattern. According to the above illustration, we chose entropy difference as a feature to differentiate the two groups. The entropy differences of selected positions are used as the feature entropy in the identification step for distinguishing the host groups with the same species of bacteria.

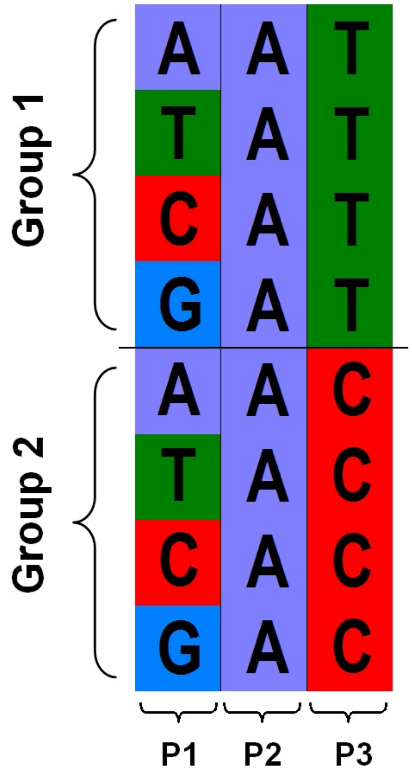

Figure 4. An example to present the different cases for the entropy calculation.

3.2.4 Sequences similarity and polymorphism analyses

Based entropy calculation, V1 region and its extended segment have been detected as the sequence markers to identify strains from poultry. The extended segment of V1 region has been called intervening sequence (IVS) and we found the proportion of gaps in IVS region is very high. Due to the incompleteness of data, it hardly for us to determine those are the real gaps or the unsequenced part of 16S rDNA. According to our selection criteria, we take V1 region as our first choice.

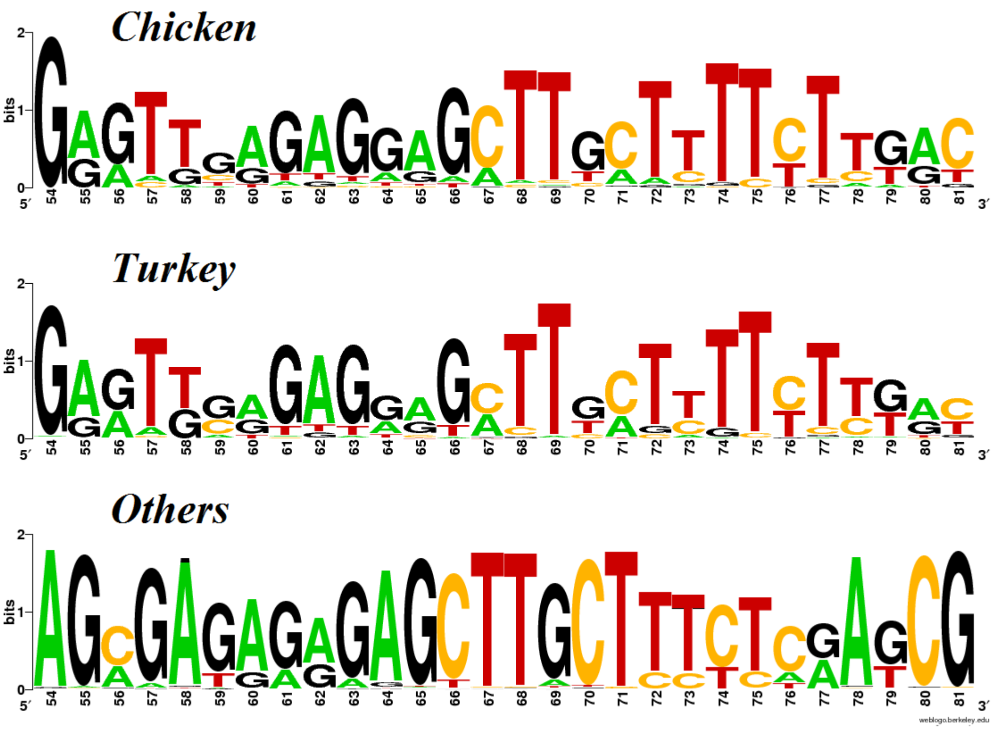

Figure 5. Sequence logos of the 16s rRNA gene V1 regions of *Faecalibacterium* from chicken, turkey and other host species. The overall height of the stack indicates the

sequence conservation at that position, while the height of symbols within the stack indicates the relative frequency of each nucleic acid at that position.

We have analyzed the polymorphism of *Faecalibacterium* 16s rDNA sequences which have the highest variations, so it supports that V1 is the region where signature sequences of a particular host may be found. In fact, there is a significant difference in the nucleotide distributions in V1 between species with poultry (including chicken and turkey) and others as hosts (Figure 5). No significant difference has been found for all other variable regions between chicken and turkey hosts, and in any variable region among all the other hosts, including human, cattle, pig, dog, and sheep. We trimmed the V1 region in all 7470 sequences by splitting them into three groups according to the different hosts, 'chicken', 'turkey' and 'others'. First, the average sequence similarity within each group was calculated and the pairwise sequence similarity defined as follows:

$$identity = \frac{number\ of\ identical\ nucleotides}{length\ of\ the\ alignment} \times 100\% \qquad (3)$$

In Table 2, within the same host group of 'chicken', 'turkey' or 'others', the V1 regions share 65.2%, 60.7% and 76.7% average pairwise sequence identities, respectively. Then we compared the average sequence identities between all groups. The 'chicken' group and the 'others' group only share 30.8% average identity, which is very close to the 30.2% identity between 'turkey' and 'others', while the 'chicken' group and 'turkey' group are very similar with 62.4% average identity.

Table 2. Sequence identities of the 16s rRNA gene V1 regions of *Faecalibacterium* between different host groups.

	Chicken	Turkey	Others
Chicken	65.2%	62.4%	30.8%
Turkey		60.7%	30.2%
Others			76.7%

The polymorphism of poultry vs. non-poultry *Faecalibacterium* 16S rDNA sequences, including the significant difference in V1 nucleotide distribution, has provided a foundation for identifying poultry host with using computational methods. We applied several simple learning methods on the sequences of V1 region and all of them can reach 100% accuracy to detect poultry host. However, no "signature sequence" of V1 region was identified to be used for design and development of a poultry feces-specific polymerase chain reaction (PCR) assay for the rapid determination of poultry fecal pollution in water. Thus we moved our focus to the IVS region from V1 region. Interestingly, the alignment comparison identified four types of insertion sequences in IVS region of some poultry Faecalibacterium 16S rDNA (Table 3). All the four sequences can form stem-loop structures, as demonstrated by using the RNA/DNA folding program (http://kinefold.curie.fr/), a characteristic of IVSs defined by Evguenieva-Hackenberg [124]. The presence of IVS in rDNA is not common but widespread in bacteria. The occurrence of IVS is more commonly found in 23S rDNA than in 16S rDNA of bacteria [125, 126].

Table 3. The IVSs and IVS -containing sequences

IVS	IVS sequence (length in bp)	GenBank ID number of the retrieved sequence containing the IVS	Host	Sample site
IVS-1	5′-gagtgatttt tctactccga gccttttgca gcgtcaatca atgcgaagca ttgatttagg cttatttagt aagctgacac atgcggatgg ttgggagtag aaaaatcgct -3′ (110 bp)	DQ456040.1	Turkey	Iowa, USA
		AF376209.1 and AF376211.1	Chicken	Delaware, USA
IVS-2	5′-gaaagatttt tctactccga gttcttcgcg ggtctttaag gagagcgtcg atcaatgcga agcatcgaag atgcgagcat tgatccaggc tttatttaga agactaacac aaaggtggag cagagagctg ggagtaggaa aatctttt -3′ (148 bp)	DQ456203.1, DQ456153.1, DQ342331.1, DQ456303.1, DQ456188.1, DQ456182.1, DQ456167.1, DQ456161.1, and DQ342331.1	Turkey	Iowa, USA
		AF376206.1	Chicken	Jiangsu, China
		JF781724.1 and JF781902.1	Chicken	Kentucky, USA
IVS-3	5′-tcgatcaatg cgaagcatcg aagatgcgag cattgatcca ggctttattt agaagactaa cacaaaggtg gagcagagag ctgggagtag gaaaatcttt t-3′ (110 bp)	DQ342331.1, DQ456459.1, DQ456303.1, DQ456209.1, DQ456203.1, DQ456188.1, DQ456182.1, DQ456167.1, DQ456161.1, and DQ456153.1	Turkey	Iowa, USA
		AF376206.1	Chicken	Jiangsu, China
		JF781902.1 and JF781724.1	Chicken	Kentucky, USA
IVS-4	5′-ggaagatttt tctactccgg gttctttgct tggctttaaa agagcgtcaa tcaatgcgga gcattgattc aggcttttta aagaagacta acacagagat ggagcggaga gctgggagta ggaaaatctt tt-3′ (132 bp)	EU009819.1	Turkey	Iowa, USA

Table 4. The prevalence and host specificity of IVSs

Sample source	Sample location*	No. of sample sites	No. of samples tested	No. of positive samples (%)			
				PCR-p1 (IVS-1)		PCR-p2 (IVS-2)	
				1 ng reaction^{-1}	10 ng reaction^{-1}	1 ng reaction^{-1}	10 ng reaction^{-1}
Feces							
Chicken	MO	1 farm	24	15 (62.5)	18 (75)	16 (66.7)	24 (100)
Turkey	MO	2 farms	28	22 (78.5)	22 (78.5)	9 (32.1)	23 (82.1)
Beef cattle	MO	3 farms	26	0	0	0	0
Dairy cattle	MO	2 farms	26	0	0	0	0
Dog	MO	4 locations	32	0	0	0	0
Goose	MO	5 locations	21	0	0	0	0
Horse	MO	3 farms	30	0	1 (3.3)	0	0
Human	MO	direct collection	32	0	0	0	0
Sheep	MO	1 farm	26	0	1 (3.8)	0	0
Swine	MO	3 farms	32	0	0	0	0
Wastewater							
Chicken lagoon	MO	5 farms	10	ND	ND	5 (50)	10 (100)
Turkey lagoon	MO	5 farms	10	ND	ND	3 (30)	10 (100)
Cattle lagoon	MO	3 farms	3	ND	ND	0	0
Cattle lagoon	MS	2 farms	2	ND	ND	0	0
Swine lagoon	MO	5 farms	5	ND	ND	0	0
Swine lagoon	MS	5 farms	5	ND	ND	0	0
Sewage	MO	4 treatment plants	20	ND	ND	0	0
Sewage	MS	9 treatment plants	12	ND	ND	0	0

Table 5. The PCR assays, targets, and primers

PCR assay	Target	Forward primer	Reverse primer	Annealing Temperature (°C)	Amplicon size (bp)	Reference
PCR-p1	IVS-1	FaCH-F1: 5’-tactccgagccttttgc-3’	FaCH-R1: 5’-gcgatttttctactccca-3’	55	97	This study
PCR-p2	IVS-2	FaCH-F2: 5’-tactccgagttcttcgcg-3’	FaCH-R2: 5’-gattttcctactcccagc-3’	55	132	This study
PCR-p3	IVS-3	FaCH-F3: 5’-tcgatcaatgcgaagcatcgaa-3’	FaCH-R3: 5’-aagattttcctactcccagctctctg-3	60	100	This study
PCR-p4	IVS-4	FaCH-F4: 5’-actccgggttctttgcttggct-3’	FaCH-R4: 5’-actcccagctctccgctcca-3’	60	106	This study
Control PCR	16S rDNA	Bac1070F: 5’-atggctgtcgtcagct-3’	Bac1392R: 5'-gacgggcggtgtgta-3’	45	323	Ferris et al. 1996

3.2.5 Experimental validation

Twenty-four chickens, 28 turkeys, and 225 non-poultry fecal samples representing eight animal species were collected in Missouri (Table 4). Human stool specimens were obtained from healthy adult donors. Dog fecal samples were collected from private house pets. Farm animal and goose samples were collected as certainly as possible from separate individuals. Sewage was collected from inflow to waste treatment plants and sewage lines. Sixty-seven wastewater samples were collected from locations in Missouri and Mississippi (Table 4). All samples were stored at -80°C before use and DNA was extracted as previously described [102], using the PowerSoil® DNA extraction kit (MoBio Laboratories, Carlsbad, CA).

PCR primers used in this study are listed in Table 5 and were designed based on the sequences containing newly identified genetic markers, using the Primer-BLAST program (http://www.ncbi.nlm.nih.gov/tools/primer-blast/). Specificity of each potential PCR primer set was examined via corresponding PCR amplification, using pooled fecal DNA for each host type listed in Table 5. Each of the composite samples contains equal amount of fecal DNA extracted from 20 fecal samples collected from 20 individual animals. Ten ng of composite DNA was used for each PCR reaction. The PCR assays were performed using 40 cycles with the following thermocycle profile: initial denaturation at 95°C for 4 minutes; denaturation at 94°C for 1 minute; annealing at the annealing temperature (Table 5) for 1 minute; elongation at 72°C for 30 seconds and final elongation at 72°C for 7 minutes. Total PCR reaction volume was 50 µl. PCR products were separated by electrophoresis in a 2% agarose gel.

In order to test the specificity of the newly identified, the designed PCR assays has been validated by using 52 fecal samples from the target (poultry) species, 225 from the non-target animal/human species, and 67 wastewater samples, as detailed in the Table 4. One ng and 10 ng of total DNA extracted from each sample were used as the template to run the PCR. PCR conditions were the same as those previously mentioned. PCR reaction, using the fecal DNA sample, confirmed to contain the target DNA sequences, which served as the positive control. PCR reaction without DNA template served as the negative control. All PCR assays were repeated at least in duplicate. To exclude potential false negative results due to the presence of PCR inhibitors in the samples, all DNA samples were tested using the Control PCR assay (Table 5) before use.

3.3 Results and discussions

3.3.1 Phylogenetically uniqueness of poultry Faecalibacterium 16S rDNA sequences and discovery of IVSs

All Faecalibacterium 16S rDNA sequences associated with poultry are phylogenetically unique and clustered in one single branch of the phylogentic tree (Figure 3). IVSs have been discovered in some strains only from poultry and they show the highly conserved patterns. Reportedly, an IVS of 23S rDNA can be unique to a bacterial species and has been used for detection of *Edwardsiella ictaluri,* a member of Enterobacteriaceae [127]. The function of ribosomal IVSs remains unknown, but their formation may be the result of bacterial adaption to a close working relationship with their host species [40]. In other words, a ribosomal IVS is potentially host specific. This leads to a new discovery which could add a new dimension to DNA-based MST methods leading to a better understanding of ribosomal IVs—their function and their effect on host species. If this is the case, ribosomal IVSs can be used as genetic markers for the development of new DNA-based MST methods.

3.3.2 The newly identified IVSs are highly associated with poultry and unique to genus *Faecalibacterium*

Using the four newly identified IVSs to run a BLAST search against the NCBI database (http://www.ncbi.nlm.nih.gov/blast/Blast.cgi) retrieved DNA sequences that contain these IVSs (Table 3). All the retrieved sequences are 16S rDNAs of uncultured bacteria from chicken or turkey intestinal or fecal samples and belong to genus *Faecalibacterium* of phylum *Firmicutes,* as classified by the Classifier program of RDP 10 (http://rdp.cme.msu.edu/classifier/ classifier.jsp). Result of this *in silico* analysis suggests that the newly identified IVSs are highly associated with poultry and unique to genus *Faecalibacterium***.**

To examine the host specificity and prevalence of the four IVSs, four PCR primer sets (Table 5) were designed based on the IVSs’ sequences (Table 3). Two PCR assays, PCR-p1 and PCR-p2, were developed, targeting at IVS-1 and -2 respectively. Based on the PCR-p1 assay, the IVS-1 could be detected in 75 % of the chicken and 78.5% of the turkey fecal samples, at the level of 10 ng fecal DNA per PCR reaction; by contrast, most of the non-poultry fecal samples were negative with no PCR-p1 amplification except for one horse and sheep fecal sample tested; in a similar test using PCR-p2 assay, the IVS-2 could be detected in 100% of the chicken and about 82.1% of the turkey fecal samples, but not in any non-poultry fecal samples (Table 4). Poultry-feces specificity of the IVS-2 was further tested using 67 wastewater samples, where PCR-p2 was positive for all chicken and turkey lagoon samples, but negative for all other types of wastewater samples (Table 4). As a PCR quality control, the Control PCR (Table 5) assay was used in this study and generated the expected size of amplicon from all the fecal samples tested, excluding the possibility of false negative results due to significant PCR inhibitors that might be present in the fecal and wastewater DNA samples. DNA sequencing analysis of the amplicons of PCR-p1 and PCR-p2 confirmed that all the amplicons contained the expected IVS.

Thus, the prevalence of the IVS-1 and IVS-2 in poultry feces is higher than any previously reported poultry feces-specific genetic markers, where the prevalence varies from 6% to 75% [91, 114]. Although neither IVS-1 nor IVS-2 can distinguish chicken from turkey fecal materials, both can differentiate feces of chicken or turkey from that of geese (Table 4). Further experiments will be needed to determine if IVS-1 and -2 can distinguish feces of chicken or turkey from that of other wild birds. Interestingly, IVS-2 appeared in feces of poultry reared in China and the United States, as demonstrated by previous (Table 3) and this study (Table 4), suggesting that this genetic marker may have a “global” distribution, a highly desired characteristic for use in MST. In conclusion, the IVS-2 may be a useful genetic marker for identification of poultry feces.

3.3.3 The potential use of *Faecalibacterium* and ribosomal IVSs in MST

Bacteria of genera *Bacteroides* and *Faecalibacterium* are known to be dominant in feces tracts of animals that are the major sources of fecal pollution in water [118, 128]; they are both host-associated and obligate anaerobes, capable of little or no multiplication in the environment. Despite the similarity between *Bacteroides* bacteria and *Faecalibacterium* bacteria, much of the research has been focused on use *Bacteroides* bacteria as the alternative FIB [108, 109]. This research attempts to explore the potential of *Faecalibacterium* bacteria as a new alternative FIB for MST. Bacteria of *Faecalibacterium* have been found to be among the dominant genus of fecal bacteria at least in intestinal tract of humans[128], cattle, swine, and poultry; the host specificity of *Faecalibacterium* has been demonstrated previously by [102] and this study.

16S rDNA of gut bacteria is an important MST genetic marker. Many PCR-based MST methods have been developed based on 16S rDNA sequences [106, 108, 109], but cross-reactivity may occur since 16S rDNA sequences are highly conserved. As the alternatives, intergenic spacer region of 16S-23S rDNA [111, 129] and genes directly involved in host-microbe interactions [130, 131] have been proposed to be used in MST. Our study demonstrates that some ribosomal IVSs may be useful as MST genetic markers.

4. RISK ASSESSMENT AND SEQUENCE MARKERS DETECTION OF GASTRIC CANCER

4.1 Introduction

Immediately after birth, humans undergo a life-long process of colonization by foreign microorganisms. Although we benefit from some host-bacterial associations, bacterial pathogens have long been known to play important roles in the development of different diseases [132] including cancer [133]. The host-bacteria interactions include many complicated mechanisms, such as co-evolution, the response of the host immune system [134], the adaption of bacteria to the host and so on. There are challenges in discovering associations between bacteria and diseases. For example: given the same host and same bacterial species, why will different subspecies or strains cause different diseases and how can one differentiate the virulence by bacterial sequences? Although many publications have discussed the roles of bacterial pathogens in the development of diseases, a standard computational method for detecting disease-related sequence markers and identifying virulent strains is still lacking. Genus *Helicobacter* is a well-studied model for its relationship between bacterial infection and cancer [135]. Here, we are using *Helicobacter pylori* (*H. pylori*) as an example to introduce a method for identification of disease-specific bacteria.

H. pylori is a Gram-negative helix-shaped bacterium inhabiting the human stomach and infecting more than half of the world's population [136-138]. Recent studies have shown that it is associated with gastroduodenal diseases, including duodenal ulcers

[139], gastric ulcers [140] and chronic gastritis. More importantly, it is a significant risk factor for developing gastric cancer [141-143]. It has been classified as a Class 1 human carcinogen by the World Health Organization since 1994 [136].

As a marker of *H. pylori*, the Cytotoxin-associated gene A (cagA) has been revealed by further analysis to be the major virulence factor. *H. pylori* strains carrying the cagA gene increase the risk factor of gastroduodenal diseases by three folds over cagA-negative strains [141, 144, 145]. CagA, which is encoded by the cagA gene, is a 125-140 kDa protein. It contains 1142-1320 amino acids and has a variable region at the C-terminal region in which various short sequences (such as EPIYA motif) repeat 1-7 times. After *H. pylori* colonizing on the surface of the gastric epithelium, CagA can be translocated into the gastric epithelial cell through a type IV secretion system. Once injected into the host cell, CagA localizes to the plasma membrane and can be phosphorylated by Src-family tyrosine kinases on the specific tyrosine residues of a five-amino-acid (EPIYA) motif [146-149]. Tyrosine-phosphorylated CagA then binds specifically to SHP-2 tyrosine phosphatase [146, 150] to activate a phosphorylase, which causes the cascade effect that interferes with the signal transduction pathway of the host cell, leading to a restructuring of the host cell cytoskeleton and formation of hummingbird phenotype [146, 151]. At the same time through activating mitogen-activated protein kinase (MAPK), extracellular signal-regulated kinase (ERK) [152] and focal adhesion kinase (FAK), CagA also can cause cell dissociation and infiltrative tumor growth [153-156]. Such a process makes CagA a most important virulence factor in *H. pylori* [157].

Within the variable region of CagA, there are some different intervening sequences between those EPIYA motifs. One copy of EPIYA plus intervening sequence is identified as an EPIYA segment. Four unique types of EPIYA segments have been found in CagA, defined as EPIYA-A, -B, -C and -D [146]. The CagA isolated from East Asian countries, designated as East Asian CagA, contains EPIYA-A, EPIYA-B and EPIYA-D motifs. The CagA from Western countries, EPIYA-D, is replaced by EPIYA-C. Stronger phosphorylation motif binding activity of the EPIYA-D motif leads to greater morphological changes than what the EPIYA-C motif can cause in infected cells [146]. It is this EPIYA-D motif's increased binding activity and resultant morphological changes that identifies it as a potential factor to explain the higher incidence of gastric cancer in East Asian countries [158, 159].

Previous studies revealed a variation in the number of EPIYA motif repeats for both East Asian and Western CagA, which can affect biological activities. Yamaoka et al. [160] found that in Columbia and USA, the ability of cagA-positive *H. pylori* to cause gastric mucosal atrophy and intestinal metaplasia might be related to the number of EPIYA motifs in the CagA strain. Argent et al. [151] came to the same conclusion later. However, contrary opinions were published by Lai et al. [161] based on findings of no relationship between the number of EPIYA motifs in the CagA strain and clinical disease within 58 isolates from Taiwan. Considering the size and geographic limitation of these studies, the validity of this conclusion is questionable. Aside from the number of the EPIYA motif repeats, the sequence difference of strains in variable regions also could cause a significant difference of virulence, which might relate to the different pathogenic abilities of *H. pylori* [162].

Because of the complex and variant sequences in CagA, the relationships between the polymorphism of CagA and clinical diseases become a very interesting research problem. However, the molecular mechanisms that underlie different gastroduodenal diseases caused by cagA-positive *H. pylori* infection remain unknown. Until now most studies are still limited to the discovery or evaluation of the correlation between the number of CagA EPIYA motifs and diseases [163].

In this chapter, we propose a systematic method to analyze not only the number of EPIYA motifs in CagA sequences but also the specific sequence patterns of intervening regions. First, we introduce entropy calculation to detect the residues within the variable region of CagA as the gastric cancer biomarkers. Then we employ a supervised learning procedure to classify cancer and non-cancer by using the information of detected residues in CagA as the features. We choose support vector machines (SVM) as a binary classifier and compare our method with others. Our approach not only proves our hypothesis that the sequence of variable region of CagA contains information to distinguish different diseases, but also provides a useful tool to predict the correlation between the novel CagA strains and diseases and to detect the biomarker as well.

4.2 Methods and materials

4.2.1 Data sources

We searched the National Center for Biotechnology Information (NCBI), the Swiss-prot/Tremble and DDBJ protein database and obtained 535 strains of *H. pylori* CagA protein. Among them, there are 287 East Asian subtype strains and 248 Western subtype strains. In the East Asian subtype group, 47 out of 287 strains are from gastric cancer patients and the rest are from other diseases. In the Western subtype group, there are 37 strains from the gastric cancer patients, and the remainders are from other diseases or the normal controls, including 24 strains from volunteers whose health (disease) status was unknown.

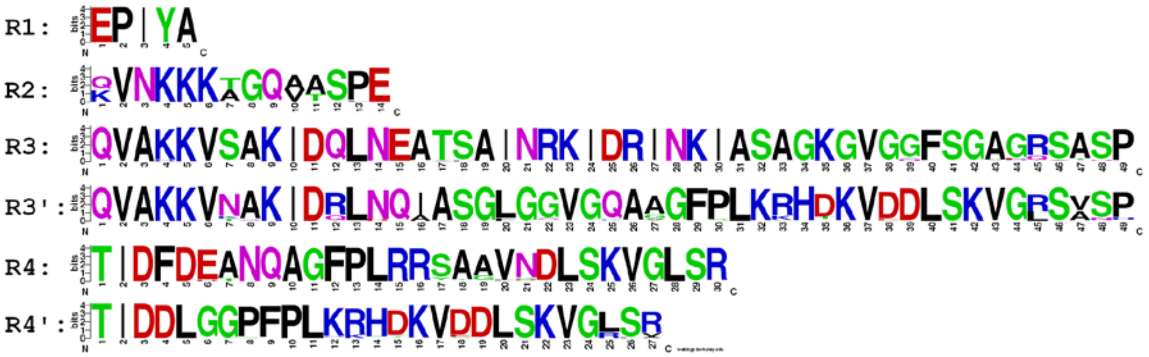

Figure 6. Profiles of the CagA repeat regions.

4.2.2 Data preprocessing

Based on the previous description in Ref. [150], we named the EPIYA motif and the following intervening regions R1, R2, R3, R3', R4 and R4' (Figure 6). Figure 7 shows the position relation between the EPIYA motif (R1) and other intervening regions by using the CagA types A-B-D (East Asian subtype) and A-B-C (Western subtype) as examples. R2 is relatively conserved across both subtypes, but there are significant differences between the intervening regions R3 and R3', as well as between R4 and R4'. The East Asian subtype and the Western subtype were treated as two independent groups. Their data was then processed and the results were analyzed within each group individually.

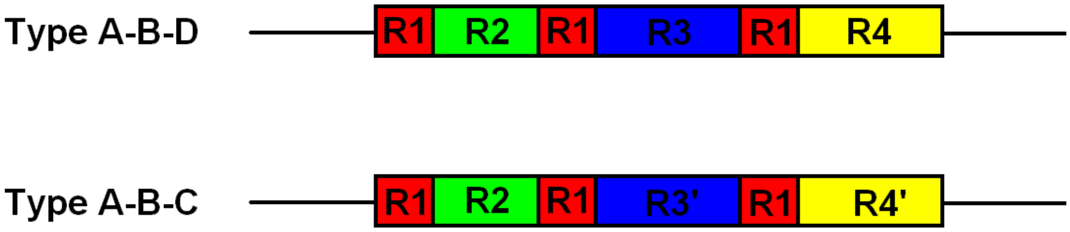

Figure 7. Structure examples: A-B-D and A-B-C types of CagA sequences (not on a proportional scale to sequence length).

All intervening regions were extracted from the CagA sequences and put into the corresponding subtype groups, and then the multiple sequence alignments were applied for each group individually by using Clustal X version 2.0.3 [35]. The sequences profiles (Figure 6) was built by using the Weblogo 3 [164].

4.2.3 Workflow

Figure 8 shows the workflow of the classification/prediction procedure:

- Select one strain as the test strain.

- Apply a bootstrap procedure to the rest of the strains to get the training strains.
- Calculate the feature entropy for the test strain based on training strains and save it as the test data.
- Calculate the feature entropy for each strain in the training strain set based on training strains and save them as the training data.
- Generate classification model by using the training data.
- Classify the test data according to the classification model.
- Repeat this procedure five times, and then calculate the average as the final result.

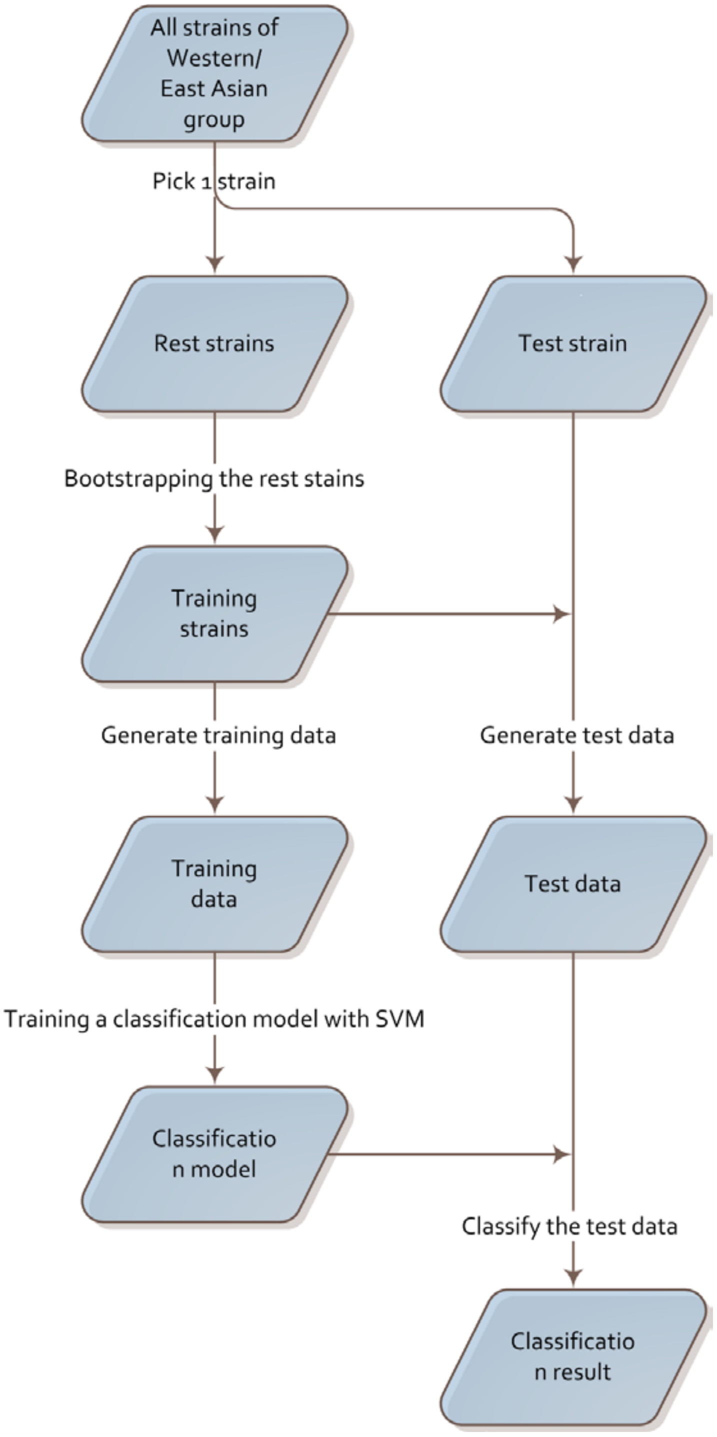

Figure 8. Workflow of classification/prediction procedure for one specific CagA sequence.

4.2.4 Bootstrapping

A major issue in building a classification model in this case is the big difference of the sample sizes between cancer and non-cancer groups, which could cause bias in the classification results. A bootstrapping procedure was applied to address this issue. In each subtype group, for each training/test data sets, all non-cancer samples were included, and then strains were continuously drawn from the cancer group on a random basis until reaching the same size of the non-cancer group. In this case, all the available data were used although cancer samples were utilized multiple times given their smaller size compared to the non-cancer group. This procedure was applied five times to generate five independent training sets for each test sequence. The classification/prediction result is the average of those five independent results.

4.2.5 Residue detection and feature-entropy calculation

Since CagA is related to almost all gastroduodenal diseases and simple analysis of EPIYA motif repeats does not yield any statistically significant differences among those diseases, the information indicating a specific disease might be hidden in the intervening regions. This research assumes that there is a set of residues or residue combinations that could be useful as a marker of a specific disease. This study focuses on the gastric cancer and uses the cancer/non-cancer groups as the example.

We used the similar approach as presented in chapter 3. Based on the aligned sequences for each intervening region, specific residues were identified by comparing the difference of combinatorial entropy [122] between the cancer and non-cancer groups. This procedure includes the following steps:

First of all, we divide the given multiple alignments for all intervening regions into two groups: gastric cancer group and non-cancer group. For each column of multiple alignments, we compute the background entropy (Eq. 1) and the combinatorial entropy (Eq. 2), described in chapter 3.

Then the entropy difference between the combinatorial entropy and the background entropy is calculated:

$$\Delta E = C_i - B_i \tag{3}$$

Considering the three extreme cases illustrated in Figure 4, according to the calculation results of the entropy difference for the above three cases, the combinatorial entropy is $C_i = 0$ for both 'globally conserved' and 'locally conserved' cases. For 'randomly and uniformly distributed' case, C_i gets the maximum value. We can distinguish the 'conserved' and 'randomly and uniformly distributed' cases based on the combinatorial entropy, but it does not help pick 'locally conserved' case from all 'conserved' cases. When we consider the background entropy at the same time, B_i gets the maximum value, 0 and medium value for the 'randomly and uniformly distributed' case, 'globally conserved' case, 'locally conserved' case, respectively. Finally, the differences for the above three cases are: $\Delta E_1 = 0$, $\Delta E_2 = 0$, and ΔE_3 gets the minimum value. Hence, the entropy difference is a proper measurement for detecting a 'locally conserved' sequence pattern.

Based on the above calculation, it can be determined that correct grouping can minimize the entropy difference for those residues belonging to the 'locally conserved' case. To perform a test, one sequence is selected while the rest of the sequences are divided into a gastric cancer group and a non-cancer group. For all selected residues, the selected sequence is placed into the gastric cancer group to calculate the entropy difference ΔE_{CA}, and then it is placed into non-cancer group to get the corresponding entropy difference ΔE_{NON-CA}. Finally, $\Delta E' = \Delta E_{CA} - \Delta E_{NON-CA}$ is obtained for all selected residues that are used as the feature entropy.

4.2.6 Cross-validation

Because the data size is small, a leave-one-out (LOO) cross-validation procedure was performed. This is not only an assessment of the classifier performance on training/test data, but also an estimate of prediction power for novel cases.

4.2.7 SVM

We chose SVM as binary classifier and used the feature-entropy vectors to train and test the classifier. In the case of two-class soft margin classification, the decision function is a weighted linear combination defined as follows:

$$f(x) = \sum_{x_i \in S} \alpha_i y_i K(x, x_i) + b \qquad (4)$$

where $K(x, x_i)$ represents a user-defined kernel function that measures the similarities between the input feature vector x and the feature vectors x_i in the training dataset S. α_i is the weight assigned to the training feature vector x_i and y_i indicates whether a CagA strain has been labeled with the positive class (+1) or negative class (-1). The primal optimization problem takes the form:

minimize
$$t(w,\xi)=\frac{1}{2}\|w\|^2+\frac{C}{m}\sum_{i=1}^{m}\xi_i \tag{5}$$

subject to

$$y_i\left(\left\langle\Phi\left(x_i\right),w\right\rangle+b\right)\geq 1-\xi_i \; and \; \xi_i\geq 0\left(i=1,\ldots,m\right) \tag{6}$$

where $\Phi^T\left(x_i\right)\Phi\left(x_j\right)=K(x_i,x_j)$. *m* is the total number of strains. ξ_i is a slack variable which measures the degree of misclassification of the datum. C is a cost parameter which allows for trading off training error against model complexity. *w* is the normal vector and *b* is the offset.

After comparing the results of polynomial, tanh and Gaussian radial basis kernels, the result obtained with the RBF kernel worked the best, where the Gaussian radial basis kernels (RBF: $\exp(-\gamma\|x_i-x_j\|^2)$) are for general-purpose learning when there is no prior knowledge about the data. The SVM$^{\text{Light}}$ package (http://svmlight.joachims.org/) [165] was employed to build our application. The parameters C and γ were tuned to get the best model for the training data as shown in the following. All other SVM parameters were set to their default values.

4.2.8 Performance evaluation

In order to evaluate the performance of classifier, a variety of performance measures are applied: accuracy, sensitivity and specificity. A true positive (TP) is a cancer-related sequence classified as such, while a false positive (FP) is a non-cancer related sequence classified as cancer-related, a false negative (FN) is a cancer related sequence classified as non-cancer related and a true negative (TN) is a non-cancer related sequence classified as non-cancer related. The accuracy, sensitivity (Sn), specificity (Sp) and Matthews correlation coefficient (MCC) of classification is defined as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{7}$$

$$Sp = \frac{TN}{FP + TN} \tag{8}$$

$$Sn = \frac{TP}{TP + FN} \tag{9}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{10}$$

Since there are only two parameters (C, γ) for the RBF kernel and they are independent, we applied a grid-search to determine the optimal parameters of classifier. We used a harmonic means of sensitivity and specificity as the objective function to optimize the performance of the model for the training set, which is defined as follows:

$$F = \frac{2(Sp \times Sn)}{Sp + Sn} \tag{11}$$

4.3 Results

4.3.1 Residue detection and feature calculation

Table 6 lists all detected key residues by calculating the entropy difference in each intervening region for both Western and East Asian subtypes. Although there are some geographic variations of CagA sequences between the Western and East Asian subtypes, some common residues could still be found to distinguish the cancer and non-cancer groups. It suggests that those residues might be very important in determining the virulence of CagA and the relation between CagA and some specific diseases.

Table 6. Detected residues by calculating entropy difference for each intervening region between Western and East Asian subtypes.

Intervening regions	Western subtype	East Asian subtype
R2	1, 8, 10, 12	1, 8, 10, 12
R3'/R3	7, 8, 14, 35, 38	2, 4, 17, 31, 39, 48
R4'/R4	9, 14	16

The residue positions are shown in Figure 9. A previous study [162] reveals that the different EPIYA segments can bind to the different kinases, e.g., EPIYA-R2 and EPIYA-R3/R3' bind to the C-terminal Src kinase (Csk) while EPIYA-R4 and EPIYA-R4' bind to the SHP-2 kinase to cause the hummingbird phenotype. The CagA-Csk interaction down-regulates CagA-SHP-2 signaling that perturbs cellular functions to control the virulence of CagA. It is found that most detected residues belong to R2 and R3/R3' regions and few residues in R4/R4' regions have been detected. This may be because R4/R4' has more conserved sequence than R2, and R4/R4' is shorter than R3/R3'. We suggest that the different residue patterns in R2 or R3/R3' regions might change the ability of down-regulating CagA-SHP-2 signaling, therefore changing the virulence of CagA.

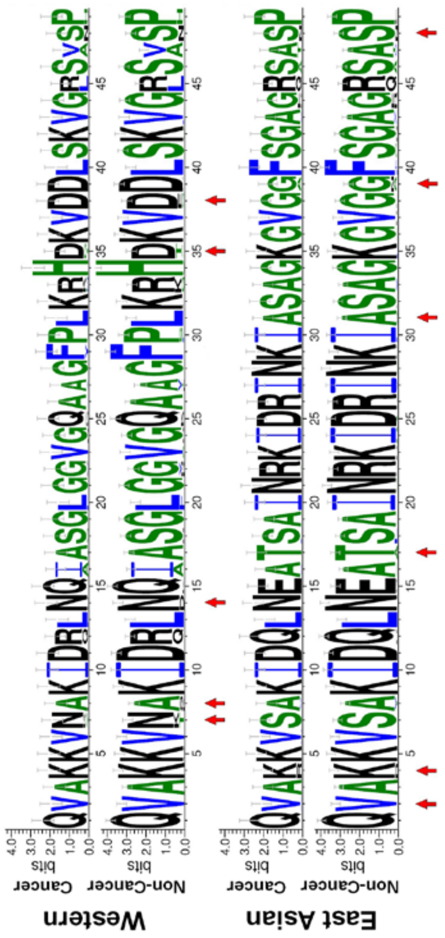

Figure 9. Comparison of sequence profiles between cancer and non-cancer groups and selected features (residues marked with arrows) based on entropy calculation in the R3'/R3 region.

Ren et al. found that CagA multimerizes in mammalian cells [166]. This multimerization is independent to the tyrosine phosphorylation, but it is related to the "FPLxRxxxVxDLSKVG" motif which is named CM motif in the R3' intervening region. Since the multimerization is a prerequisite for the CagA-SHP-2 signaling complex and subsequent deregulation of SHP-2, the CM motif plays an important role in cagA-positive *H. pylori*-mediated gastric pathogenesis. With multiple CM motifs *H. pylori* strains are much likely associated with severe gastroduodenal diseases [166, 167], but this observation cannot explain why different gastroduodenal diseases can be developed with the exact same number of CM motifs. Our study detected two residues in the CM motif of R3' intervening region, which might lead to the change of multimerization, thus changing the virulence of CagA. This is in consistent with a previous discovery [168] that the sequence difference between the East Asian CM and the Western CM determines the binding affinity between CagA and SHP-2.

While the key residues detected can reveal some difference between cancer and non-cancer groups, no single residue can be a marker for cancer as shown in Figure 9. This research predicts that one special combination of all or partial detected residues could have a high correlation with one particular disease. To verify, several linear statistical models, e.g. linear regression and logistic regression, were applied to the detected features to evaluate the importance of each residue and the correlation between selected residues and cancer. However, none of above models were able to produce a statistically significant result. Since the features cannot be fitted by simple linear models for predicting cancer, applying a machine learning method to analyze and classify these data becomes necessary.

4.3.2 Parameter training for classification

Using the Western subtype group as the example, a loose grid-search was first performed on $C = 2^{-5},...,2^{10}$ and $\gamma = 2^{-10},...,2^{5}$ (Figure 10A) and found that the best (C,γ) is around $\left(2^{2},2^{-5}\right)$ to get the highest F value with the LOO cross-validation rate 76%. Then a finer grid search was conducted on the neighborhood $\left(2^{2},2^{-5}\right)$ and a better F value was obtained with 79.7% LOO cross-validation at $\left(2^{2.25},2^{-4.75}\right)$. The same procedure was utilized for the East Asian subtype group and the best LOO cross-validation rate 72.6% was reached at $\left(2^{2.25},2^{-4.75}\right)$.

Since there are no previous studies or computational methods on the same topic, evaluating the performance of this research's new method is difficult. To assess the information content of the sequences in terms of their discerning power to predict cancer, a random shuffling procedure was employed to build the control group. First, all sequences from the Western subtype were placed together to build a sequence pool. Second, we randomly picked the same number of sequences as cancer group from the sequence pool and treated the rest of the sequences as the non-cancer group. Then, the whole training procedure was applied to newly shuffled data to find the best (C,γ). The above steps were repeated five times to generate five independent shuffled data sets. The one with the highest *F* value, which equals 46.6% was selected and its contour plot is shown in Figure 10B. This randomly shuffling evaluation was also applied to the East Asian subtype data and the best *F* value was at 54.3%. Comparing the two plots shows the significant difference of *F* values between the data with correct grouping of cancer and non-cancer cases in training and the best randomly shuffled data. The result suggests that the intervening regions are informative to distinguish between the cancer and non-cancer groups and our method can use the information effectively.

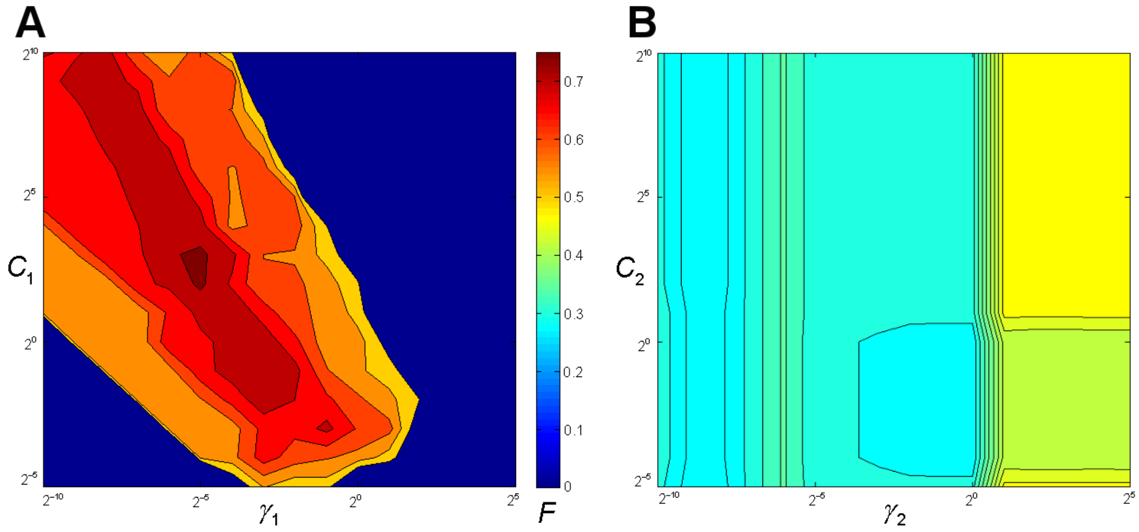

Figure 10. Grid-search for determining the optimal parameters (C,γ) of classifier, with color indicating the F value. (A) The contour plot of F value resulting from a loose grid-search on a hyper parameter range for the Western subtype group. (B) The contour plot of F value resulting from a loose grid-search on a hyper parameter range for a randomly shuffled Western subtype group with the highest F value.

4.3.3 Classification performance

There are mainly three categories of sequence classification methods: feature based, sequence distance based and model based. The method that we described in this dissertation belongs to the feature-based category. We selected two of the most popular sequence classification tools as the representative methods of other two categories for comparison. BLAST [53] was chosen for the sequence distance based category, since it is the most widely used sequence comparison tool. For the model-based category, the hidden Markov model is the typical method for sequence analysis and its widely used tool, HMMER [169], was selected. For the classification procedure of both BLAST and HMMER, we used the default parameters of the tools, applied the same LOO cross-validation as our method, and used the same evaluation formulas listed in the Method section.

Table 7. Classification performance

Subtype	No. of cancer cases	No. of non-cancer cases	Method	Sn	Sp	Accuracy	F value	MCC
Western	37	211	Entropy-SVM	0.86	0.74	0.76	0.80	0.45
			BLAST	0.22	0.77	0.69	0.34	-0.01
			HMMER	0.94	0.005	0.14	0.009	-0.16
East Asian	47	240	Entropy-SVM	0.74	0.71	0.71	0.73	0.35
			BLAST	0.17	0.75	0.65	0.28	-0.07
			HMMER	1	0.003	0.19	0.05	0.06

Table 7 lists the classification results for all three methods. The SVM method performs significantly better than the other two approaches. BLAST achieved close accuracy to the Entropy-SVM method, but it predicted many false negatives with low sensitivity. HAMMER achieved high sensitivity but with little specificity. Considering *F* values and *MCC* values, the prediction results from BLAST and HAMMER are almost random.

The classification result and the contour plot (Figure 10) strongly support our hypothesis, i.e., the information of the selected residues in intervening regions can be used to classify the relation between CagA sequences and gastric cancer, although the difference between the profiles of cancer and non-cancer groups is not very strong.

Comparison among different diseases

Table 8. Number of strains in each disease

	GC	AG	CG	GU	DU	Other	Total
East Asian	47	49	45	47	79	20	287
Western	37	8	44	14	50	95	248

GC: Gastric cancer; AG: Atrophic gastritis; CG: Chronic gastritis; GU: Gastric ulcer; DU: Duodenal ulcer

H. pylori infection is associated with most gastroduodenal diseases, among which gastric cancer is the most severe one causing more than 700,000 deaths worldwide every year [170]. Since *H. pylori* is a main risk factor of gastric cancer (GC), discovery of the mechanism of *H. pylori* mediating GC becomes a top priority task in this field. Comparing to other diseases, the diagnosis information of GC from public data is relatively accurate, and it is another important reason to focus on GC in this chapter. Our studies are not limited to GC, though. We also tried to evaluate the relations between the variance of CagA sequences and different diseases.

Since most data were collected from public databases without accurate diagnosis information, before applying our method to CagA data, we manually curated the disease annotations for all strains by reviewing the literature. Table 8 lists the distributions of major diseases for both the Western and the East Asain subtype groups. Due to the limitation of strain numbers of some diseases, such as atrophic gastritis (AG) and gastric ulcer (GU), we eventually picked chronic gastritis (CG) and duodenal ulcer (DU) as the control groups for evaluation. The DU group in the East Asian subtype contains 79 strains, and a bootstrapping procedure was applied to all other groups to make the same number of strains as the East Asian DU group. This step guarantees all comparisons on the same scale, since the value of combinatorial entropy depends on the number of sequences. We used Formula (3) to calculate the entropy difference of each position between GC and CG/DU groups, and then added up all entropy differences as the total difference between GC and CG/DU groups, as shown in Table 9. By comparing results between two groups within the same geographic subtype (East Asian or Western subtype), it is consistent with the clinical view that gastritis has stronger relations to cancer than to DU [171] (generally, gastritis cases might contain some unreported or undiagnosed chronic atrophic gastritis and intestinal metaplasia cases, with which patients have a high risk to develop GC). By considering the same disease-pair between two geographic subtypes, it also explained the virulent difference between the East Asian and the Western subtypes. In addition, due to the high similarity between different disease groups of the East Asian subtype, even with more data, we still cannot reach the same classification accuracy as the Western subtype group.

Table 9. Total entropy difference between gastric cancer and two other diseases groups

	GC vs CG	GC vs DU
Western	-166.65	-536.42
East Asian	-57.18	-244.03

GC: Gastric cancer; CG: Chronic gastritis; DU: Duodenal ulcer

Based on the above results, CagA sequences show potential to distinguish multiple gastroduodenal diseases. In order to evaluate the classification performance, we used DU group to replace non-Cancer group, and then applied the whole classification procedure again without bootstrapping, since those two diseases groups have comparable sizes. Table 10 shows the classification results. Although from the clinical point of view, DU has the negtive correlation with GC among all gastroduodenal diseases [172], the classification performance of two subtype groups was only slightly improved. Thus cancer-related CagA strains might have some unique sequence patterns comparing to all other gastroduodenal diseases. Hence, tuning a subset of the control group may not be able to improve the classification accuracy.

Table 10. Classification performance between gastric cancer and duodenal ulcer groups for both the Western and the East Asian subtypes

Subtype	No. of GC cases	No. of DU cases	Sn	Sp	Accuracy	F value	MCC
Western	37	50	0.86	0.82	0.84	0.84	0.68
East Asian	47	79	0.68	0.73	0.71	0.71	0.41

4.4 Discussion

Although research indicates that there are sequence markers to differentiate between cancer group and non-cancer group, the major profiles of those two groups are too similar to distinguish by using traditional methods since the CagA sequences are overall highly conserved. Therefore, we focused on identifying the informative residues, quantifying information of these selected residues, and then using it to design a classifier that can predict whether a new sequence belongs to the cancer group or the non-cancer group. This method not only sheds light on the relations between CagA sequences and gastric cancer, but also may provide a useful tool for gastric cancer diagnosis or prognosis.

One possible explanation of the performance difference between the Western subtype group and the East Asian subtype group is that there is no real 'control' group for the East Asian subtype, and that all strains are from the patients with other diseases. In contrast, the Western subtype group contains 24 strains from volunteers (the normal controls). The mechanisms of *H. pylori* causing the different gastroduodenal diseases are still unclear, however it is likely that various gastroduodenal diseases caused by *H. pylori* infection share some sequence patterns in the intervening regions. Small variations of amino acids in those important residues might lead to the virulence variance of CagA strains resulting in different gastroduodenal diseases. While CagA could be a marker for detecting potential cancer risk, using CagA alone to distinguish all gastroduodenal diseases is not realistic. As a future study, we will develop new models that differentiate various gastroduodenal diseases from cagA and other genes.

5. EFFECTOR PREDICTION IN HOST-PATHOGEN INTERACTION

5.1 Introduction

As a complex and interesting relation between organisms in ecology and evolution, host-pathogen interaction is a basis of infectious diseases [173]. Pathogens span a broad spectrum of biological species, including viruses, bacteria, fungi, protozoa, and multicellular parasites. In all these cases, a pathogen causing an infection usually exhibits an extensive interaction with the host during pathogenesis. The cross-talks between a host and a pathogen allow the pathogen to successfully invade the host organism, to breach its immune defence, as well as to replicate and persist within the organism. One of the most important and therefore widely studied groups of host-pathogen interactions is the interaction between pathogen protein (effector) and host cells. Effectors are secreted from pathogens' secretion systems. So far five types of secretion systems have been identified (Type I-V). Among them, T3SS (Type III Secretion System) and T4SS (Type IV Secretion System) can cross bacterial cell walls and host eukaryotic cell membranes to deliver effectors into host cells directly without going through extracellular matrix [174]. Those effectors can manipulate host cell functions once entering host cell [174]. Identifying effectors and exploring their molecular mechanisms not only are critical to understand the disease mechanisms but also provide theoretical foundations for infectious disease diagnosis, prognosis and treatment [175, 176].

Besides to be an important virulence factor of *H. pylori*, as a well-studied effector, CagA one of the major pathogens of upper gastrointestinal diseases (e.g., peptic ulcer and gastric cancer) [177]. CagA can be delivered into gastric epithelial cells by the T4SS of *H. pylori*. Recent studies of CagA sequences found that they have a variable region within which the EPIYA (glutamic acid-proline-isoleucine-tyrosine-alanine) motif repeats from once to seven times. Tyrosine in the EPIYA motif can be phosphorylated in the host cell. The phosphorylated CagA protein binds to a phosphatase SHP-2, which will interfere with the signal transduction pathway of the host cell and manipulate cell growth, differentiation and apoptosis [150, 162, 178]. This interference causes a restructure of the host cell cytoskeleton, cell scattering as well as invasive growth of cells, and formation of hummingbird phenotype with gastric epithelial cells. Such a process not only is considered an important strategy of interaction between *H. pylori* and host cell, but also is the most significant mechanism of pathogenesis and carcinogenesis of *H. pylori* [179-181].

In recent years, studies have discovered other pathogens that can also secrete effectors to manipulate the host cells through phosphorylation during the interaction process between hosts and pathogens (e.g. *Anaplasma phagocytophilum* [182, 183] and *Bartonella henselae* [184-186]). These effectors cause rearrangements of host cell cytoskeleton, NF-kB activation and apoptosis inhibition [187]. Table 11 lists eight effectors from six pathogens. They contain 28 experimentally identified phosphorylation sequences, all of which have the similar pattern to the EPIYA motif in CagA [185]. This finding leads to our hypothesis that the EPIYA-like motif and its phosphorylation, together with its interference of host cells, may be a general mechanism of pathogenesis. Based on this novel hypothesis, we used the effectors in Table 11 to build an EPIYA-motif-based hidden Markov model (HMM), and then searched the current protein database to identify more proteins with the EPIYA motif. Through studying the distribution and features of EPIYA motif in different species and genuses, we attempted to better understand the function of EPIYA motif, especially the role of EPIYA motif during the interaction process between pathogens and hosts.

Table 11. Experimentally determined tyrosine-phosphorylated effectors and their motifs

Effector	Pathogen	Locus of protein	Motif (phosphorylated Y position)					
CagA	*H.Ppylori*	NP_207343	EPIYAKVNK	Y-899	EPIYTQVAK	Y-918	EPIYATIDD	Y-972
Ankyrin	*Anaplasma phagocytophilum*	ABB84853	ESIYEEIKD EDLYATVGA EPIYATVKK	Y-940 Y-1028 Y-1098	ESIYEEIKD ESIYADPFD	Y-967 Y-1056	ESIYEEIKD ESIYADPFA	Y-994 Y-1074
BepD	*Bartonella henselae*	YP_034066	EPLYAQVNK EPLYAQVNK	Y-32 Y-211	NPLYEGVGG NPLYEGVGG	Y-114 Y-293	NPLYEGVGS NPLYEGVGP	Y-176 Y-355
BepE	*Bartonella henselae*	YP_034067	EPLYATVNK	Y-37	ETIYTTVSS	Y-91		
BepF	*Bartonella henselae*	YP_034068	TPLYATPSP EPLYATAAP	Y-149 Y-297	EPLYATPLP EPLYATPLP	Y-213 Y-269	EPLYATPLP	Y-241
Tir	*Escherichia Eoli*	AAC38390	EHIYDEVAA	Y-474				
Tir	*Citrobacter rodentium*	AAL06376	EPIYDEVAP	Y-468				
Trap	*Chlamydie trachomatis*	YP_001654788	ENIYENIYE	Y-136	ENIYENIYE	Y-238	ENIYENIYE	Y-390

CagA: cytotoxin associated gene A [146, 148-150, 178, 180, 188, 189]; BepD: *Bartonella henselae* protein D [184-186, 190]; BepE: *Bartonella henselae* protein E [184-186, 190]; BepF: *Bartonella henselae* protein F [184-186, 190]; Tir: translocated intimin receptor

[191-193]; Trap: Translocated actin-recruiting protein [194-196]. The first five amino acids of the listed sequences in the table correspond to the EPIYA motif.

5.2 Methods and materials

Figure 11 shows the whole work flow for this project. The major modules and steps are described as follows:

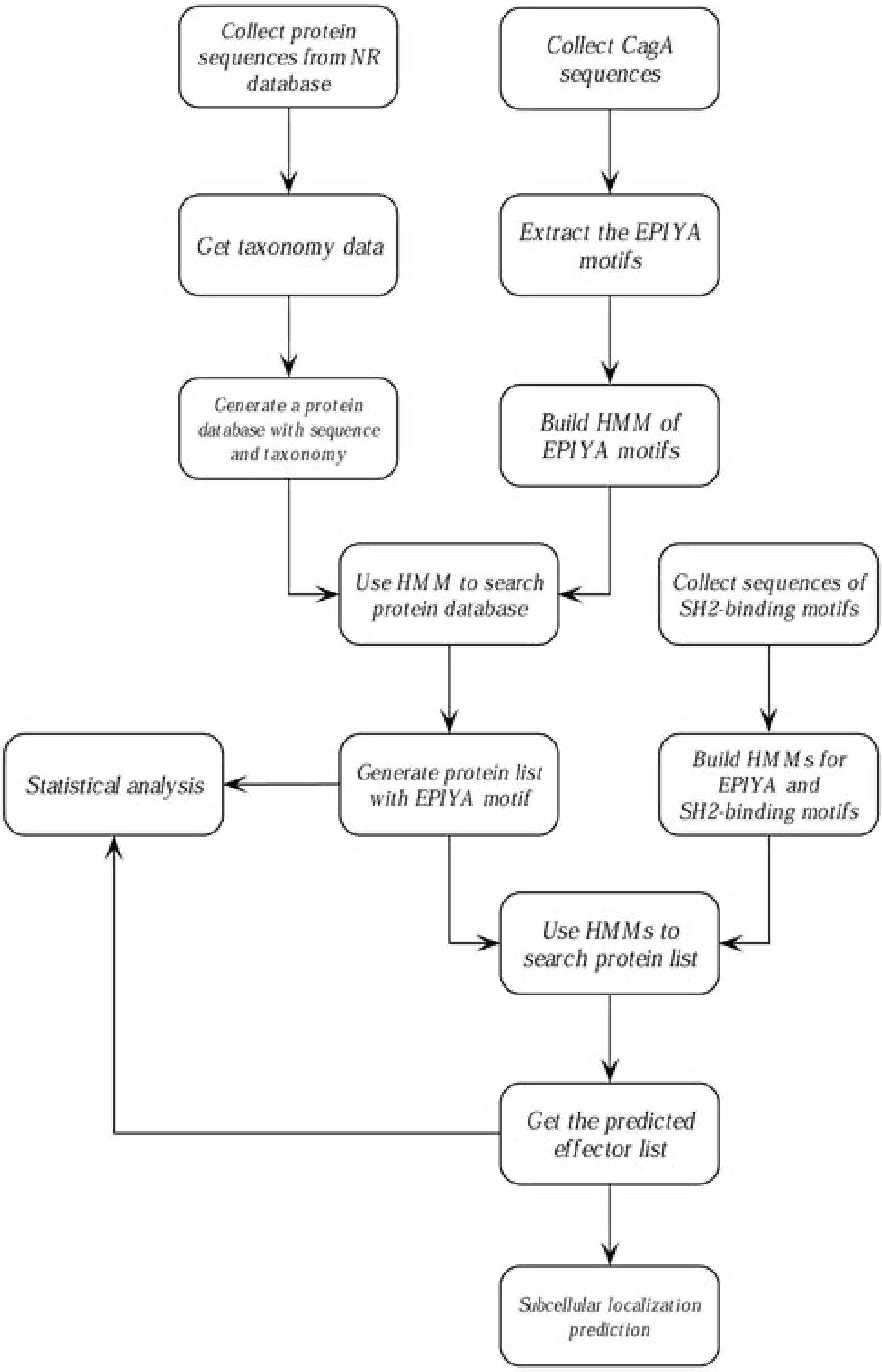
Collect protein sequences from NR database
Collect CagA sequences
Get taxonomy data
Extract the EPIYA motifs
Generate a protein database with sequence and taxonomy
Build HMM of EPIYA motifs
Use HMM to search protein database
Collect sequences of SH2-binding motifs
Statistical analysis
Generate protein list with EPIYA motif
Build HMMs for EPIYA and SH2-binding motifs
Use HMMs to search protein list
Get the predicted effector list
Subcellular localization prediction

Figure 11. Work flow of the whole project

5.2.1 Data sources

Protein sequence data: We used the NR (non-redundant) protein database at the National Center for Biotechnology Information (NCBI) in this study. All protein sequences in the FASTA format were downloaded from the NCBI site (ftp://ftp.ncbi.nih.gov/blast/db/FASTA/nr.gz; as of July 6th 2009; 9,216,047 sequences). We excluded "other" sequences and "unclassified" sequences" in the database (as labelled in http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Root).

Taxonomy data: The taxonomy data was obtained from the NCBI website (http://www.ncbi.nlm.nih.gov/Taxonomy/txstat.cgi; as of July 6th 2009).

5.2.2 Hidden Markov model

A hidden Markov model was built by using Hmmer 2.3.2 [169] (http://hmmer.janelia.org). We used selected sequences to run the command hmmbuild.exe for building and calibrating the HMM. We then used the HMM to run the command hmmsearch.exe for searching protein sequences. We used a natural cutoff of HMM score such that the last of the all known motifs is retrieved.

5.2.3 Data analysis

We used Perl (release ActivePerl 5.8.8) as the programming language to analyse the data and build the database. We applied SAS 9.0 (http://www.sas.com) as the statistic analysis tool and chose p<0.01 as the significant threshold.

5.2.4 Sequences comparison

BioEdit 7.0 (http://www.mbio.ncsu.edu/BioEdit/bioedit.html), Lasergene 7 (http://www.dnastar.com/products/lasergene.php), and Blast [53] (http://blast.ncbi.nlm.nih.gov/Blast.cgi) were used to compare and analyse the protein sequences. Sequence logos were constructed using Weblogo[164].

5.3 Results and discussion

5.3.1 Building and using hidden Markov model

Using the 28 experimentally identified phosphorylated motif sequences in Table 11, we built the sequence logo as shown in Figure 12. In this logo, the fourth position of the EPIYA motif is always tyrosine (Y), which can be phosphorylated. The first and third positions have small variations. The amino acids in the first position are primarily glutamic acid (E), together with asparagines (N). Most residues in the third position are isoleucine (I) and leucine (L), two very similar amino acids. The second and fifth positions have big variations. The second position varies from proline (P), serine (S) to asparagines (N). The fifth position mainly contains alanine (A), glutamic acid (E) and aspartic acid (D). These 28 sequences were used to build and calibrate the HMM by applying Hmmer 2.3.2 (http://hmmer.janelia.org). We then employed the HMM to search the protein non-redundant (NR) database, which contains 9,216,047 protein sequences. The search yielded 107,231 sequences containing at least one copy of EPIYA motif and 3115 sequences with multiple repeats of the EPIYA motif, where the highest number of repeats in a single protein is 29 (see Table 12).

Table 12. Distribution of protein sequences containing the EPIYA motif

Number of motif repeats in one protein	Number of protein sequences	Observed Frequency	Expected Frequency
29	1	1.09E-07	3.44E-57
14	2	2.17E-07	5.52E-28
13	2	2.17E-07	4.88E-26
12	1	1.09E-07	4.32E-24
10	1	1.09E-07	3.39E-20
9	3	3.26E-07	3.00E-18
8	6	6.51E-07	2.65E-16
7	10	1.09E-06	2.35E-14
6	32	3.47E-06	2.08E-12
5	55	5.97E-06	1.84E-10
4	173	1.88E-05	1.63E-08
3	916	9.94E-05	1.44E-06
2	1913	2.08E-04	1.28E-04
1	104116	1.13E-02	1.13E-02

Expected frequency is the expected probability if the combination of the motif in a protein sequence is random.

We found that the repeats of EPIYA motif in a protein are highly non-random. As the probability of one protein sequence having a copy of EPIYA motif is 1.13E-02 (104,116/9,216,047), the expected probabilities of one protein sequence containing 2-4 copies of EPIYA motif are $(1.13E\text{-}02)^2$=1.28E-04, $(1.13E\text{-}02)^3$=1.44E-06, and $(1.13E\text{-}02)^4$=1.63E-08, respectively, assuming the combination of the motif in a sequence is random. The observed probabilities of one sequence containing multiple copies of EPIYA motif are much larger than the expected probabilities as shown in Table 12. Hence, the repeats of the EPIYA motif may have been resulted from evolution with biological significance. This is also reflected in Table 11, where most effectors with known EPIYA motif have 2-7 motif repeats. Thus, we suggest that multiple copies of EPIYA motif in the same protein are more likely to be functional than single motif occurrence.

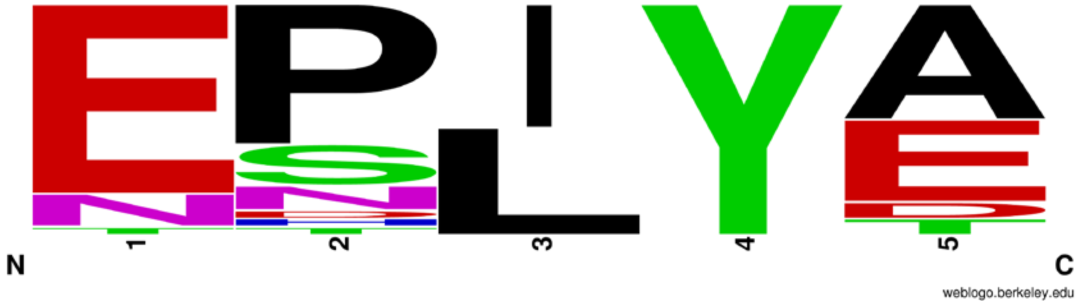

Figure 12. Work flow of the whole project

5.3.2 Distribution pattern of EPIYA motif among species

The NR database contains proteins sequences from 27,432 genuses and 121,718 species. Among them the sequences from 2675 genuses and 4646 species contain at least one copy of EPIYA motif, and 368 genuses and 587 species have proteins containing at least two copies of EPIYA motif. The proteins with the EPIYA motif are mainly distributed in lower organisms. As shown in Table 13, the probability of a genuses/species containing proteins with the EPIYA motif in archaea, viruses or bacteria is much higher than that in eukaryotes. This indicates that with evolution advanced species mostly lost the EPIYA motif together with its functions for host-pathogen interactions.

Table 13. Distribution of EPIYA-motif containing proteins at genus and species levels (as of July 6th 2009)

Groups	Number of genuses					Number of species				
	total	With copies of motif≥1 (%)		With copies of motif≥2 (%)		total	With copies of motif≥1 (%)		With copies of motif≥2 (%)	
Archaea	109	49	44.95%	19	17.43%	330	90	27.27%	28	8.48%
Viruses	623	221	35.47%	18	2.89%	6443	433	6.72%	30	0.47%
Bacteria	1198	560	46.74%	209	17.45%	6291	1398	22.22%	360	5.72%
Eukaryota	35499	1828	5.15%	122	0.34%	108654	2725	2.51%	169	0.16%
-Protista	1263	109	8.63%	18	1.43%	3747	186	4.96%	32	0.85%
-Fungi	1509	121	8.02%	40	2.65%	5772	206	3.57%	52	0.90%
-Metazoa	22309	662	2.97%	49	0.22%	62097	826	1.33%	68	0.11%
-Viridiplantae	10418	936	8.98%	15	0.14%	37038	1507	4.07%	17	0.05%
total	37429	2658	7.10%	368	0.98%	121718	4646	3.82%	587	0.48%

Data in this table presents the numbers of genuses/species with proteins containing the EPIYA motif versus total number of genuses/species in NR. The Eukaryota group is divided into protista, fungi, metzoa and viridiplantae.

We listed top 10 species and genuses with most EPIYA-motif containing proteins for the groups of archaea, viruses, bacteria, protista, fungi, metazoa and viridiplantae (Table 14). In archaea, *Methanococcus* is the genus that includes the most EPIYA-motif containing proteins. In viruses, *Potyvirus* is the highest in number of EPIYA-motif containing proteins among genuses while *Bovine Viral Diarrhea Virus* is the highest among species. The top four genuses (and the corresponding species) in bacteria are *Helicobacter* (*H. pylori*), *Clostridum* (*Clostridum botulinum*, *Clostridum perfringens*), *Bacillus* (*Bacillus cereus*) and *Anaplasma* (*Anaplasma phagocytophilum*). *Plasmodium* (*Plasmodium falciparum*) and *Tetrahymena* (*Tetrahymena thermophila*) are the top genuses in protista. In fungi and viridiplantae, the corresponding top genuses are *Candida* (*Candida tropicalis*) and *Oryza* (*Oryza sativa*), respectively. Two well-studied genuses *Drosophila* (*Drosophila melanogaster*) and *Homo* (*Homo sapiens*) take the top two in metazoa. It should be noted that the data in Table 14 are biased, with widely studied species such as *H. pylori* having the same gene sequenced many times, while some other species have incomplete proteomes. Nevertheless, this table provides some interesting reference for known and putative pathogens with effectors.

Table 14. Top 10 genuses and species containing most proteins with at least two copies of EPIYA motif for each group

Group	rank by genus	occurrence	rank by species	occurrence
Archaea	Methanococcus	11	Methanococcus maripaludis	9
	Methanocaldococcus	6	Aciduliprofundum boonei	4
	Aciduliprofundum	4	Halogeometricum borinquense	4
	Halogeometricum	4	Methanocaldococcus jannaschii	4
	Methanobrevibacter	3	Methanobrevibacter smithii	3
	Methanosarcina	3	Halorubrum lacusprofundi	2
	Thermococcus	3	Hyperthermus butylicus	2
	Haloferax	2	Methanosarcina barkeri	2
	Halorubrum	2	Methanospirillum hungatei	2
	Hyperthermus	2	Archaeoglobus fulgidus	1
Viruses	Potyvirus	14	Bovine Viral Diarrhea Virus	11
	Pestivirus	11	Zucchini yellow	8
	Orthopoxvirus	5	Bean common	5
	Simplexvirus	5	Grapevine virus	5
	Vitivirus	5	Cowpox virus	3
	Capripoxvirus	4	Lumpy skin	2

	Yatapoxvirus	3	Papiine herpesvirus	2
	Alphabaculovirus	2	Tanapox virus	2
	T4-likeviruses	2	Acidianus filamentous	1
	Alphapapillomavirus	1	Aeromonas phage	1
Bacteria	Helicobacter	1024	Helicobacter pylori	1021
	Bacillus	133	Bacillus cereus	78
	Clostridium	102	Anaplasma phagocytophilum	46
	Anaplasma	48	Bacillus thuringiensis	28
	Bacteroides	23	Clostridium botulinum	28
	Vibrio	17	Clostridium perfringens	20
	Lactobacillus	16	Cyanothece sp.	11
	Cyanothece	11	Bacteroides sp.	10
	Ureaplasma	11	Lactococcus lactis	10
	Campylobacter	10	Chlamydia trachomatis	9
Protista	Plasmodium	103	Plasmodium falciparum	47
	Tetrahymena	35	Tetrahymena thermophila	35
	Paramecium	26	Paramecium tetraurelia	26
	Entamoeba	14	Plasmodium yoelii	19
	Leishmania	14	Trichomonas vaginalis	14
	Trichomonas	14	Plasmodium vivax	12
	Cryptosporidium	11	Plasmodium knowlesi	11
	Giardia	9	Entamoeba dispar	8
	Monosiga	8	Giardia lamblia	8
	Theileria	7	Monosiga brevicollis	8
Fungi	Candida	21	Candida tropicalis	9
	Aspergillus	15	Paracoccidioides brasiliensis	9
	Pichia	12	Candida albicans	8
	Paracoccidioides	9	Pichia stipitis	6
	Ajellomyces	7	Vanderwaltozyma polyspora	6
	Vanderwaltozyma	6	Ajellomyces capsulatus	5
	Coccidioides	5	Cryptococcus neoformans	5
	Filobasidiella	5	Saccharomyces cerevisiae	5
	Saccharomyces	5	Aspergillus terreus	4
	Debaryomyces	4	Candida dubliniensis	4
Metazoa	Drosophila	178	Homo sapiens	69
	Homo	69	Mus musculus	58
	Mus	58	Drosophila melanogaster	41
	Pan	30	Pan troglodytes	29
	Branchiostoma	27	Branchiostoma floridae	27

	Caenorhabditis	21	Rattus norvegicus	21
	Rattus	21	Canis lupus	20
	Canis	20	Danio rerio	17
	Danio	17	Drosophila pseudoobscura	17
	Macaca	16	Drosophila persimilis	15
Viridiplantae	Oryza	24	Oryza sativa	23
	Physcomitrella	12	Physcomitrella patens	12
	Arabidopsis	10	Arabidopsis thaliana	10
	Populus	7	Populus trichocarpa	7
	Sorghum	7	Sorghum bicolor	7
	Ricinus	6	Ricinus communis	6
	Vitis	6	Vitis vinifera	6
	Micromonas	3	Micromonas pusilla	3
	Huperzia	2	Huperzia lucidula	2
	Ostreococcus	2	Zea mays	2

Bacterial pathogens can be divided to two types. Some can enter host cells, e.g., *Chlamydia* and *Anaplasma*. They are known as intracellular pathogens and most of them have T3SS or T4SS. Some other bacteria are extracellular pathogens with T3SS or T4SS, such as *H. pylori* and *Campylobacter*. As shown in Table 14, numerous bacteria containing multiple copies of EPIYA motif belong to pathogens, such as *Anaplasma* (*Anaplasma phagocytophilumn*, ranking 4th in the species list) and *Chlamydia* (*Chlamydia trachomatis*, ranking 10th in the species list), both of which are intracellular pathogens. *Helicobacter* (ranking 1st in both genus and species list) and *Campylobacter* (ranking 10th in the genus list) contain T4SS [197]. Some other T4SS-containing species are not listed in Table 14, such as *Wolbachia* (ranking 13th in the genus list), *Escherichia* (ranking 18th in the genus list), *Mycobacterium* (ranking 23rd in the genus list) and *Bartonella* (ranking 27th in the genus list). Furthermore, we also found that in protista most top 10 species are unicellular parasites that can live in host cells to survive and reproduce by subverting of signalling pathways and inhibiting apoptosis of host cells [198]. However, the pathogens mediators responsible for this modulation are still unknown [199]. Those intracellular protozoan parasites include *Plasmodium* (ranking 1st in the genus list), *Leishmania*, *Trichomonas*, *Cryptosporidium* and *Giardia* (correspondingly ranking 5-8th in the genus list). Table 15 lists most known pathogens including intracellular bacteria (*Mycobacteriaceae*, *Legionellales*, *Chlamydiales*, *Rickettsiales* and *Listeriaceae*), extracellular bacteria with T3SS or T4SS (*Enterobacteriaceae*, *Campylobacterales* and *Rhizobiales*) and intracellular protozoan parasites (*Apicomplexa* and *Kinetoplastida*), representing 1319 species in total with and without the EPIYA motif. We analyzed the distribution of EPIYA motif in the potential effectors in these pathogens in Figure 13. 310 out of 4646 species with the EPYIA motif belong to such pathogens, and the percentage (310/4646=6.67%) is much higher than that of such pathogens (with and without the EPIYA motif) in all species (1319/121,718=1.08%) (p-value<0.0001, odds ratio=6.54). We also found that the percentage of species with the EPYIA motif belonging to pathogens increases significantly with the increase of EPIYA motif repeats in a protein sequence.

Table 15. Known intracellular bacterial pathogens or bacteria containing III/IV type secretion system, and intracellular parasitic protozoan

Bacteria			**Protista**		
Genus	Type	Number of species	**Genus**	Type	Number of species
Enterobacteriaceae		245	**Apicomplexa**		187
Salmonella	T3SS		*Babesia*	IPP	
Yersinia pestis	T3SS		*Cryptosporidium*	IPP	
Shigella	T3SS		*Plasmodium*	IPP	
Escherichia	T3SS		*Isospora*	IPP	
Campylobacterales		76	*Toxoplasma*	IPP	
Campylobacter	T4SS		*Theileria*	IPP	
Helicobacter	T4SS				
Wolinella	T4SS				
Rhizobiales		346	**Kinetoplastida**		120
Brucella	IPB		*Leishmania*	IPP	
Bartonella	IPB		*Trypanosoma*	IPP	
Agrobacterium	T4SS				
Rickettsiales		83			
Anaplasma	IPB				
Ehrlichia	IPB				
Wolbachia	IPB				
Rickettsia	IPB				
Chlamydiae		23			
Chlamydia	IPB				
Legionellales		62			
Legionella	IPB				
Coxiella	IPB				
Rickettsiella	IPB				
Mycobacteriaceae		168			
Mycobacterium	IPB				
Listeriaceae		9			
Listeria	IPB				

IPB: intracellular parasitic bacteria; IPP: intracellular parasitic protozoan; T3SS: type III secretion system; T4SS: type IV secretion system.

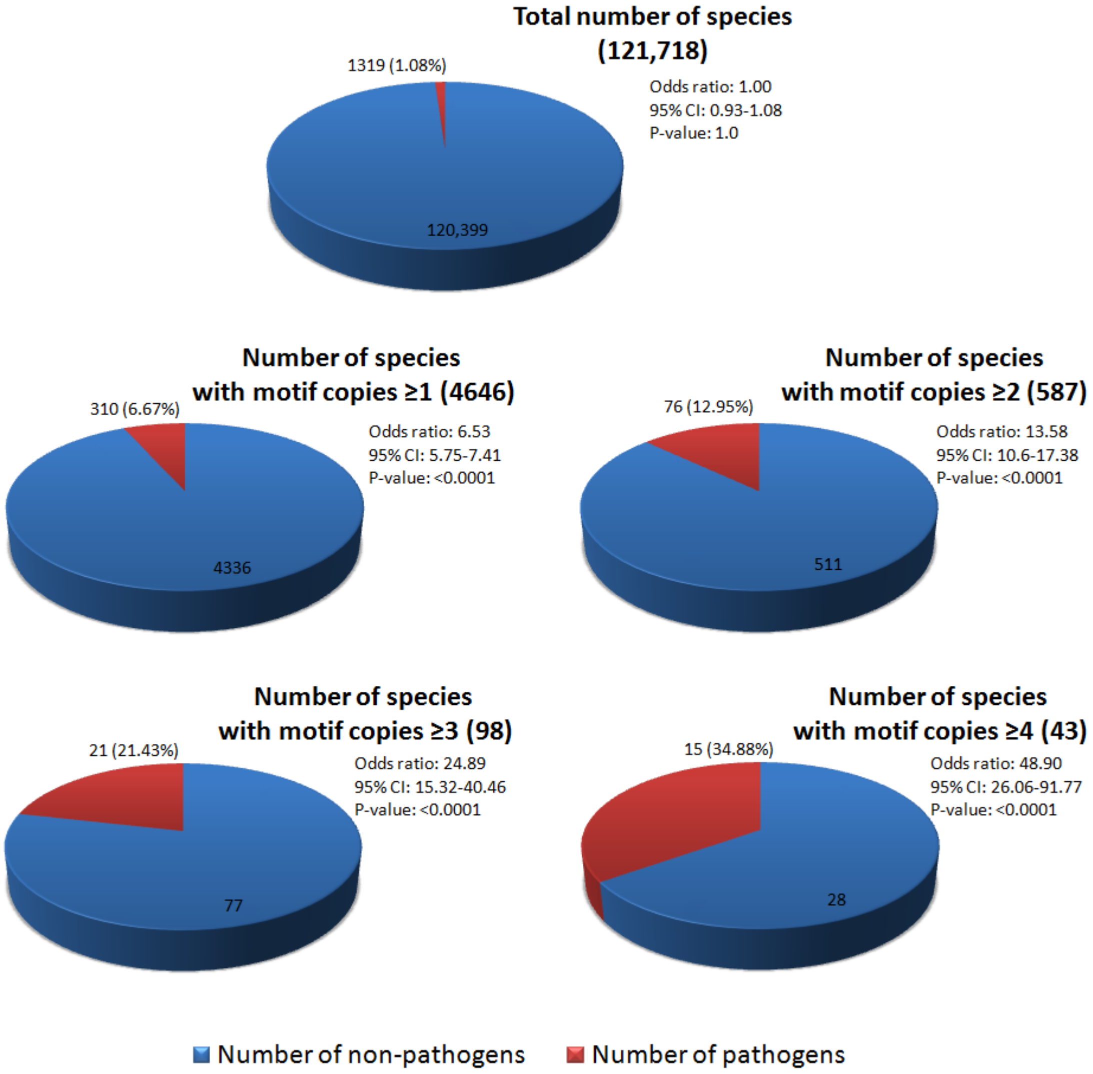

Figure 13. Relationship between number of EPIYA motif copies and number of species in known pathogens

Table 16. Distribution of top 40 protein sequences containing at least two copies of EPIYA motif

Protein Name	Number of proteins (number of genuses)							
	Total	Archaea	Viruses	Bacteria	Protista	Fungi	Metazoa	Viridiplantae
CagA	1015(1)			1015(1)				
hypothetical protein	689(186)	15(10)	10(6)	242(88)	162(19)	78(28)	127(25)	55(11)
ATP*	81(21)	2(2)		68(11)	6(4)	4(3)		1(1)
Ankryin	55(7)			51(3)			4(4)	
DNA*	52(34)	3(3)		40(25)	7(4)	1(1)		1(1)
Kinase	43(28)	5(2)		23(15)	4(3)		11(8)	
zinc finger protein	43(11)						43(11)	
TPR repeat protein	33(15)			33(15)				
Polyprotein	24(2)		24(2)					
SecA	23(14)			23(14)				
Peptidase	19(12)	1(1)		16(9)	1(1)		1(1)	
dynein heavy chain	17(13)				4(2)	1(1)	11(9)	1(1)
elongation factor 2	15(7)				10(2)	1(1)	4(4)	
Palmdelphin	14(9)						14(9)	
tRNA*	14(11)	2(1)		10(8)	1(1)	1(1)		
glycogen synthase	13(1)			13(1)				
GTP-binding	13(3)			12(2)	1(1)			
transcriptional regulator	13(8)	1(1)		9(4)		3(3)		
unc-119 homolog	13(6)						13(6)	
FAT tumor suppressor homolog 3	12(9)						12(9)	
nuclear ribonucleoprotein	12(9)				1(1)	10(7)		1(1)
4-alpha-glucanotransferase	9(1)			9(1)				
paternally expressed 3	8(6)						8(6)	
Striatin	8(7)						8(7)	
Tarp	8(1)			8(1)				
nuclear autoantigen	7(6)						7(6)	
putative mannosyltransferase	7(1)			7(1)				
Ubiquitin	7(6)				5(4)	2(2)		
26S proteasome regulatory subunit	6(3)							6(3)
cell division protein	6(4)	3(3)		1(1)				
centaurin, delta 3	6(5)						6(5)	
fat tumor suppressor homolog 2	6(5)						6(5)	
glycosyl transferase	6(5)			6(5)				
guanine nucleotide exchange factor	6(6)						6(6)	
cytochrome c oxidase subunit VI	5(4)				5(4)			
PEG3	5(5)						5(5)	
polyketide synthase	5(4)			3(3)	2(1)			
polysaccharide biosynthesis protein	5(3)			5(3)				
TatD-related deoxyribonuclease	5(1)			5(1)				
translation initiation factor	5(4)			2(2)	2(1)	1(1)		

ATP* includes ATPase, ABC transporter, ATP-binding protein, and ATP-dependent helicase; DNA* includes DNA photolyase, DNA primase, DNA repair protein, DNA-binding protein, and DNA mismatch repair protein; kinases* includes histidine kinase, protein kinase, hexokinase, serine kinase, and fyn-related kinase; tRNA* includes tRNA synthetase, tRNA formyltransferase, and tRNA ligase.

5.3.3 Distribution pattern of EPIYA-motif containing proteins

In our search result, there are totally 3115 protein sequences with at least two EPIYA motif repeats. Among them, most are CagA of *H. pylori* as this protein has been extensively studied, and 689 out of 3115 are hypothetical proteins whose functions have not been identified. Based on protein functions, the top 40 proteins ranking by occurrence under each protein function type are widely distributed. They not only exist in achaea, viruses and bacteria, but also are found in protista, metazoa and viridiplantae. Besides the known EPIYA-motif containing effectors ankryin and Tarp, they also involve enzymes related to DNA, ATP and tRNA, transcription regulators, tumor suppressors, different types of kinases, zinc-finger proteins, ubiquitin and various metabolic enzymes (Table 16).

Although many of these predicted effectors are false positives and the EPIYA motif may not be functional in them, a significant portion of them is likely to be true effectors. As known effectors, ankyrin and TPR (tetratricopeptide repeat) are related to protein-protein interaction [200, 201]. Considering the sequence similarity of the above proteins, 44 sequences of ankryin are highly similar among each other and come exclusively from *Anaplasma phagocytophilum*, *Wolbachia endosymbiont* and *Ehrlichia sp.*, all of which belong to *Rickettsiales*. Except the sequences from *Haliangium ochraceum* (ZP_03879805 and ZP_03880192), other sequences of TPR repeat-containing proteins are also similar and they are from *Trichodesmium erythraeum*, *Stigmatella aurantiaca*, *Acaryochloris marina*, *Cyanothece sp.* and *Microcoleus chthonoplaste*e. For the 20 hypothetical proteins, YP_034066 and YP_001610012 (*Bartonella*), YP_153762 and YP_002563468 (*Anaplasma*), XP_001623017 and XP_001636029 (*Nematostella*), ZP_01620341 and ZP_01622571 (*Lyngbya*), XP_001468598 and XP_001686356 (*Leishmania*) are similar pairs in sequences (with more than 30% sequence identity in each pair), and two proteins in a pair come from the same genus. The EPIYA motif in these proteins is highly conserved during evolution, and it may play similar roles as the motif in CagA.

Among proteins containing at least four copies of EPIYA motif with 286 sequences in total, most of them are from bacteria, especially from intracellular bacterial pathogens or extracellular bacterial pathogens with T3SS or T4SS, and some are from protist, e.g., intracellular protozoan parasites. Four out of eight known effectors (CagA, Ankyrin, BepD, and Tarp) are found in these sequences, and thus other proteins from bacteria and protista may also be effectors. An interesting observation is that the percentage of protein sequences having the EPIYA motif in archaea is the highest among all groups (see Table 13), but none of these archaeal proteins contain four or more copies of EPIYA motif. Previous studies revealed that CagA sequences with more EPIYA-motif occurrences are more virulent [162]. Since archaea and other organisms have relationships of either mutualism or commensalsim and till now there is no clear evidence for the existence of archea parasites [202, 203], it is unlikely that the archaeal proteins containing the EPIYA motif act as pathogen effectors. Compared to other groups, archaea is not well studied, but we can still find some interesting examples, such as *Methanobrevibacter smithii* (ranking 5th in the species list in Table 14). It is the most common commensal archaea in the human gut and plays an important role in digesting polysaccharides, while it may not benefit the host directly. We speculate that EPIYA-motif containing proteins of *Methanobrevibacter smithii* may have some biological functions in this commensal interaction.

Many functions listed in Table 16 may reflect the fact that these proteins may have multiple functions other than phosphorylation-induced signalling control in the host cell. Some of these proteins may mimic host protein functions. For example, it was suggested that CagA functions as a prokaryotic mimic of the eukaryotic Grb2-associated binder (Gab) adaptor protein [204]. Some of the predicted effectors may mimic singling proteins, such as HPK (histidine protein kinase) listed in Table 16, which is an important part of two-component signal transduction system that recognizes and transmits environmental signals [205]. Some known effectors induce protein expressions with increased expression of RNA polymerase [206]. It is not surprising to see a significant number of proteins in Table 16 are related to protein synthesis, such as RNA polymerase, elongation factor, and helicase. It is noted that CagA itself also contains an RNA polymerase domain based on a BLAST search. These connections also suggest the ancestor proteins of the predicted effectors. Many of the proteins listed in Table 16 are ancient house-keeping genes. The predicted effectors might have evolved from these house-keeping genes by mimicking the host genes. Furthermore, over evolution some of the effectors or their ancestors might have evolved into genes with different functions unrelated to host-pathogen interactions, such as EPIYA-motif containing proteins in archaea and metazoan.

5.3.4 Building HMM based on KK, R4, Tarp and Tir motifs

It is known that biological effects of CagA induced by phosphorylation depend on the binding to the SH2 domain. Different combinations of the five amino acids after phosphorylated tyrosine (pY) will bind different SH2 domains and cause different downstream effects. There are two known motifs in CagA that could bind to SH2 domain - EPIYAKVNK and EPIYATIDD(F), which are referred to as KK motif and R4 motif [151, 207], respectively. For the Tir protein, we retrieved all sequences in *Escherichia coli* and *Citrobacter rodentium* with the pattern of EHIYDEVAA(P) and built a motif (named as Tir motif). We did the same for Tarp protein, yielding the motif ENIYENIYE (named as Tarp motif). We extracted sequences of known KK, R4, Tir and Tarp motifs in proteins CagA, Tir and Tarp (including their variants), and then built HMMs one by one. Figure 14 shows the sequence logo of each motif. Comparing with the sequence logo in Figure 12, the 9-mer motifs are more conserved and specific. This will help reduce the false positive rate in identifying putative effectors, while the downstream SH2 binding partners are also predicted at the same time.

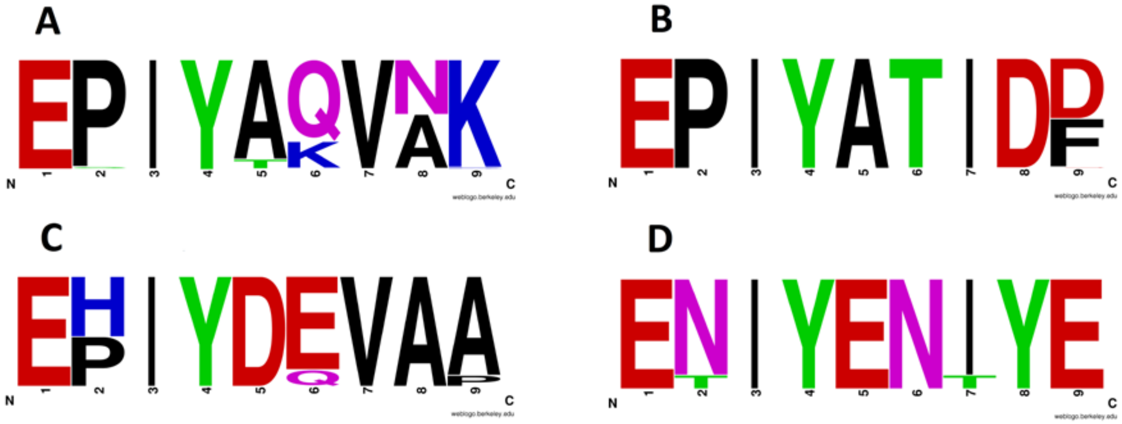

Figure 14. Sequence logos for KK, R4, Tir and Tarp motifs

A: the logo was built with 1705 KK motif sequences extracted from 842 CagA protein sequences; B: the logo was built with 979 R4 motif sequences extracted from 842 CagA protein sequences; C: the logo was built with 20 Tir motif sequences extracted from 20 Tir protein sequences of Escherichia Eoli and Citrobacter rodentium; D: the logo was built with 16 Tarp motif sequences extracted from 7 Tarp protein sequences of Chlamydie trachomatis.

5.3.5 Search results by using HMMs based on KK, R4, Tarp, and Tir motifs

Using HMMs based on KK, R4, Tarp and Tir motifs to search the protein sequences containing the EPIYA motif as described above, we found that the results are widespread in many species. In this chapter we only focus on the results in bacteria and protista. As shown in Table 17, CagA KK (EPIYAKVNK) motif exists in some known phosphorylation effectors, e.g., Beps (BepD, BepE, BepH) and Tir. CagA R4 (EPIYATIDD) motif exists in Tarp. Both KK and R4 motifs exist in ankyrin. Among 8 proteins for building our EPIYA-motif based HMM, BepF is the only one containing neither KK nor R4 motif. The Tir protein just have one motif -Tir motif in Table 11, while BAF52548 (Tir of *E. coli*) contains two motifs, i.e., Tir (EHIYDEVAA) motif and EPIYAKIQR, similar to the KK motif. Tarp protein (YP_001654788) of *Chlamydia trachomatis* contains not only the Tarp motif (ENIYENIYE), but also another motif ENIYESIDD, which is similar to the R4 motif.

Table 17. Sequences containing KK and R4 motifs in known effectors

KK Motif	Species	Protein	pY prosition	Locus
EPIYAKVNK	*H.pylori*	cagA	Y-899	NP_207343
EPIYTQVAK	*H.pylori*	cagA	Y-918	NP_207343
EPIYAKIQR	*E.coli*	Tir	Y-477	BAF52548
EPIYATVKK	*Anaplasma phagocytophilum*	Ankyrin	Y-1094	ABB84853
EPLYAQVNK	*Bartonella henselae*	BepD protein	Y-28	YP_034066
EPLYATVNK	*Bartonella henselae*	BepE protein	Y-33	YP_034067
EDLYATVGA	*Anaplasma phagocytophilum*	Ankyrin	Y-1024	ABB84853
R4 Motif	**Species**	**Protein**	**pY prosition**	**Locus**
EPIYATIDD	*H.pylori*	cagA	Y-972	NP_207343
ENIYESIDD	*Chlamydia trachomatis*	Tarp	Y-189	YP_001654788
ESIYEEIKD	*Anaplasma phagocytophilum*	Ankyrin	Y-990	ABB84853

It reveals that although these proteins are not similar in global sequences (weak similarity exists between Beps sequences), they share the same or similar motifs with significant functional relationships.

5.3.6 Prediction of new effectors

Based on the above HMMs of KK, R4, Tir and Tarp motifs, we predicted some new pathogen effectors and we assessed them based on the literature. The details of predicted effectors have been listed as follow.

(1) *Bartonella tribocorum*: Since BepH contains the EPLYAQVNK (YP_001610013, Y-8) motif (KK motif), we predicted it as a phosphorylation effector like BepD-F secretory proteins.

(2) *Lawsonia intracellularis*: *Lawsonia intracellularis* is an obligate intracellular bacterial pathogen, which infects a wide range of animals, mainly pigs, and causes proliferative enteropathy - a type of contagious diseases [208, 209]. Its symptoms are acute, including diarrhea, loss of appetite and stunting. After an initial close association with the cell membrane of the enterocytes, *Lawsonia intracellularis* is endocytosed into host cell [210]. Infected host cells are inhibited in maturation, continue to undergo mitosis and proliferation, and at last form hyperplastic crypts, but the mechanism is unknown [211]. The genome sequence of *Lawsonia intracellularis* indicates that it may possess a type III secretion system, which may assist the bacterium during cell invasion and evasion of the host's immune system and could be a mechanism for inducing cellular proliferation [212, 213], but its effectors secreted by T3SS was never reported. Current database contains 20 proteins of *Lawsonia intracellularis* with the EPIYA motif and all of them are from strain PHE/MN1-00. The maximum sequence identity between any two of these 20 proteins is 22% and most of them are enzymes, e.g. ATP synthase. Among them, in the HMM search result by using the R4 motif, we found that

hypothetical protein LI0666 (YP_595041) contains two copies of EPIYA motif (EPIYAEIKT Y-149, EPIYAEIKT Y-186), which are similar to the R4 and Tir motifs, respectively. Thus, we speculate that this protein might be the effector of *Lawsonia intracellularis* to interact with intestinal epithelial cells.

(3) *Ehrlichia sp.:* It belongs to the same family *Ehrilichiaceae* as *Anaplasma* [214]. Ankyrin of *Ehrlichia sp.* and ankyrin of *Anaplasma* share 89% sequence identity. Ankyrin (T08612) of *Ehrlichia sp.* contains six copies of 9-mer motifs including the KK and R4 motifs, and thus it is a likely effector of *Ehlichia* to interact with host.

(4) *Wolbachia*: *Wolbachia* belongs to *Rickettsiales. Wolbachia* is a symbiotic bacterium existing in the sex organ of many insects. Though ankyrin (AAY54257) of *Wolbachia* and ankryin of *Anaplasma* share only 15% sequence identity, they contain almost exactly the same motifs. Hypothetical protein WD0942 (NP_966676), which is not similar to ankyrin in sequence, has two motifs, and one of them is EPIYATVPK(Y-318) similar the KK motif. EsorChan1 (AAP34173) contains the motif EPIYDEVYD (Y-77) similar to the Tir motif. Therefore, the above three proteins, especially the first two, are potential effectors of *Wolbachia* [215].

(5) *Pasterurella multocida*: As the major pathogen to cause swine infectious atrophic rhinitis, it secretes toxin filamentous hemagglutinin containing six copies of EPIYA motif. Based on the BLAST search results, we found that the filamentous hemagglutinin (AAK61595) of *Pasterurella multocida*, filamentous hemagglutinin of *Bordetella pertussis* and *Bordetella Parapertussis* share ~30% sequence identity [216-218]. Filamentous hemagglutinin, the major virulence factor of *Bordetella pertussis*, not

only has adhesion function, but also plays a critical role in immunomodulation. Since filamentous hemagglutinin has the sequences EDIYATINK (Y-2792), which is similar to the KK motif, EHIYADIRD (Y-2550) and ENLYAEISD (Y-2651), both of which are similar to the R4 motif, and EHLYAEINE (Y-2387), which is similar to the Tir motif, we suggest that filamentous hemagglutinin being the effector of *Pasterurella multocida* and it might be secreted by the TPS (Two-Partner Secretion) system [217]. PfhB2 (NP_244996) has four sequences that are similar to KK, R4 and Tir motifs, and thus it might be another candidate of effector in *Pasterurella multocida.*

(6) *Haemophilus ducreyi*: *Haemophilus ducreyi* is a facultative anaerobic Gram-negative coccobacillus and could cause the sexually transmitted disease chancroid. Large supernatant protein2 (NP_873623) of *Haemophilus ducreyi* has six copies of EPIYA motif. Its sequence and filamentous hemagglutinin of *Bordetella pertussis* share 41% sequence identity. Its sequence EPVYADLHF and EPVYADLRF are similar to the R4 motif. Hence, we suggest large supernatant protein2 (NP_873623) is a potential effector of *Haemophilus ducreyi* and it could be secreted by T4SS [219]. The effector can lead to immunosuppression, inhibition of proliferation, and permanent changes in host cells [220-222].

(7) *Haemophilus somnus*: *Haemophilus somnus* can survive in host cells and is the cause of a variety of systemic diseases in cattle, including thrombotic-meningoencephalitis, pneumonia, arthritis, myocarditis, septicemia and other reproductive diseases [223, 224]. Cysteine protease domain YopT-type (YP_001784809) and filamentous hemagglutinin of *Bordetella pertussis* share 42% sequence identity. The

sequence EPIYATLDK (Y-2933) in YP_001784809 is similar to the KK motif, EHIYEQIGE (Y-2358) similar to the Tarp motif, and EPVYDKVSA (Y-2287) similar to the Tir motif. Thus, YP_001784809 might be the effector of *Haemophilus somnus* to cause immunosuppression [225].

(8) *Chlamydophila pneumonia*: Hypothetical protein CPj0472 (NP_300527) contains three copies of EPIYA motif. EPIYANTPE (Y-647) is similar to the KK motif, EPIYEEIGG (Y-346) is similar to the Tir motif and EPIYDEIPW (Y-681) is similar to the R4 motif. Although we did not find any similar protein through BLAST search, hypothetical protein CPj0472 (NP_300527) is a good candidate for the effector of *Chlamydophila pneumonia*.

(9) *Leishmania major*: *Leishmania major* could parasitize into phagocyte of human or other mammals and is responsible for the disease leishmaniasis, which is a serious zoonosis. *Leishmania major* have 6 proteins containing at least two copies of EPIYA motif and 4 of them are proteins with unknown functions. Among these 6 proteins, Cytochrome C oxidase subunit VI (XP_001683136) contains two copies of EPIYA motif. One is at position Y-107 with sequence EPLYQPVKK, which is similar to the KK motif. Another one is at position Y-130 with sequence EPLYDVDAA, which is similar to the Tir motif. Hence, XP_001683136 might be an effector. Hypothetical protein (XP_001686159) has three copies of EPIYA motif and the sequences are all EPLYAVTIE, which is similar to KK and R4 motifs. Hypothetical protein (XP_001686160) also has three copies of EPIYA motif and the sequences are all EPLYAVTID, which is similar to the R4 motif. In addition, hypothetical protein XP_001686159 and XP_001686160 share 43% sequence identity. Hypothetical protein (XP_001686356) has 29 copies EPIYA motif (the

one with most EPIYA motifs in our data) and all sequences are the same as EPLYAVTLE, which is similar to the R4 motif. Microtubule-associated protein (XP_001687515) contains two copies of EPIYA motif. One is at Y-1543 and another is at Y-1589. The sequences for both of them are ESIYAKDYK, which is similar to the KK motif. Thus, we predict Hypothetical protein (XP_001686159), hypothetical protein (XP_001686160), hypothetical protein (XP_001686356) and Microtubule-associated protein (XP_001687515) might also be the effectors of *Leishmania major*. For another potential effector hypothetical protein (XP_001683914), although it contains two copies of EPIYA motif (ESLYE is at Y-1006 and EHLYD is at Y-1047), they are not similar to KK, R4, Tarp or Tir motif and hence less likely to be an effector than the above five proteins.

(10) *Plasmodium falciparum*: *Plasmodium falciparum* can invade human liver cells and RBC to cause dangerous infection malaria. It contains many proteins with the EPIYA motif and 47 proteins with at least two copies of EPYIA motif. Among them, Plasmodium exported protein (XP_001347309) has three copies of motif which are all similar to the Tarp motif. The sequences and the corresponding pY sites are ESIYKNKLK (Y-331), ESIYKNKLK (Y-359), and ESIYKNKLE (Y-387). Thus we predict it as the effector of *Plasmodium falciparum*. Conserved *Plasmodium* protein (XP_001347469) has eight copies of EPIYA motif, RNA pseudouridylate synthase (XP_001350676) has nine copies of EPIYA motifs and hypothetical protein (XP_001351018) has three copies of EPIYA motif, but none of them contains any of KK, R4, Tir and Tarp motifs, and therefore is less likely to be the effector than Plasmodium exported protein (XP_001347309).

5.3.7 Protein subcellular localization prediction

Based on our effectors prediction results, we applied the subcellular localization prediction for all bacteria effectors by using CELLO v.2.5 [226] (http://cello.life.nctu.edu.tw). First, we chose ‘Gram negative’ as the organisms to perform the prediction, since those effectors are all from Gram-negative bacteria. 9 out 11 effectors were predicted as extracellular or outermembrane (Table 18). Then we reapplied the prediction by choosing ‘Eukaryotes’ as the organisms and the results of all 11 effectors are nuclear (Table 18). The above results show that most our predicted effectors have the same attributes as the real effectors.

Table 17. Subcellular localization prediction results of predict bacteria effectors

Species	Effector	Gram-negative			Eukaryotes		
		1st Prediction	2nd Prediction	3rd Prediction	1st Prediction	2nd Prediction	3rd Prediction
Bartonella tribocorum	BepH (YP_001610013, Y-8)	Cytoplasmic 1.669 *	Periplasmic 1.432 *	Extracellular 1.363 *	Nuclear 4.119 *	Extracellular 0.637	Cytoplasmic 0.115
Lawsonia intracellularis	hypothetical protein LI0666 (YP_595041)	Extracellular 2.545 *	Periplasmic 1.335	OuterMembrane 1.335	Nuclear 3.154 *	Mitochondrial 0.943	Cytoplasmic 0.485
Ehrlichia sp.	ankyrin (T08612)	Periplasmic 2.003 *	OuterMembrane 1.214	Extracellular 1.131	Nuclear 1.419 *	Cytoplasmic 1.373 *	Mitochondrial 0.941
Wolbachia	ankyrin (AAY54257)	Extracellular 2.511 *	OuterMembrane 1.723	Periplasmic 0.42	Nuclear 3.542 *	Cytoplasmic 0.566	Mitochondrial 0.248
	Hypothetical protein WD0942 (NP_966676)	Extracellular 1.709 *	OuterMembrane 1.471 *	Periplasmic 1.275 *	Nuclear 4.265 *	Mitochondrial 0.439	Cytoplasmic 0.192
	EsorChan1 (AAP34173)	Cytoplasmic 3.122 *	Periplasmic 0.727	Extracellular 0.585	Nuclear 2.181 *	Cytoplasmic 1.470	Extracellular 0.646
Pasterurella multocida	filamentous hemagglutinin (AAK61595)	OuterMembrane 2.344 *	Extracellular 1.723 *	InnerMembrane 0.594	Nuclear 2.622 *	Cytoplasmic 0.996	PlasmaMembrane 0.479
	PfhB2 (NP_244996)	OuterMembrane 2.182 *	Extracellular 1.632 *	InnerMembrane 0.6	Nuclear 2.374 *	Cytoplasmic 0.990	PlasmaMembrane 0.482
Haemophilus ducreyi	Large supernatant protein2 (NP_873623)	Extracellular 2.318 *	OuterMembrane 1.802 *	InnerMembrane 0.593	Nuclear 2.638 *	Cytoplasmic 0.550	PlasmaMembrane 0.484
Haemophilus somnus	Cysteine protease domain YopT-type (YP_001784809)	Extracellular 2.245 *	OuterMembrane 1.844 *	InnerMembrane 0.593	Nuclear 2.551 *	Cytoplasmic 0.944	PlasmaMembrane 0.482
Chlamydophila pneumonia	Hypothetical protein CPj0472 (NP_300527)	OuterMembrane 2.415 *	Extracellular 1.224	Periplasmic 0.577	Nuclear 3.099 *	Cytoplasmic 0.949	PlasmaMembrane 0.286

5.4 Conclusions

In this charpter, we showed that the EPIYA motif might be a ubiquitous functional site for effectors that play an important role in pathogenicity for mediating host-pathogen interactions. Most known effectors have more than one copy of EPIYA motif. The predicted effector sequences of pathogens from the same genus are likely homologous, and those from different genuses are rarely homologous although they often share common motifs. Most pathogens are intracellular bacteria or long-term chronic infection of extracellular bacteria, e.g., *H. pylori*. Usually effectors are secreted by T3SS or T4SS to enter host cells, and then interfere signal transduction pathway of the host cell to disturb host cell functions, which mainly involve actin polymerization, cell proliferation, apoptosis and immunosuppression, so as to improve the abilities of survival and propagation of microorganism with host-pathogen interaction.

Our study predicted many putative effectors. We grouped the phosphorylated EPIYA motifs into four types, KK, R4, Tir and Tarp based on the sequence features of the five amino acids after Y, and then we used them individually to build the HMM. After using the HMMs to search our database and considering the known pathogenic characteristics of pathogens, we predicted some effectors of bacteria and also suggested that using our method will discover more effectors with the EPIYA motif. Besides the discovery in bacteria, we also found that there were many protein sequences containing the EPIYA motif in protist pathogens. Intracellular protozoan parasites can live in host cells to survive and reproduce by subverting of host cell signalling [198], to induce downstream effects, e.g., inhibiting apoptosis of host cells, restructuring of the host cell cytoskeleton, and so on. However, the pathogen mediators responsible for this modulation are still unknown [199]. Based on this study, we hypothesize that during the interaction process between protist and host, there is a secretion system that can secrete effectors to disturb the signal transduction pathway of infected host and to control the apoptosis of host cells.

Our predictions provide useful hypotheses for further studies on exploring pathogenic mechanisms in the host-pathogen interactions. It also has the scientific and clinical implications for prevention and treatment of infectious diseases, as it may provide some guidance for vaccine/drug development. Having said that, it is noted that the EPIYA-motif containing protein does not exist in all intracellular bacteria, and therefore EPIYA-motif mediating interaction is only one type of various host-pathogen mechanisms. Furthermore, our prediction result is based on computation and definitely contains false positives, and thus it requires further experimental validations.

6. SUMMARY AND FUTURE WORK

6.1 Conclusions

In the post-genome era, as an exciting new field, metagenomics offers a powerful lens to explore the world of microorganisms. With the explosion of metagenomic data, it brings many surprises in microbial genomics and microbiology, and is becoming the standard way to help better understand the genome diversity and bacteria dynamics, interactions between bacteria and their hosts/environments, and the pathogenicity of pathogens. The unprecedented amount of genome data not only provides new insights into many aspects of complex microbial communities, but also poses major challenges for computational analysis. In fact, computational methods for massive genomic sequence analysis have restricted the development of microbial genomics.

Among computational problems in metagenomic data analysis, bacterial characterization (i.e. identification and classification) is a very important and essential step of analyzing genomic data. In this dissertation, we systematically studied most computational methods for general bacteria identification. Due to the limitations in theory, most of them only can be successfully applied at genus level identification. However, many practical problems demand high accuracy at species level, sub-species level, sometime even at ecotype level. Furthermore we propose two challenging problems and introduce possible solutions with our methodologies. And we also present some extension work on effector prediction based on our study of *H. pylori*.

In chapter 2, we have reviewed the most widely used bacterial identification algorithms, e.g. sequence search, phylogeny, frequencies of length-N motifs, naïve Bayes classifier, Markov model and SVM. With the improvement of new tools, the accuracy is getting improved. However, the significant drop of identification performance with the decrease of the read length and the lacking of accuracy on species-level identification are the two common drawbacks for all current algorithms. It is crucial to improve the performance in using short reads for bacterial identification on the species or sub-species level.

In chapter 3, we have addressed a specific bacterial identification problem, detections of host-specific bacteria. It means that we are trying to identify the different subtypes/subspecies of one particular bacterium. While current algorithms cannot solve it, we provided promising examples for tackling these issues, which may point a helpful direction to pursue for future studies. In this case, we used entropy difference to detect host-specific genetic markers for *Faecalibacterium*. Then two primers designed based on detected markers have been successfully validated by experiments. The discovery of *Faecalibacterium* 16S rDNA IVS has provided a foundation to design and develop of a poultry feces-specific PCR assay for the rapid determination of poultry fecal pollution in water.

Identification of host-specific bacteria is not only a special issue of bacterial identification, but also a way to discover bacteria-host interaction. In terms of bacteria-host interaction, the mechanism of same pathogen causing different diseases is a very common and challenging problem. Detection of disease-associated sequence markers in pathogenic bacteria, discussed in chapter 4, requires an even higher differentiation power requires even more accuracy in differentiation than in identification of host-specific bacteria. The clinical importance of differentiating disease-associated strains from the nonpathogenic ones calls for more computer scientists to develop new computational methods, inference algorithms, and standard tools to solve these challenging problems.

In chapter 5, based on our knowledge of *H. pylori*, we speculate that the EPIYA-like motif and its phosphorylation, together with its interference of host cells, may be a general mechanism of pathogenesis. Then we used EPIYA-motif-based HMM to search the current protein database, further identify more proteins with the EPIYA motif. Through studying the distribution and features of EPIYA motif in different species and genuses, we could better understand the function of EPIYA motif. Our predictions not only provide useful hypotheses for further studies on exploring pathogenic mechanisms in the host-pathogen interactions, but also have the scientific and clinical implications for prevention and treatment of infectious diseases, as it may provide some guidance for vaccine/drug development.

6.2 Limitations and future works

6.2.1 General 16S rRNA-based classifier

According to our previous research in chapter 3 and 4, the result shows that our method has the potential ability to achieve high identification accuracy on sub-species level or ecotype level. Besides the better accuracy, our approach also has the following advantages:

- We can utilize partial 16s rRNA sequence for classification, and the accuracy will not be affected dramatically.
- We can assign bacterial sequence to new taxonomy at species level, and for some species we can even identify sub-species.
- For two groups of sequences, e.g. from two genera or two species, our approach can tell the most important sites to distinguish those sequence groups.

But as a position-related method, multiple sequence alignment is a necessary pre-processing step and the alignment quality has a significant impact on identification accuracy. As we known, large scaled MSA has the extremely expensive computational complexity Therefore, finding a high-efficient MSA algorithm or modify current algorithm to reduce the computational complexity will be the high priority task for making our method practical.

6.2.2 Fecal pollution tracking with using host-specific genetic markers

Although we successfully detected a genetic marker which is specific for *Faecalibacterium* strains from poultry hosts, it cannot solve the fecal pollution tracking problem completely. We have to find other new markers to distinguish other hosts in practice. At first, we applied the method presented in chapter 3 to *E. coli* genes, but unfortunately the variation between strains from different hosts is not as significant as *Faecalibacterium*. We speculate that the host-specific difference of *E. coli* might be at genome level rather than gene level, so a new method is necessary to introduce more information in genetic marker detection, such as, gene rearrangement. With the dramatic increasing of metagenomic data, another possible solution to find new host-specific markers is that screening all possible bacteria by using an automatic pipeline.

6.2.3 Gastric cancer marker identification and further research on host-pathogen interactions

We already did some studies on host-pathogen interactions and trying to find the cancer-related sequence marker on CagA sequences of *H. pylori*. We know that the molecular mechanism of development of gastric cancer is very complicated and only one pathogen gene cannot dominate the disease. Due to the limitation of data, we have to start our research from the most studied gene, CagA. With the increasing of cancer genome data and improvement of studies of other genes of *H. pylori*, the following further researches become possible and they will help us to better understand the host-pathogen interactions and molecular mechanisms of *H. pylori*-mediated gastric cancer:

- Analyze another important virulence factor, VacA, by using the similar method as we applied to CagA, and detect the cancer-related residues for VacA gene.
- Use both VacA and CagA sequence data for analysis.
- Use both gastric cancer and *H. pylori* genome data for analysis.

REFERENCES

1. Whitman WB, Coleman DC, Wiebe WJ: **Prokaryotes: the unseen majority**. *Proc Natl Acad Sci U S A* 1998, **95**(12):6578-6583.
2. Curtis TP, Sloan WT, Scannell JW: **Estimating prokaryotic diversity and its limits**. *Proc Natl Acad Sci U S A* 2002, **99**(16):10494-10499.
3. Fredrickson JK, Zachara JM, Balkwill DL, Kennedy D, Li SM, Kostandarithes HM, Daly MJ, Romine MF, Brockman FJ: **Geomicrobiology of high-level nuclear waste-contaminated vadose sediments at the hanford site, washington state**. *Appl Environ Microbiol* 2004, **70**(7):4230-4241.
4. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP *et al*: **A core gut microbiome in obese and lean twins**. *Nature* 2009, **457**(7228):480-484.
5. Dinsdale EA, Pantos O, Smriga S, Edwards RA, Angly F, Wegley L, Hatay M, Hall D, Brown E, Haynes M *et al*: **Microbial ecology of four coral atolls in the Northern Line Islands**. *PLoS One* 2008, **3**(2):e1584.
6. Lorenz P, Eck J: **Metagenomics and industrial applications**. *Nat Rev Microbiol* 2005, **3**(6):510-516.
7. Ishige T, Honda K, Shimizu S: **Whole organism biocatalysis**. *Curr Opin Chem Biol* 2005, **9**(2):174-180.
8. Andries K, Verhasselt P, Guillemont J, Gohlmann HW, Neefs JM, Winkler H, Van Gestel J, Timmerman P, Zhu M, Lee E *et al*: **A diarylquinoline drug active on the ATP synthase of Mycobacterium tuberculosis**. *Science* 2005, **307**(5707):223-227.
9. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM *et al*: **Whole-genome random sequencing and assembly of Haemophilus influenzae Rd**. *Science* 1995, **269**(5223):496-512.
10. Nishida H, Kondo S, Nojiri H, Noma K, Oshima K: **Evolutionary mechanisms of microbial genomes**. *Int J Evol Biol* 2011, **2011**:319479.
11. Schloss PD, Handelsman J: **Status of the microbial census**. *Microbiol Mol Biol Rev* 2004, **68**(4):686-691.
12. Petrosino JF, Highlander S, Luna RA, Gibbs RA, Versalovic J: **Metagenomic pyrosequencing and microbial identification**. *Clin Chem* 2009, **55**(5):856-866.
13. Wooley JC, Ye Y: **Metagenomics: Facts and Artifacts, and Computational Challenges***. *J Comput Sci Technol* 2009, **25**(1):71-81.
14. Pallen MJ, Wren BW: **Bacterial pathogenomics**. *Nature* 2007, **449**(7164):835-842.
15. Fricke WF, Rasko DA, Ravel J: **The role of genomics in the identification, prediction, and prevention of biological threats**. *PLoS Biol* 2009, **7**(10):e1000217.

16. Medini D, Serruto D, Parkhill J, Relman DA, Donati C, Moxon R, Falkow S, Rappuoli R: **Microbiology in the post-genomic era**. *Nat Rev Microbiol* 2008, **6**(6):419-430.
17. Welch RA, Burland V, Plunkett G, 3rd, Redford P, Roesch P, Rasko D, Buckles EL, Liou SR, Boutin A, Hackett J *et al*: **Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli**. *Proc Natl Acad Sci U S A* 2002, **99**(26):17020-17024.
18. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI: **The human microbiome project**. *Nature* 2007, **449**(7164):804-810.
19. Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, Gill SR, Nelson KE, Relman DA: **Diversity of the human intestinal microbial flora**. *Science* 2005, **308**(5728):1635-1638.
20. Woese CR, Kandler O, Wheelis ML: **Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya**. *Proc Natl Acad Sci U S A* 1990, **87**(12):4576-4579.
21. Relman DA, Falkow S, LeBoit PE, Perkocha LA, Min KW, Welch DF, Slater LN: **The organism causing bacillary angiomatosis, peliosis hepatis, and fever and bacteremia in immunocompromised patients**. *N Engl J Med* 1991, **324**(21):1514.
22. Winker S, Woese CR: **A definition of the domains Archaea, Bacteria and Eucarya in terms of small subunit ribosomal RNA characteristics**. *Syst Appl Microbiol* 1991, **14**(4):305-310.
23. Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA *et al*: **Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms**. *Proc Natl Acad Sci U S A* 1998, **95**(6):3140-3145.
24. Lin GN, Zhang C, Xu D: **Polytomy identification in microbial phylogenetic reconstruction**. *BMC Systems Biology* 2011, **Submitted**.
25. Bansal AK, Meyer TE: **Evolutionary analysis by whole-genome comparisons**. *J Bacteriol* 2002, **184**(8):2260-2272.
26. Van de Peer Y, Chapelle S, De Wachter R: **A quantitative map of nucleotide substitution rates in bacterial rRNA**. *Nucleic Acids Res* 1996, **24**(17):3381-3391.
27. Peterson DA, Frank DN, Pace NR, Gordon JI: **Metagenomic approaches for defining the pathogenesis of inflammatory bowel diseases**. *Cell Host Microbe* 2008, **3**(6):417-427.
28. Garrity G: **Bergey's Manual of Systematic Bacteriology, Vol. 2 (Parts A, B & C; Three-Volume Set)**: Springer; 2005.
29. Pace NR: **A molecular view of microbial diversity and the biosphere**. *Science* 1997, **276**(5313):734-740.
30. Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar, Buchner A, Lai T, Steppi S, Jobb G *et al*: **ARB: a software environment for sequence data**. *Nucleic Acids Res* 2004, **32**(4):1363-1371.
31. Hugenholtz P: **Exploring prokaryotic diversity in the genomic era**. *Genome Biol* 2002, **3**(2):REVIEWS0003.

32. Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, Kulam-Syed-Mohideen AS, McGarrell DM, Marsh T, Garrity GM *et al*: **The Ribosomal Database Project: improved alignments and new tools for rRNA analysis**. *Nucleic Acids Res* 2009, **37**(Database issue):D141-145.
33. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL: **Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB**. *Appl Environ Microbiol* 2006, **72**(7):5069-5072.
34. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glockner FO: **SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB**. *Nucleic Acids Res* 2007, **35**(21):7188-7196.
35. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R *et al*: **Clustal W and Clustal X version 2.0**. *Bioinformatics* 2007, **23**(21):2947-2948.
36. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: **MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods**. *Mol Biol Evol* 2011.
37. DeSantis TZ, Jr., Hugenholtz P, Keller K, Brodie EL, Larsen N, Piceno YM, Phan R, Andersen GL: **NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes**. *Nucleic Acids Res* 2006, **34**(Web Server issue):W394-399.
38. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput**. *Nucleic Acids Res* 2004, **32**(5):1792-1797.
39. Schloss PD: **The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies**. *PLoS Comput Biol* 2010, **6**(7):e1000844.
40. Baker GC, Smith JJ, Cowan DA: **Review and re-analysis of domain-specific 16S primers**. *J Microbiol Methods* 2003, **55**(3):541-555.
41. Luna RA, Fasciano LR, Jones SC, Boyanton BL, Jr., Ton TT, Versalovic J: **DNA pyrosequencing-based bacterial pathogen identification in a pediatric hospital setting**. *J Clin Microbiol* 2007, **45**(9):2985-2992.
42. Chakravorty S, Helb D, Burday M, Connell N, Alland D: **A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria**. *J Microbiol Methods* 2007, **69**(2):330-339.
43. Crielaard W, Zaura E, Schuller AA, Huse SM, Montijn RC, Keijser BJ: **Exploring the oral microbiota of children at various developmental stages of their dentition in the relation to their oral health**. *BMC Med Genomics* 2011, **4**:22.
44. Wade WG: **Has the use of molecular methods for the characterization of the human oral microbiome changed our understanding of the role of bacteria in the pathogenesis of periodontal disease?** *J Clin Periodontol* 2011, **38 Suppl 11**:7-16.

45. Schmalenberger A, Schwieger F, Tebbe CC: **Effect of primers hybridizing to different evolutionarily conserved regions of the small-subunit rRNA gene in PCR-based microbial community analyses and genetic profiling**. *Appl Environ Microbiol* 2001, **67**(8):3557-3563.

46. Wu GD, Lewis JD, Hoffmann C, Chen YY, Knight R, Bittinger K, Hwang J, Chen J, Berkowsky R, Nessel L *et al*: **Sampling and pyrosequencing methods for characterizing bacterial communities in the human gut using 16S sequence tags**. *BMC Microbiol* 2010, **10**:206.

47. Liu Z, DeSantis TZ, Andersen GL, Knight R: **Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers**. *Nucleic Acids Res* 2008, **36**(18):e120.

48. Claesson MJ, O'Sullivan O, Wang Q, Nikkila J, Marchesi JR, Smidt H, de Vos WM, Ross RP, O'Toole PW: **Comparative analysis of pyrosequencing and a phylogenetic microarray for exploring microbial community structures in the human distal intestine**. *PLoS One* 2009, **4**(8):e6669.

49. Sacchi CT, Whitney AM, Mayer LW, Morey R, Steigerwalt A, Boras A, Weyant RS, Popovic T: **Sequencing of 16S rRNA gene: a rapid tool for identification of Bacillus anthracis**. *Emerg Infect Dis* 2002, **8**(10):1117-1123.

50. Gori F, Folino G, Jetten MS, Marchiori E: **MTR: taxonomic annotation of short metagenomic reads using clustering at multiple taxonomic ranks**. *Bioinformatics* 2011, **27**(2):196-203.

51. Rosen GL, Essinger SD: **Comparison of statistical methods to classify environmental genomic fragments**. *IEEE Trans Nanobioscience* 2010, **9**(4):310-316.

52. Foerstner KU, von Mering C, Hooper SD, Bork P: **Environments shape the nucleotide composition of genomes**. *EMBO Rep* 2005, **6**(12):1208-1213.

53. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**(3):403-410.

54. Wommack KE, Bhavsar J, Ravel J: **Metagenomics: read length matters**. *Appl Environ Microbiol* 2008, **74**(5):1453-1463.

55. Andersson AF, Lindberg M, Jakobsson H, Backhed F, Nyren P, Engstrand L: **Comparative analysis of human gut microbiota by barcoded pyrosequencing**. *PLoS One* 2008, **3**(7):e2836.

56. Dalevi D, Ivanova NN, Mavromatis K, Hooper SD, Szeto E, Hugenholtz P, Kyrpides NC, Markowitz VM: **Annotation of metagenome short reads using proxygenes**. *Bioinformatics* 2008, **24**(16):i7-13.

57. Koski LB, Golding GB: **The closest BLAST hit is often not the nearest neighbor**. *J Mol Evol* 2001, **52**(6):540-542.

58. Pignatelli M, Aparicio G, Blanquer I, Hernandez V, Moya A, Tamames J: **Metagenomics reveals our incomplete knowledge of global diversity**. *Bioinformatics* 2008, **24**(18):2124-2125.

59. Manichanh C, Chapple CE, Frangeul L, Gloux K, Guigo R, Dore J: **A comparison of random sequence reads versus 16S rDNA sequences for estimating the**

biodiversity of a metagenomic library. *Nucleic Acids Res* 2008, **36**(16):5180-5188.

60. Huson DH, Auch AF, Qi J, Schuster SC: **MEGAN analysis of metagenomic data**. *Genome Res* 2007, **17**(3):377-386.

61. Clemente JC, Jansson J, Valiente G: **Flexible taxonomic assignment of ambiguous sequencing reads**. *BMC Bioinformatics* 2011, **12**:8.

62. Clemente JC, Jansson J, Valiente G: **Accurate taxonomic assignment of short pyrosequencing reads**. *Pac Symp Biocomput* 2010:3-9.

63. Vinga S, Almeida J: **Alignment-free sequence comparison-a review**. *Bioinformatics* 2003, **19**(4):513-523.

64. Wang Q, Garrity GM, Tiedje JM, Cole JR: **Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy**. *Appl Environ Microbiol* 2007, **73**(16):5261-5267.

65. Brady A, Salzberg SL: **Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models**. *Nat Methods* 2009, **6**(9):673-676.

66. Kotamarti RM, Hahsler M, Raiford D, McGee M, Dunham MH: **Analyzing taxonomic classification using extensible Markov models**. *Bioinformatics* 2010, **26**(18):2235-2241.

67. McHardy AC, Martin HG, Tsirigos A, Hugenholtz P, Rigoutsos I: **Accurate phylogenetic classification of variable-length DNA fragments**. *Nat Methods* 2007, **4**(1):63-72.

68. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T: **Cytoscape 2.8: new features for data integration and network visualization**. *Bioinformatics* 2011, **27**(3):431-432.

69. Beck D, Settles M, Foster JA: **OTUbase: an R infrastructure package for operational taxonomic unit data**. *Bioinformatics* 2011.

70. Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M: **CAMERA: a community resource for metagenomics**. *PLoS Biol* 2007, **5**(3):e75.

71. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A *et al*: **The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes**. *BMC Bioinformatics* 2008, **9**:386.

72. Gerlach W, Stoye J: **Taxonomic classification of metagenomic shotgun sequences with CARMA3**. *Nucleic Acids Res* 2011.

73. Giongo A, Crabb DB, Davis-Richardson AG, Chauliac D, Mobberley JM, Gano KA, Mukherjee N, Casella G, Roesch LF, Walts B *et al*: **PANGEA: pipeline for analysis of next generation amplicons**. *ISME J* 2010, **4**(7):852-861.

74. Horton M, Bodenhausen N, Bergelson J: **MARTA: a suite of Java-based tools for assigning taxonomic status to DNA sequences**. *Bioinformatics* 2010, **26**(4):568-569.

75. Devulder G, Perriere G, Baty F, Flandrois JP: **BIBI, a bioinformatics bacterial identification tool**. *J Clin Microbiol* 2003, **41**(4):1785-1787.

76. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI *et al*: **QIIME allows analysis of high-throughput community sequencing data**. *Nat Methods* 2010, **7**(5):335-336.

77. Wu D, Hartman A, Ward N, Eisen JA: **An automated phylogenetic tree-based small subunit rRNA taxonomy and alignment pipeline (STAP)**. *PLoS One* 2008, **3**(7):e2566.

78. Kosakovsky Pond S, Wadhawan S, Chiaromonte F, Ananda G, Chung WY, Taylor J, Nekrutenko A: **Windshield splatter analysis with the Galaxy metagenomic pipeline**. *Genome Res* 2009, **19**(11):2144-2153.

79. Rosen GL, Reichenberger ER, Rosenfeld AM: **NBC: the Naive Bayes Classification tool webserver for taxonomic classification of metagenomic reads**. *Bioinformatics* 2011, **27**(1):127-129.

80. Schatz MC: **CloudBurst: highly sensitive read mapping with MapReduce**. *Bioinformatics* 2009, **25**(11):1363-1369.

81. Underwood A, Green J: **Call for a quality standard for sequence-based assays in clinical microbiology: necessity for quality assessment of sequences used in microbial identification and typing**. *J Clin Microbiol* 2011, **49**(1):23-26.

82. Teng JL, Yeung MY, Yue G, Au-Yeung RK, Yeung EY, Fung AM, Tse H, Yuen KY, Lau SK, Woo PC: **In silico analysis of 16S ribosomal RNA gene sequencing based methods for identification of medically important aerobic Gram-negative bacteria**. *J Med Microbiol* 2011.

83. Woo PC, Teng JL, Yeung JM, Tse H, Lau SK, Yuen KY: **Automated identification of medically important bacteria by 16S rRNA gene sequencing using a novel comprehensive database 16SpathDB**. *J Clin Microbiol* 2011.

84. Lecomte J, St-Arnaud M, Hijri M: **Isolation and identification of soil bacteria growing at the expense of arbuscular mycorrhizal fungi**. *FEMS Microbiol Lett* 2011, **317**(1):43-51.

85. Schloss PD, Handelsman J: **Toward a census of bacteria in soil**. *PLoS Comput Biol* 2006, **2**(7):e92.

86. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, Fernandes GR, Tap J, Bruls T, Batto JM *et al*: **Enterotypes of the human gut microbiome**. *Nature* 2011.

87. Fierer N, Lauber CL, Zhou N, McDonald D, Costello EK, Knight R: **Forensic identification using skin bacterial communities**. *Proc Natl Acad Sci U S A* 2010, **107**(14):6477-6481.

88. Janda JM, Abbott SL: **16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls**. *J Clin Microbiol* 2007, **45**(9):2761-2764.

89. Silverman AP, Kool ET: **Quenched autoligation probes allow discrimination of live bacterial species by single nucleotide differences in rRNA**. *Nucleic Acids Res* 2005, **33**(15):4978-4986.

90. Robertson GA, Thiruvenkataswamy V, Shilling H, Price EP, Huygens F, Henskens FA, Giffard PM: **Identification and interrogation of highly informative single**

nucleotide polymorphism sets defined by bacterial multilocus sequence typing databases. *J Med Microbiol* 2004, **53**(Pt 1):35-45.

91. Lu J, Santo Domingo J, Shanks OC: **Identification of chicken-specific fecal microbial sequences using a metagenomic approach**. *Water Res* 2007, **41**(16):3561-3574.

92. Yoder JS, Centers for Disease C, Prevention: **Surveillance for waterborne disease and outbreaks associated with recreational water use and other aquatic facility-associated health events-- United States, 2005-2006 ; and, Surveillance for waterborne disease and outbreaks associated with drinking water and water not intended for drinking-- United States, 2005-2006**. Atlanta, GA: Coordinating Center for Health Information and Service, Centers for Disease Control and Prevention (CDC), U.S. Dept. of Health and Human Services; 2008.

93. Xu J, Gordon JI: **Honor thy symbionts**. *Proc Natl Acad Sci U S A* 2003, **100**(18):10452-10459.

94. Ahmed W, Neller R, Katouli M: **Host species-specific metabolic fingerprint database for enterococci and Escherichia coli and its application to identify sources of fecal contamination in surface waters**. *Appl Environ Microbiol* 2005, **71**(8):4461-4468.

95. Bernhard AE, Field KG: **A PCR assay To discriminate human and ruminant feces on the basis of host differences in Bacteroides-Prevotella genes encoding 16S rRNA**. *Appl Environ Microbiol* 2000, **66**(10):4571-4574.

96. Fong TT, Griffin DW, Lipp EK: **Molecular assays for targeting human and bovine enteric viruses in coastal waters and their application for library-independent source tracking**. *Appl Environ Microbiol* 2005, **71**(4):2070-2078.

97. Carson CA, Christiansen JM, Yampara-Iquise H, Benson VW, Baffaut C, Davis JV, Broz RR, Kurtz WB, Rogers WM, Fales WH: **Specificity of a Bacteroides thetaiotaomicron marker for human feces**. *Appl Environ Microbiol* 2005, **71**(8):4945-4949.

98. Bonjoch X, Balleste E, Blanch AR: **Enumeration of bifidobacterial populations with selective media to determine the source of waterborne fecal pollution**. *Water Res* 2005, **39**(8):1621-1627.

99. Sorensen DL, Eberl SG, Dicksa RA: **Clostridium perfringens as a point source indicator in non-point polluted streams**. *Water Research* 1989, **23**(2):191-197.

100. Marti R, Dabert P, Ziebal C, Pourcher AM: **Evaluation of Lactobacillus sobrius/L. amylovorus as a new microbial marker of pig manure**. *Appl Environ Microbiol* 2010, **76**(5):1456-1461.

101. Ufnar JA, Wang SY, Ufnar DF, Ellender RD: **Methanobrevibacter ruminantium as an indicator of domesticated-ruminant fecal pollution in surface waters**. *Appl Environ Microbiol* 2007, **73**(21):7118-7121.

102. Zheng G, Yampara-Iquise H, Jones JE, Andrew Carson C: **Development of Faecalibacterium 16S rRNA gene marker for identification of human faeces**. *J Appl Microbiol* 2009, **106**(2):634-641.

103. Wiggins BA, Cash PW, Creamer WS, Dart SE, Garcia PP, Gerecke TM, Han J, Henry BL, Hoover KB, Johnson EL *et al*: **Use of antibiotic resistance analysis for representativeness testing of multiwatershed libraries**. *Appl Environ Microbiol* 2003, **69**(6):3399-3405.

104. Parveen S, Portier KM, Robinson K, Edmiston L, Tamplin ML: **Discriminant analysis of ribotype profiles of Escherichia coli for differentiating human and nonhuman sources of fecal pollution**. *Appl Environ Microbiol* 1999, **65**(7):3142-3147.

105. Dombek PE, Johnson LK, Zimmerley ST, Sadowsky MJ: **Use of repetitive DNA sequences and the PCR To differentiate Escherichia coli isolates from human and animal sources**. *Appl Environ Microbiol* 2000, **66**(6):2572-2577.

106. Field KG, Samadpour M: **Fecal source tracking, the indicator paradigm, and managing water quality**. *Water Res* 2007, **41**(16):3517-3538.

107. Ksoll WB, Ishii S, Sadowsky MJ, Hicks RE: **Presence and sources of fecal coliform bacteria in epilithic periphyton communities of Lake Superior**. *Appl Environ Microbiol* 2007, **73**(12):3771-3778.

108. Roslev P, Bukh AS: **State of the art molecular markers for fecal pollution source tracking in water**. *Appl Microbiol Biotechnol* 2011, **89**(5):1341-1355.

109. Stoeckel DM, Harwood VJ: **Performance, design, and analysis in microbial source tracking studies**. *Appl Environ Microbiol* 2007, **73**(8):2405-2415.

110. Ahmed W, Stewart J, Gardner T, Powell D, Brooks P, Sullivan D, Tindale N: **Sourcing faecal pollution: a combination of library-dependent and library-independent methods to identify human faecal pollution in non-sewered catchments**. *Water Res* 2007, **41**(16):3771-3779.

111. Dickerson JW, Jr., Hagedorn C, Hassall A: **Detection and remediation of human-origin pollution at two public beaches in Virginia using multiple source tracking methods**. *Water Res* 2007, **41**(16):3758-3770.

112. McQuaig SM, Scott TM, Lukasik JO, Paul JH, Harwood VJ: **Quantification of human polyomaviruses JC Virus and BK Virus by TaqMan quantitative PCR and comparison to other water quality indicators in water and fecal samples**. *Appl Environ Microbiol* 2009, **75**(11):3379-3388.

113. Doyle MP, Erickson MC: **Reducing the carriage of foodborne pathogens in livestock and poultry**. *Poult Sci* 2006, **85**(6):960-973.

114. Weidhaas JL, Macbeth TW, Olsen RL, Harwood VJ: **Correlation of quantitative PCR for a poultry-specific brevibacterium marker gene with bacterial and chemical indicators of water pollution in a watershed impacted by land application of poultry litter**. *Appl Environ Microbiol* 2011, **77**(6):2094-2102.

115. Kortbaoui R, Locas A, Imbeau M, Payment P, Villemur R: **Universal mitochondrial PCR combined with species-specific dot-blot assay as a source-tracking method of human, bovine, chicken, ovine, and porcine in fecal-contaminated surface water**. *Water Res* 2009, **43**(7):2002-2010.

116. Duncan SH, Hold GL, Harmsen HJ, Stewart CS, Flint HJ: **Growth requirements and fermentation products of Fusobacterium prausnitzii, and a proposal to**

reclassify it as Faecalibacterium prausnitzii gen. nov., comb. nov. *Int J Syst Evol Microbiol* 2002, **52**(Pt 6):2141-2146.

117. Tap J, Mondot S, Levenez F, Pelletier E, Caron C, Furet JP, Ugarte E, Munoz-Tamayo R, Paslier DL, Nalin R *et al*: **Towards the human intestinal microbiota phylogenetic core**. *Environ Microbiol* 2009, **11**(10):2574-2584.

118. Dowd SE, Callaway TR, Wolcott RD, Sun Y, McKeehan T, Hagevoort RG, Edrington TS: **Evaluation of the bacterial diversity in the feces of cattle using 16S rDNA bacterial tag-encoded FLX amplicon pyrosequencing (bTEFAP)**. *BMC Microbiol* 2008, **8**:125.

119. Leser TD, Amenuvor JZ, Jensen TK, Lindecrona RH, Boye M, Moller K: **Culture-independent analysis of gut bacteria: the pig gastrointestinal tract microbiota revisited**. *Appl Environ Microbiol* 2002, **68**(2):673-690.

120. Zhu XY, Zhong T, Pandya Y, Joerger RD: **16S rRNA-based analysis of microbiota from the cecum of broiler chickens**. *Appl Environ Microbiol* 2002, **68**(1):124-137.

121. Price MN, Dehal PS, Arkin AP: **FastTree 2--approximately maximum-likelihood trees for large alignments**. *PLoS One* 2010, **5**(3):e9490.

122. Reva B, Antipin Y, Sander C: **Determinants of protein function revealed by combinatorial entropy optimization**. *Genome Biol* 2007, **8**(11):R232.

123. Zhang C, Xu S, Xu D: **Detection and application of CagA sequence markers for assessing risk factor of gastric cancer caused by Helicobacter pylori**. In: *Bioinformatics and Biomedicine (BIBM), 2010 IEEE International Conference on: 18-21 Dec. 2010 2010*; 2010: 485-488.

124. Evguenieva-Hackenberg E: **Bacterial ribosomal RNA in pieces**. *Mol Microbiol* 2005, **57**(2):318-325.

125. Pronk LM, Sanderson KE: **Intervening sequences in rrl genes and fragmentation of 23S rRNA in genera of the family Enterobacteriaceae**. *J Bacteriol* 2001, **183**(19):5782-5787.

126. Pei AY, Oberdorf WE, Nossa CW, Agarwal A, Chokshi P, Gerz EA, Jin Z, Lee P, Yang L, Poles M *et al*: **Diversity of 16S rRNA genes within individual prokaryotic genomes**. *Appl Environ Microbiol* 2010, **76**(12):3886-3897.

127. Williams ML, Lawrence ML: **Verification of an Edwardsiella ictaluri-specific diagnostic PCR**. *Lett Appl Microbiol* 2010, **50**(2):153-157.

128. Dick LK, Bernhard AE, Brodeur TJ, Santo Domingo JW, Simpson JM, Walters SP, Field KG: **Host distributions of uncultivated fecal Bacteroidales bacteria reveal genetic markers for fecal source identification**. *Appl Environ Microbiol* 2005, **71**(6):3184-3191.

129. D'Elia TV, Cooper CR, Johnston CG: **Source tracking of Escherichia coli by 16S-23S intergenic spacer region denaturing gradient gel electrophoresis (DGGE) of the rrnB ribosomal operon**. *Can J Microbiol* 2007, **53**(10):1174-1184.

130. Shanks OC, Domingo JW, Lu J, Kelty CA, Graham JE: **Identification of bacterial DNA markers for the detection of human fecal pollution in water**. *Appl Environ Microbiol* 2007, **73**(8):2416-2422.

131. Yampara-Iquise H, Zheng G, Jones JE, Carson CA: **Use of a Bacteroides thetaiotaomicron-specific alpha-1-6, mannanase quantitative PCR to detect human faecal pollution in water**. *J Appl Microbiol* 2008, **105**(5):1686-1693.

132. Hacker J, Hentschel U, Dobrindt U: **Prokaryotic chromosomes and disease**. *Science* 2003, **301**(5634):790-793.

133. Ullman TA, Itzkowitz SH: **Intestinal inflammation and cancer**. *Gastroenterology* 2011, **140**(6):1807-1816 e1801.

134. Round JL, Mazmanian SK: **The gut microbiota shapes intestinal immune responses during health and disease**. *Nat Rev Immunol* 2009, **9**(5):313-323.

135. Franco AT, Friedman DB, Nagy TA, Romero-Gallo J, Krishna U, Kendall A, Israel DA, Tegtmeyer N, Washington MK, Peek RM, Jr.: **Delineation of a carcinogenic Helicobacter pylori proteome**. *Mol Cell Proteomics* 2009, **8**(8):1947-1958.

136. IARC: **Schistosomes, liver flukes and *Helicobacter pylori*. in "Monographs on the evaluation of carcinogenic risks to humans" IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. Lyon, 7-14 June 1994**. 1994, **61**:1-241.

137. Linz B, Balloux F, Moodley Y, Manica A, Liu H, Roumagnac P, Falush D, Stamer C, Prugnolle F, van der Merwe SW *et al*: **An African origin for the intimate association between humans and *Helicobacter pylori***. *Nature* 2007, **445**(7130):915-918.

138. Suerbaum S, Michetti P: ***Helicobacter pylori* infection**. *N Engl J Med* 2002, **347**(15):1175-1186.

139. Covacci A, Censini S, Bugnoli M, Petracca R, Burroni D, Macchia G, Massone A, Papini E, Xiang Z, Figura N *et al*: **Molecular characterization of the 128-kDa immunodominant antigen of *Helicobacter pylori* associated with cytotoxicity and duodenal ulcer**. *Proc Natl Acad Sci U S A* 1993, **90**(12):5791-5795.

140. Ernst PB, Gold BD: **The disease spectrum of *Helicobacter pylori*: the immunopathogenesis of gastroduodenal ulcer and gastric cancer**. *Annu Rev Microbiol* 2000, **54**:615-640.

141. Blaser MJ, Perez-Perez GI, Kleanthous H, Cover TL, Peek RM, Chyou PH, Stemmermann GN, Nomura A: **Infection with *Helicobacter pylori* strains possessing cagA is associated with an increased risk of developing adenocarcinoma of the stomach**. *Cancer Res* 1995, **55**(10):2111-2115.

142. Peek RM, Jr, Blaser MJ: ***Helicobacter pylori* and gastrointestinal tract adenocarcinomas**. *Nat Rev Cancer* 2002, **2**(1):28-37.

143. Uemura N, Okamoto S, Yamamoto S, Matsumura N, Yamaguchi S, Yamakido M, Taniyama K, Sasaki N, Schlemper RJ: ***Helicobacter pylori* infection and the development of gastric cancer**. *N Engl J Med* 2001, **345**(11):784-789.

144. Gwack J, Shin A, Kim CS, Ko KP, Kim Y, Jun JK, Bae J, Park SK, Hong YC, Kang D *et al*: **CagA-producing *Helicobacter pylori* and increased risk of gastric cancer: a nested case-control study in Korea**. *Br J Cancer* 2006, **95**(5):639-641.

145. Xie XF, Ito M, Sumii M, Tanaka S, Yoshihara M, Chayama K: ***Helicobacter pylori*-associated gastritis is related to babA2 expression without heterogeneity of

the 3' region of the cagA genotype in gastric biopsy specimens. *Pathobiology* 2007, **74**(5):309-316.

146. Higashi H, Tsutsumi R, Fujita A, Yamazaki S, Asaka M, Azuma T, Hatakeyama M: **Biological activity of the *Helicobacter pylori* virulence factor CagA is determined by variation in the tyrosine phosphorylation sites**. *Proc Natl Acad Sci U S A* 2002, **99**(22):14428-14433.

147. Odenbreit S, Puls J, Sedlmaier B, Gerland E, Fischer W, Haas R: **Translocation of *Helicobacter pylori* CagA into gastric epithelial cells by type IV secretion**. *Science* 2000, **287**(5457):1497-1500.

148. Selbach M, Moese S, Hauck CR, Meyer TF, Backert S: **Src is the kinase of the *Helicobacter pylori* CagA protein in vitro and in vivo**. *J Biol Chem* 2002, **277**(9):6775-6778.

149. Stein M, Bagnoli F, Halenbeck R, Rappuoli R, Fantl WJ, Covacci A: **c-Src/Lyn kinases activate *Helicobacter pylori* CagA through tyrosine phosphorylation of the EPIYA motifs**. *Mol Microbiol* 2002, **43**(4):971-980.

150. Higashi H, Tsutsumi R, Muto S, Sugiyama T, Azuma T, Asaka M, Hatakeyama M: **SHP-2 tyrosine phosphatase as an intracellular target of *Helicobacter pylori* CagA protein**. *Science* 2002, **295**(5555):683-686.

151. Argent RH, Kidd M, Owen RJ, Thomas RJ, Limb MC, Atherton JC: **Determinants and consequences of different levels of CagA phosphorylation for clinical isolates of *Helicobacter pylori***. *Gastroenterology* 2004, **127**(2):514-523.

152. Fu H, Hu Z, Wen J, Wang K, Liu Y: **TGF-beta promotes invasion and metastasis of gastric cancer cells by increasing fascin1 expression via ERK and JNK signal pathways**. *Acta Biochim Biophys Sin (Shanghai)* 2009, **41**(8):648-656.

153. Amieva MR, Vogelmann R, Covacci A, Tompkins LS, Nelson WJ, Falkow S: **Disruption of the epithelial apical-junctional complex by *Helicobacter pylori* CagA**. *Science* 2003, **300**(5624):1430-1434.

154. Churin Y, Al-Ghoul L, Kepp O, Meyer TF, Birchmeier W, Naumann M: ***Helicobacter pylori* CagA protein targets the c-Met receptor and enhances the motogenic response**. *J Cell Biol* 2003, **161**(2):249-255.

155. Seo JH, Lim JW, Kim H, Kim KH: ***Helicobacter pylori* in a Korean isolate activates mitogen-activated protein kinases, AP-1, and NF-kappaB and induces chemokine expression in gastric epithelial AGS cells**. *Lab Invest* 2004, **84**(1):49-62.

156. Tsutsumi R, Takahashi A, Azuma T, Higashi H, Hatakeyama M: **Focal adhesion kinase is a substrate and downstream effector of SHP-2 complexed with *Helicobacter pylori* CagA**. *Mol Cell Biol* 2006, **26**(1):261-276.

157. Satomi S, Yamakawa A, Matsunaga S, Masaki R, Inagaki T, Okuda T, Suto H, Ito Y, Yamazaki Y, Kuriyama M *et al*: **Relationship between the diversity of the cagA gene of *Helicobacter pylori* and gastric cancer in Okinawa, Japan**. *J Gastroenterol* 2006, **41**(7):668-673.

158. Jones KR, Joo YM, Jang S, Yoo YJ, Lee HS, Chung IS, Olsen CH, Whitmire JM, Merrell DS, Cha JH: **Polymorphism in the CagA EPIYA Motif Impacts Development of Gastric Cancer**. *J Clin Microbiol* 2009.

159. McClain MS, Shaffer CL, Israel DA, Peek RM, Jr., Cover TL: **Genome sequence analysis of *Helicobacter pylori* strains associated with gastric ulceration and gastric cancer**. *BMC Genomics* 2009, **10**:3.

160. Yamaoka Y, El-Zimaity HM, Gutierrez O, Figura N, Kim JG, Kodama T, Kashima K, Graham DY: **Relationship between the cagA 3' repeat region of *Helicobacter pylori*, gastric histology, and susceptibility to low pH**. *Gastroenterology* 1999, **117**(2):342-349.

161. Lai YP, Yang JC, Lin TZ, Wang JT, Lin JT: **CagA tyrosine phosphorylation in gastric epithelial cells caused by *Helicobacter pylori* in patients with gastric adenocarcinoma**. *Helicobacter* 2003, **8**(3):235-243.

162. Naito M, Yamazaki T, Tsutsumi R, Higashi H, Onoe K, Yamazaki S, Azuma T, Hatakeyama M: **Influence of EPIYA-repeat polymorphism on the phosphorylation-dependent biological activity of *Helicobacter pylori* CagA**. *Gastroenterology* 2006, **130**(4):1181-1190.

163. Xia Y, Yamaoka Y, Zhu Q, Matha I, Gao X: **A comprehensive sequence and disease correlation analyses for the C-terminal region of CagA protein of *Helicobacter pylori***. *PLoS One* 2009, **4**(11):e7736.

164. Crooks GE, Hon G, Chandonia JM, Brenner SE: **WebLogo: a sequence logo generator**. *Genome Res* 2004, **14**(6):1188-1190.

165. Joachims T: **Making large-scale support vector machine learning practical**. In: *Advances in Kernel Methods: Support Vector Machines.* Edited by Schölkopf: MIT Press, Cambridge, MA; 1999.

166. Ren S, Higashi H, Lu H, Azuma T, Hatakeyama M: **Structural basis and functional consequence of *Helicobacter pylori* CagA multimerization in cells**. *J Biol Chem* 2006, **281**(43):32344-32352.

167. Lu HS, Saito Y, Umeda M, Murata-Kamiya N, Zhang HM, Higashi H, Hatakeyama M: **Structural and functional diversity in the PAR1b/MARK2-binding region of *Helicobacter pylori* CagA**. *Cancer Sci* 2008, **99**(10):2004-2011.

168. Sicinschi LA, Correa P, Peek RM, Camargo MC, Piazuelo MB, Romero-Gallo J, Hobbs SS, Krishna U, Delgado A, Mera R *et al*: **CagA C-terminal variations in *Helicobacter pylori* strains from Colombian patients with gastric precancerous lesions**. *Clin Microbiol Infect* 2010, **16**(4):369-378.

169. Eddy SR: **Profile hidden Markov models**. *Bioinformatics* 1998, **14**(9):755-763.

170. Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM: **Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008**. *International journal of cancer Journal international du cancer* 2010, **127**(12):2893-2917.

171. Hansson LE, Nyren O, Hsing AW, Bergstrom R, Josefsson S, Chow WH, Fraumeni JF, Jr., Adami HO: **The risk of stomach cancer in patients with gastric or duodenal ulcer disease**. *N Engl J Med* 1996, **335**(4):242-249.

172. Zhang Z: **The risk of gastric cancer in patients with duodenal and gastric ulcer: research progresses and clinical implications**. *Journal of gastrointestinal cancer* 2007, **38**(1):38-45.

173. Stavrinides J, McCann HC, Guttman DS: **Host-pathogen interplay and the evolution of bacterial effectors**. *Cell Microbiol* 2008, **10**(2):285-292.

174. Cambronne ED, Roy CR: **Recognition and delivery of effector proteins into eukaryotic cells by bacterial secretion systems**. *Traffic* 2006, **7**(8):929-939.

175. Cossart P, Boquet P, Normark S, Rappuoli R: **Cellular microbiology emerging**. *Science* 1996, **271**(5247):315-316.

176. Niebuhr K, Dramsi S: **EMBO-EBNIC workshop on cellular microbiology 'Host cell-pathogen interactions in infectious disease'**. *Cell Microbiol* 1999, **1**(1):79-84.

177. Maeda S, Mentis AF: **Pathogenesis of Helicobacter pylori infection**. *Helicobacter* 2007, **12 Suppl 1**:10-14.

178. Mimuro H, Suzuki T, Tanaka J, Asahi M, Haas R, Sasakawa C: **Grb2 is a key mediator of helicobacter pylori CagA protein activities**. *Mol Cell* 2002, **10**(4):745-755.

179. Hatakeyama M: **SagA of CagA in Helicobacter pylori pathogenesis**. *Curr Opin Microbiol* 2008, **11**(1):30-37.

180. Tammer I, Brandt S, Hartig R, Konig W, Backert S: **Activation of Abl by Helicobacter pylori: a novel kinase for CagA and crucial mediator of host cell scattering**. *Gastroenterology* 2007, **132**(4):1309-1319.

181. Saadat I, Higashi H, Obuse C, Umeda M, Murata-Kamiya N, Saito Y, Lu H, Ohnishi N, Azuma T, Suzuki A *et al*: **Helicobacter pylori CagA targets PAR1/MARK kinase to disrupt epithelial cell polarity**. *Nature* 2007, **447**(7142):330-333.

182. Ijdo JW, Carlson AC, Kennedy EL: **Anaplasma phagocytophilum AnkA is tyrosine-phosphorylated at EPIYA motifs and recruits SHP-1 during early infection**. *Cell Microbiol* 2007.

183. Lin M, den Dulk-Ras A, Hooykaas PJ, Rikihisa Y: **Anaplasma phagocytophilum AnkA secreted by type IV secretion system is tyrosine phosphorylated by Abl-1 to facilitate infection**. *Cell Microbiol* 2007, **9**(11):2644-2657.

184. Dehio C: **Infection-associated type IV secretion systems of Bartonella and their diverse roles in host cell interaction**. *Cell Microbiol* 2008.

185. Schulein R, Guye P, Rhomberg TA, Schmid MC, Schroder G, Vergunst AC, Carena I, Dehio C: **A bipartite signal mediates the transfer of type IV secretion substrates of Bartonella henselae into human cells**. *Proc Natl Acad Sci U S A* 2005, **102**(3):856-861.

186. Schmid MC, Schulein R, Dehio M, Denecker G, Carena I, Dehio C: **The VirB type IV secretion system of Bartonella henselae mediates invasion, proinflammatory activation and antiapoptotic protection of endothelial cells**. *Mol Microbiol* 2004, **52**(1):81-92.

187. Backert S, Meyer TF: **Type IV secretion systems and their effectors in bacterial pathogenesis**. *Curr Opin Microbiol* 2006, **9**(2):207-217.

188. Poppe M, Feller SM, Romer G, Wessler S: **Phosphorylation of Helicobacter pylori CagA by c-Abl leads to cell motility**. *Oncogene* 2006.

189. Tsutsumi R, Higashi H, Higuchi M, Okada M, Hatakeyama M: **Attenuation of Helicobacter pylori CagA x SHP-2 signaling by interaction between CagA and C-terminal Src kinase**. *J Biol Chem* 2003, **278**(6):3664-3670.

190. Schulein R, Dehio C: **The VirB/VirD4 type IV secretion system of Bartonella is essential for establishing intraerythrocytic infection**. *Mol Microbiol* 2002, **46**(4):1053-1067.

191. Phillips N, Hayward RD, Koronakis V: **Phosphorylation of the enteropathogenic E. coli receptor by the Src-family kinase c-Fyn triggers actin pedestal formation**. *Nat Cell Biol* 2004, **6**(7):618-625.

192. Blasutig IM, New LA, Thanabalasuriar A, Dayarathna TK, Goudreault M, Quaggin SE, Li SS, Gruenheid S, Jones N, Pawson T: **Phosphorylated YDXV motifs and Nck SH2/SH3 adaptors act cooperatively to induce actin reorganization**. *Mol Cell Biol* 2008, **28**(6):2035-2046.

193. Gruenheid S, DeVinney R, Bladt F, Goosney D, Gelkop S, Gish GD, Pawson T, Finlay BB: **Enteropathogenic E. coli Tir binds Nck to initiate actin pedestal formation in host cells**. *Nat Cell Biol* 2001, **3**(9):856-859.

194. Jewett TJ, Dooley CA, Mead DJ, Hackstadt T: **Chlamydia trachomatis tarp is phosphorylated by src family tyrosine kinases**. *Biochem Biophys Res Commun* 2008, **371**(2):339-344.

195. Swanson KA, Crane DD, Caldwell HD: **Chlamydia trachomatis species-specific induction of ezrin tyrosine phosphorylation functions in pathogen entry**. *Infect Immun* 2007, **75**(12):5669-5677.

196. Clifton DR, Fields KA, Grieshaber SS, Dooley CA, Fischer ER, Mead DJ, Carabeo RA, Hackstadt T: **A chlamydial type III translocated protein is tyrosine-phosphorylated at the site of entry and associated with recruitment of actin**. *Proc Natl Acad Sci U S A* 2004, **101**(27):10166-10171.

197. Bereswill S, Kist M: **Recent developments in Campylobacter pathogenesis**. *Curr Opin Infect Dis* 2003, **16**(5):487-491.

198. Gregory DJ, Olivier M: **Subversion of host cell signalling by the protozoan parasite Leishmania**. *Parasitology* 2005, **130 Suppl**:S27-35.

199. Carmen JC, Sinai AP: **Suicide prevention: disruption of apoptotic pathways by protozoan parasites**. *Mol Microbiol* 2007, **64**(4):904-916.

200. Blatch GL, Lassle M: **The tetratricopeptide repeat: a structural motif mediating protein-protein interactions**. *Bioessays* 1999, **21**(11):932-939.

201. Li J, Mahajan A, Tsai MD: **Ankyrin repeat: a unique motif mediating protein-protein interactions**. *Biochemistry* 2006, **45**(51):15168-15178.

202. Cavicchioli R, Curmi PM, Saunders N, Thomas T: **Pathogenic archaea: do they exist?** *Bioessays* 2003, **25**(11):1119-1128.

203. Lepp PW, Brinig MM, Ouverney CC, Palm K, Armitage GC, Relman DA: **Methanogenic Archaea and human periodontal disease**. *Proc Natl Acad Sci U S A* 2004, **101**(16):6176-6181.

204. Botham CM, Wandler AM, Guillemin K: **A transgenic Drosophila model demonstrates that the Helicobacter pylori CagA protein functions as a eukaryotic Gab adaptor**. *PLoS Pathog* 2008, **4**(5):e1000064.

205. Calva E, Oropeza R: **Two-component signal transduction systems, environmental signals, and virulence**. *Microb Ecol* 2006, **51**(2):166-176.

206. Jasmer DP, Goverse A, Smant G: **Parasitic nematode interactions with mammals and plants**. *Annu Rev Phytopathol* 2003, **41**:245-270.

207. Xu SF, Zhang GX, Shi RH, Hao B, Miao Y: **polymorphism of variable region of CagA protein**. *Chin J Gastroenterol* 2007, **12**(06):357-361.

208. Drolet R, Larochelle D, Gebhart CJ: **Proliferative enteritis associated with Lawsonia intracellularis (ileal symbiont intracellularis) in white-tailed deer**. *J Vet Diagn Invest* 1996, **8**(2):250-253.

209. Horiuchi N, Watarai M, Kobayashi Y, Omata Y, Furuoka H: **Proliferative enteropathy involving Lawsonia intracellularis infection in rabbits (Oryctlagus cuniculus)**. *J Vet Med Sci* 2008, **70**(4):389-392.

210. McOrist S, Jasni S, Mackie RA, Berschneider HM, Rowland AC, Lawson GH: **Entry of the bacterium ileal symbiont intracellularis into cultured enterocytes and its subsequent release**. *Res Vet Sci* 1995, **59**(3):255-260.

211. Smith DG, Lawson GH: **Lawsonia intracellularis: getting inside the pathogenesis of proliferative enteropathy**. *Vet Microbiol* 2001, **82**(4):331-345.

212. Kroll JJ, Roof MB, Hoffman LJ, Dickson JS, Harris DL: **Proliferative enteropathy: a global enteric disease of pigs caused by Lawsonia intracellularis**. *Anim Health Res Rev* 2005, **6**(2):173-197.

213. Alberdi MP, Watson E, McAllister GE, Harris JD, Paxton EA, Thomson JR, Smith DG: **Expression by Lawsonia intracellularis of type III secretion system components during infection**. *Vet Microbiol* 2009, **139**(3-4):298-303.

214. Lin M, Rikihisa Y: **Obligatory intracellular parasitism by Ehrlichia chaffeensis and Anaplasma phagocytophilum involves caveolae and glycosylphosphatidylinositol-anchored proteins**. *Cell Microbiol* 2003, **5**(11):809-820.

215. Iturbe-Ormaetxe I, Burke GR, Riegler M, O'Neill SL: **Distribution, expression, and motif variability of ankyrin domain genes in Wolbachia pipientis**. *J Bacteriol* 2005, **187**(15):5136-5145.

216. May BJ, Zhang Q, Li LL, Paustian ML, Whittam TS, Kapur V: **Complete genomic sequence of Pasteurella multocida, Pm70**. *Proc Natl Acad Sci U S A* 2001, **98**(6):3460-3465.

217. Clantin B, Hodak H, Willery E, Locht C, Jacob-Dubuisson F, Villeret V: **The crystal structure of filamentous hemagglutinin secretion domain and its implications for the two-partner secretion pathway**. *Proc Natl Acad Sci U S A* 2004, **101**(16):6194-6199.

218. Inatsuka CS, Julio SM, Cotter PA: **Bordetella filamentous hemagglutinin plays a critical role in immunomodulation, suggesting a mechanism for host specificity**. *Proc Natl Acad Sci U S A* 2005, **102**(51):18578-18583.

219. Juhas M, Crook DW, Dimopoulou ID, Lunter G, Harding RM, Ferguson DJ, Hood DW: **Novel type IV secretion system involved in propagation of genomic islands**. *J Bacteriol* 2007, **189**(3):761-771.

220. Ahmed HJ, Johansson C, Svensson LA, Ahlman K, Verdrengh M, Lagergard T: **In vitro and in vivo interactions of Haemophilus ducreyi with host phagocytes**. *Infect Immun* 2002, **70**(2):899-908.

221. Cortes-Bratti X, Chaves-Olarte E, Lagergard T, Thelestam M: **The cytolethal distending toxin from the chancroid bacterium Haemophilus ducreyi induces cell-cycle arrest in the G2 phase**. *J Clin Invest* 1999, **103**(1):107-115.

222. Svensson LA, Henning P, Lagergard T: **The cytolethal distending toxin of Haemophilus ducreyi inhibits endothelial cell proliferation**. *Infect Immun* 2002, **70**(5):2665-2669.

223. Gomis SM, Godson DL, Wobeser GA, Potter AA: **Intracellular survival of Haemophilus somnus in bovine blood monocytes and alveolar macrophages**. *Microb Pathog* 1998, **25**(5):227-235.

224. Lederer JA, Brown JF, Czuprynski CJ: **"Haemophilus somnus," a facultative intracellular pathogen of bovine mononuclear phagocytes**. *Infect Immun* 1987, **55**(2):381-387.

225. Howard MD, Boone JH, Buechner-Maxwell V, Schurig GG, Inzana TJ: **Inhibition of bovine macrophage and polymorphonuclear leukocyte superoxide anion production by Haemophilus somnus**. *Microb Pathog* 2004, **37**(5):263-271.

226. Yu CS, Chen YC, Lu CH, Hwang JK: **Prediction of protein subcellular localization**. *Proteins* 2006, **64**(3):643-651.

VITA

Chao Zhang was born on April 16, 1978 in Beijing, China. He graduated from the High School Affiliated to Capital Normal University in 1996. He received his Bachelor's degree in Automation in 2000 from Beijing Institute of Technology and the Master of Science in Computer Science from University of Missouri in 2008. After getting his master degree, he continued to work with Dr. Dong Xu as a research assistant for his Ph.D. degree in Bioinformatics and Computational Biology. He also was admitted by the Department of Statistics, University of Missouri in the summer of 2009 as a graduate student. His research interests are cancer biology, analysis of host-pathogen interaction and computational analysis for metagenomic data.