Chapter 15
Bioinformatics Applications in Clinical Microbiology

Chao Zhang, Shunfu Xu, and Dong Xu

Abstract The human body is believed to house over 100 trillion microbes. These microbial communities have a tremendously influential impact on their human hosts. Although increasing evidence indicated a key role for the specific microbial species in carcinogenesis, such as *Helicobacter pylori* (*H. pylori*), Epstein-Barr virus, *Human papillomavirus*, and *Hepatitis C virus*, the underlying roles of human microbiome in cancers are still unclear. Using the bioinformatics algorithms and tools to integrate the microbiological data and clinical data could be very helpful to better understand the mechanisms of diseases. During the past decade, we have kept working on microbiome research and utilized bioinformatics methods to discover host-pathogen interactions, relationships between microbiome dynamics and diseases, and correlations between bacterial sequence variation and clinical outcomes. In this chapter, we use *H. pylori* as an example to demonstrate the procedure of related data integration, virulence classification, and prognosis model construction.

Keywords Microbiome • *Helicobacter pylori* • CagA • Gastric cancer • Bioinformatics • SVM

15.1 Introduction

As the most abundant domain of all living organisms on earth, bacteria are estimated to have more than five nonillion (10^{30}) individuals worldwide (Whitman et al. 1998), and these small single-cell organisms can be found everywhere. They

C. Zhang
Department of Medicine and Institute for Computational Biomedicine, Weill Cornell Medical College, New York, NY, USA

S. Xu
Department of Gastroenterology, Nanjing Medical University, Nanjing, Jiangsu, China

D. Xu (✉)
Department of Computer Science & C.S. Bond Life Science Center, University of Missouri, Columbia, MO, USA
e-mail: XuDong@missouri.edu

D.-Q. Wei et al. (eds.), *Translational Bioinformatics and Its Application*, Translational Medicine Research, DOI 10.1007/978-94-024-1045-7_15

are playing very important roles in our life, and we actually benefit from the microorganisms in many cases, e.g., food production, human health (Turnbaugh et al. 2009), environmental biotechnologies (Dinsdale et al. 2008), and chemical industry (Lorenz and Eck 2005). 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 two million people every year (Andries et al. 2005).

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 (Petrosino et al. 2009). The Human Microbiome Project (Turnbaugh et al. 2007) undoubtedly provides new insight into many aspects of complex microbial communities, such as metabolic capabilities of microorganisms, coevolution of bacteria and host, interactions among microbial cells, and so on (Medini et al. 2008). 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. In our previous study, we reviewed the major computational methods on metagenomic/genomic analyses and the future computational challenges on general microbial identification (Zhang et al. 2012a, 2015), and we will focus on bioinformatics applications in clinical microbiology in this chapter.

Immediately after birth, humans undergo a lifelong 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 many diseases (Hacker et al. 2003) including cancer (Ullman and Itzkowitz 2011). The host-bacteria interactions include many complicated mechanisms, and discovering associations between bacteria and diseases in a clinical setting is even more challenging. Due to the explosion of metagenomic/genomic data, DNA sequence-based identification and classification are becoming more and more important in exploring microbial diversity in clinical research. For example, *Bradyrhizobium enterica* was discovered in cord colitis syndrome with shotgun DNA sequencing of biopsy specimens (Bhatt et al. 2013). Recently we also found that the *Helicobacter pylori* (*H. pylori*) infection can change the gastric microbiome according to whole genome sequencing (WGS) on endoscopic biopsy. WGS gives a much more accurate identification on *H. pylori* infection than traditional methods, such as ELISA test and C-13 breath test. Besides *H. pylori* infection identification, we also spent much effort on discovering the molecular mechanisms that underlie different gastroduodenal diseases caused by *H. pylori* infection.

In this chapter, we use *H. pylori* as the example to describe how we utilized computational methods to discover the relationships between *H. pylori* virulence factor and diseases and built a potential model for clinical diagnosis or prognosis. At first, we collected and curated the data from public databases, and then through studying the distribution and polymorphism of EPIYA motif in CagA sequences, we attempted to better understand the function of EPIYA motif, especially the role

of EPIYA motif during the interaction process between *H. pylori* and hosts. We also constructed a computational model to access gastric cancer risk by using detected important residues in CagA intervening sequences.

15.2 Public Data Collection and Curation

H. pylori is a Gram-negative helix-shaped bacterium inhabiting the human stomach for possibly more than thousands of years. By far as one of the oldest known human pathogens, it infects more than half of the world's population (Suerbaum and Michetti 2002). *H. pylori* has shown a strong correlation with all gastroduodenal diseases, including duodenal ulcers (Covacci et al. 1993), gastric ulcers (Ernst and Gold 2000), and chronic gastritis, especially being an important risk factor for developing gastric cancer (Uemura et al. 2001). *H. pylori* is becoming more and more important not only for gastroenterologists and pathologists but also for phylogenists who use it as the evidence to study human's origin and migration (Linz et al. 2007).

As one of the most important model bacteria, the data of *H. pylori* have been increasing dramatically in recent years. As of January 2014, 399 genome-sequencing projects are almost complete or "in progress." 37,304 nucleotide sequences, 65,684 protein sequences, 61 primers, and 9953 publications were collected from several major databases, e.g., NCBI databases (http://www.ncbi.nlm.nih.gov), EBI databases (http://www.ebi.ac.uk), DDBJ (http://www.ddbj.nig.ac.jp), and PDB (http://www.pdb.org). We searched the above databases with the keywords "Helicobacter pylori" and "H. pylori" and then verified all results based on the taxonomy information. References were collected from PubMed (http://www.ncbi.nlm.nih.gov/pubmed).

As we know, geographical diversity and disease diversity are two most significant features and hottest topics on *H. pylori* research. Without these types of information, the sequences of *H. pylori* are not very useful for studying the underline mechanisms of *H. pylori* causing gastric diseases. Based on our research experience, collecting *H. pylori* data from various sources is laborious and difficult, and currently no database/website can provide the corresponding accurate information, and collecting comprehensive information of *H. pylori* specifically for a particular country or disease is even more time-consuming. We manually curated the information not only based on the records from the above major databases but also by reviewing related literature.

15.3 Data Deposit

Based on the information we collected and curated, a web-based database, HPbase (www.hpbase.org) has been built for providing a one-stop shop for *H. pylori* data from multiple sources together with multiple embedded search/analysis tools for querying the database. This website is not only for depositing collected public data but also for providing curated information, new data generated by users, and other features derived from original experimental data. By continuously accumulating and updating the data, we anticipate that HPbase will serve as an important resource for studying *H. pylori* and gastroduodenal diseases.

15.3.1 Implementation

The web interface is constructed using PHP, CSS, and the JavaScript jQuery framework for a flexible user interaction with the system. The HPbase database is implemented through a MySQL relational database as the backend data storage system. A Java-based tool was developed to periodically synchronize data with major sources, and it is also used to import related manually curated diseases and geographical information into the MySQL database.

15.3.2 Other Information

Besides the basic information we collected from other sources, we generated sequence profiles for all 65,684 protein sequences by running PSIBLAST (Altschul et al. 1997) (2007 release version) three rounds against nonredundant (NR) database (as of 2013) with the e-value cutoff of 0.001, and then we predicted secondary structures by using PSIPRED (McGuffin et al. 2000) with the sequence profiles generated above. We also predicted 3D structures for most of proteins, including all major ones, e.g., CagA and VacA, by using our in-house software MUFOLD (Zhang et al. 2010), which integrates whole and partial template information to cover both template-based and ab initio predictions in the same package. The predicted secondary and tertiary structural information could help users to better understand the interaction between human proteins and *H. pylori* proteins.

15.3.3 Browsing Data

Users can search *H. pylori* data by different entries, such as GI number, accession number, strain ID, keywords, disease type, geographical information, and so on

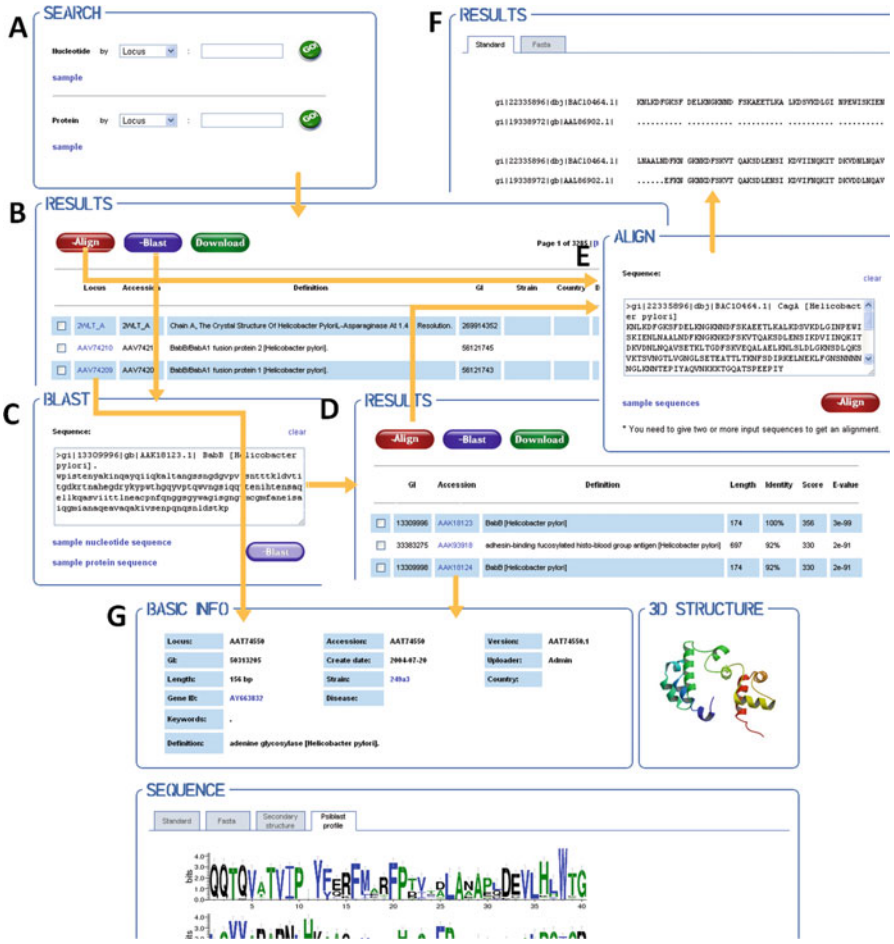

Fig. 15.1 Screenshots and basic workflow of HPbase. (**a**) Dialog for entering search criteria. (**b**) Table for displaying searching results. (**c**) Input dialog for BLAST. (**d**) Table for displaying BLAST results. (**e**) Input dialog for sequence alignment. (**f**) Dialog for displaying alignments between different sequences. (**g**) Page for displaying detail information of one particular protein/DNA sequence

(Fig. 15.1a). After submission of the search condition, the results will be displayed in a new page as a list of records with some brief information, including GI number, accession number, definition, strain ID, keywords, corresponding diseases, and geographical information. As shown in Fig. 15.1b, users can simply pick all sequence records or part of them in search results to download in the FASTA or Genbank/GenPept format for further analyses provided by HPbase, e.g., BLAST and multiple sequence alignment (MSA) as in Fig. 15.1c and e. Furthermore, users can also navigate detailed information of any particular nucleotide/protein record in

the result page by clicking on the corresponding "locus" link, and it will redirect to the nucleotide/protein detail page as shown in Fig. 15.1g. It not only provides the brief information as other major databases do but also includes manually curated information, e.g., disease-related and geographical information and computational information, e.g., PSIBLAST sequence profiles, secondary structures, and 3D structures for proteins. In this page, the sequence will be displayed with several formats. PSIBLAST sequence profile is represented as a sequence logo, and it is generated by using the WebLogo (Crooks et al. 2004) for the top 100 alignments of the last PSIBLAST round with no gap in the query sequence. Protein secondary structures are colored with the FASTA format. Jmol (http://www.jmol.org) is used as a viewer for displaying protein 3D structures. Users are also encouraged to add their own comments to each nucleotide/protein record and use the reference voting function to improve the correlation between each sequence record and its references, which could be helpful for others to better understand *H. pylori*.

15.3.4 Other Tools

Some further functions have also been embedded into the HPbase website to improve the power of search and data analysis. As shown in Fig. 15.1c, a BLAST utility was integrated as one useful feature, and two different BLAST programs have been included, e.g., BLASTn and BLASTp. By selecting gene entries from search results or uploading a protein/nucleotide sequence, users can retrieve identical or similar nucleotides/peptides in the database through BLAST according to user-defined parameters, which can be freely chosen including E-value, number of alignments, mutation matrix, and so on. As shown in Fig. 15.1d, a typical result page contains the collected information including GI number, accession, definition, length of sequence, E-value, identity, score, and alignment. Users can download records and further execute BLAST for database search or MUSCLE (Edgar 2004) for MSA by selecting records of their own interest from the BLAST results. Users can also upload their own sequences to perform multiple sequence alignment. In addition, the entire sequence data can be downloaded directly in the FASTA or Genbank/GenPept format. Users can also download data for one particular "strain," "disease," or "country." Some statistical analysis of the most important virulence factor – CagA from our previous work (Zhang et al. 2012b) – is also included in the website, including the relations between CagA sequence subtypes and diseases, the geographical diversity of CagA sequences, and the geographical diversity of different diseases.

15.4 Computational Model for CagA

15.4.1 Motivation

As one of the most important virulence markers of *H. pylori*, the cytotoxin-associated gene A (CagA) has been revealed to be related to the gastric disease occurrence. *H. pylori* strains carrying the CagA gene increase the risk factor of gastroduodenal diseases by threefold over CagA-negative strains (Blaser et al. 1995). CagA contains 1142–1320 amino acids, and at the C-terminal region, it has a variable region in which various short sequences (EPIYA motif) repeat 1–7 times. After colonizing on the surface of the gastric epithelium, *H. pylori* translocated into the gastric epithelial cell through type IV secretion system. Once injected into the host cell, CagA could localize to the plasma membrane. Src family tyrosine kinases can phosphorylate CagA on the specific tyrosine residues of a five-amino-acid (EPIYA) motif (Odenbreit et al. 2000). Then tyrosine-phosphorylated CagA binds specifically to SHP-2 tyrosine phosphatase (Higashi et al. 2002) 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 (Argent et al. 2004). At the same time through activating mitogen-activated protein kinase (MAPK), extracellular signal-regulated kinase (ERK) (Fu et al. 2009), and focal adhesion kinase (FAK), CagA also can cause cell dissociation and infiltrative tumor growth (Amieva et al. 2003).

CagA protein carries two unique features. One is the geographical diversity. 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, EPIYA-B, EPIYA-C, and EPIYA-D (Higashi et al. 2002). Among them, EPIYA-D motif only can be found from the East Asian subtype, and for the CagA from Western countries, EPIYA-D is replaced by EPIYA-C. EPIYA-D has stronger phosphorylation motif binding activity which leads to greater morphological changes than what the EPIYA-C motif can cause in infected cells (Higashi et al. 2002). And it explains the higher incidence of gastric cancer in East Asian countries (Jones et al. 2009).

Another feature of CagA is the variation in the number of EPIYA motif copies. Many studies attempted to reveal the relations between number of EPIYA motif repeats and clinical diseases (Lai et al. 2003). Although increasing of number of EPIYA motif copies will affect biological activities, due to the sample size limitation and geographic limitations of studies, none of the studies can draw a statistically significant conclusion about the relation.

Aside from the number of the EPIYA motif repeats, the sequence difference of strains in variable regions could also cause a significant difference of virulence, which might relate to the different pathogenic abilities of *H. pylori* (Naito et al. 2006). We speculate not only the number of EPIYA motif repeats, but also

polymorphism of CagA sequences will affect the virulence of *H. pylori* and then cause the different diseases. In this study, 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 noncancer group. This method not only sheds light on the relations between CagA sequences and gastric diseases but also may provide a potentially useful tool for gastric cancer diagnosis or prognosis.

15.4.2 Data Preprocessing

According to our collected data, 535 strains of *H. pylori* CagA protein with disease information will be used for this study. Among them, 287 strains belong to the East Asian subgroup, and the rest 248 are Western 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. Due to the significant difference between two subgroups, the East Asian subtype and the Western subtype were treated as two independent groups and analyzed within each group individually.

CagA sequences of each subtype were put into the corresponding disease groups, and then the multiple sequence alignments were applied for each group individually by using Clustal X version 2.0.3 (Larkin et al. 2007). Based on the aligned sequences, for each column of multiple alignments, we computed the background entropy B_i and the combinatorial entropy C_i based on the disease groups for each column i as follow:

$$C_i = \sum_k \ln \frac{N_k!}{\prod\limits_{\alpha=1\ldots20} N_{\alpha,i,k}!}$$

$$B_i = \sum_k \ln \frac{N_k!}{\prod\limits_{\alpha=1\ldots20} \widetilde{N}_{\alpha,i,k}!}$$

$$\widetilde{N}_{\alpha,i,k} = N_k N_{\alpha,i}/N$$

where N_k represents the number of sequences in group k, $N_{\alpha,i,k}$ indicates the number of residues of type α in the column i of group k, $N_{\alpha,i}$ is the number of residues of type α in the column i, and N represents the total number of aligned sequences. Then the entropy difference between the combinatorial entropy and the background entropy was calculated as feature values:

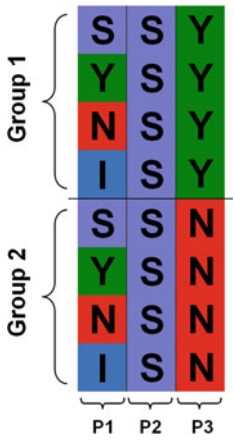

Fig. 15.2 An example to present different cases for the entropy calculation

$$\Delta E = C_i - B_i$$

Figure 15.2 illustrates the entropy concept using three extreme cases for a given column of residues from sequence alignment. Case P1 is the so-called randomly distributed or uniformly distributed, and the amino acids are "randomly distributed" over all groups and without significantly conserved pattern. Case P2 represents a "globally conserved" pattern, and all the amino acids are all almost same across different groups. In "locally conserved" case P3, some specific amino acids are only conserved in particular groups, and different groups have different conserved pattern.

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 or uniformly distributed" case, C_i gets the maximum value. "Conserved" and "randomly distributed" cases can be distinguished based on the value of combinatorial entropy, but it won't help pick "locally conserved" case from all "conserved" cases. Then we look at the value of background entropy, B_i gets the maximum value, 0 and medium value for the "randomly and uniformly distributed" case, "globally conserved" case, and "locally conserved" case, respectively. Finally, "locally conserved" case could be selected based on the differences between combinatorial entropy and the background entropy. The value of differences for the above three cases are $\Delta E_1 = 0$, $\Delta E_2 = 0$, and ΔE_3 gets the minimum value.

15.4.3 Modeling

The training/identification procedure has been implemented based on the workflow shown as follows (Fig. 15.3):

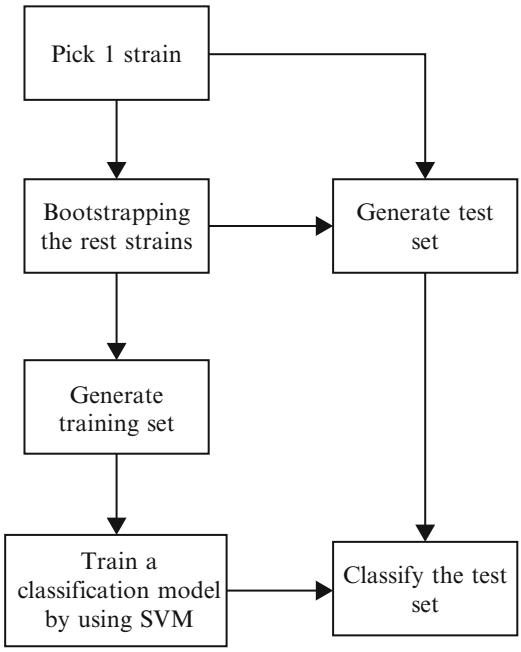

Fig. 15.3 Workflow of classification/prediction procedure for one specific CagA sequence

- 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.

A bootstrapping procedure was applied to avoid the classification bias, since the extremely unbalanced number of cases from different disease group. Usually gastric cancer cases will be much less frequent than other diseases, such as ulcer or gastritis. So we used all samples from noncancer group, and stains from the cancer group were continuously drawn on a random basis until getting the same number of samples as noncancer group. We also repeated this process five times to generate five independent training sets for each test strain, and the final decision is based on the average of five independent classification results. Due to the same reason, traditional n-fold cross validation won't fit our data. Then 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.

SVM^{Light} package V6.02 (http://svmlight.joachims.org/) (Joachims 1999) has been employed as the classifier, and radial basis function (RBF) has been chosen as kernel function. Two parameters were tuned to obtain the optimal F-vlaue by using grid search with above-generated training data. The feature values of each test stain were then fed into the optimized model to get the classification decision. Overall classification performances were evaluated by using the following measurements accuracy (Acc), sensitivity (*Sn*), specificity (*Sp*), Matthews correlation coefficient (*MCC*), and F-value:

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

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

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

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

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

where a true positive (*TP*) is a cancer-related sequence classified as such, while a false positive (*FP*) is a noncancer-related sequence classified as cancer related, a false negative (*FN*) is a cancer-related sequence classified as noncancer related, and a true negative (*TN*) is a noncancer-related sequence classified as noncancer related.

15.4.4 Comparison with Other Methods

Two popular identification methods, BLAST (Altschul et al. 1990) and HMMER (Eddy 1998), were selected as the representative methods for comparison. We applied the same evaluation procedure and measurements to above two tools as our method, such as LOO cross validation. The default parameters have been used for both BLAST and HMMER. Comparing the results for three methods, our method achieved 76% and 71% classification accuracy for Western and East Asian subtypes, respectively, which performed significantly better than the rest of the two methods (Table 15.1).

Table 15.1 Classification performances of different methods

Subtype	No. of cancer cases	No. of noncancer cases	Method	Sn	Sp	Accuracy	F-value	MCC
Western	37	211	Entropy	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	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

15.4.5 Discussion

It was found that CagA multimerizes in mammalian cells (Ren et al. 2006). This multimerization is independent to the tyrosine phosphorylation, but it is related to the "FPLxRxxxVxDLSKVG" motif, which is named CM motif following EPIYA-C motif. The CM motif plays an important role in CagA-positive *H. pylori*-mediated gastric pathogenesis, since the multimerization is a prerequisite for the CagA-SHP-2 signaling complex and subsequent deregulation of SHP-2. With multiple CM motifs, *H. pylori* strains are much likely associated with severe gastroduodenal diseases (Lu et al. 2008), 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, which might lead to the change of multimerization, thus changing the virulence of CagA. This is in consistent with a previous discovery (Sicinschi et al. 2010) that the sequence difference between the East Asian CM and the Western CM determines the binding affinity between CagA and SHP-2.

However, we also found that there is no simple relation between any single residue and cancer occurrence, and hence, it is not possible to just use one single residue to be the marker for identifying cancer. We speculate that one special combination of all or partial important residues could have a high correlation with one particular disease. The classification result strongly supports 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 noncancer groups is not very strong.

15.5 Summary

We described the procedures for collecting, curating, and depositing public data into a web-based database. With a user-friendly interface, those data could be easily downloaded, browsed, and searched by different entries. Some computational information (PSIBLAST sequence profile, protein secondary structures, and 3D

structures) have also been integrated into the database. This database is not only useful for our research but also could benefit the *H. pylori* and gastroduodenal disease research community.

Based on the curated CagA data, an entropy-based calculation was used to detect key residues of CagA intervening sequences as the gastric cancer biomarker. For each residue, both combinatorial entropy and background entropy were calculated, and the entropy difference was used as the criterion for feature residue selection. The feature values were then fed into SVM with the RBF kernel, and two parameters were tuned to obtain the optimal F-value by using a grid search. Two other popular sequence classification methods, the BLAST and HMMER, were also applied to the same data for comparison. Our study indicates that small variations of amino acids in those important residues might lead to the virulence variance of CagA strains resulting in different gastroduodenal diseases. This study provides not only a useful tool to predict the correlation between the novel CagA strain and diseases but also a general new framework for detecting biological sequence biomarkers in population studies.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10. doi:10.1016/S0022-2836(05)80360-2.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389–402.

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):1430–4. doi:10.1126/science.1081919.

Andries K, Verhasselt P, Guillemont J, Gohlmann HW, Neefs JM, Winkler H, Van Gestel J, Timmerman P, Zhu M, Lee E, Williams P, de Chaffoy D, Huitric E, Hoffner S, Cambau E, Truffot-Pernot C, Lounis N, Jarlier V. A diarylquinoline drug active on the ATP synthase of Mycobacterium tuberculosis. Science. 2005;307(5707):223–7. doi:10.1126/science.1106753.

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):514–23.

Bhatt AS, Freeman SS, Herrera AF, Pedamallu CS, Gevers D, Duke F, Jung J, Michaud M, Walker BJ, Young S, Earl AM, Kostic AD, Ojesina AI, Hasserjian R, Ballen KK, Chen YB, Hobbs G, Antin JH, Soiffer RJ, Baden LR, Garrett WS, Hornick JL, Marty FM, Meyerson M. Sequence-based discovery of Bradyrhizobium enterica in cord colitis syndrome. N Engl J Med. 2013;369 (6):517–28. doi:10.1056/NEJMoa1211115.

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):2111–5.

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):5791–5.

Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004;14(6):1188–90. doi:10.1101/gr.849004.

Dinsdale EA, Pantos O, Smriga S, Edwards RA, Angly F, Wegley L, Hatay M, Hall D, Brown E, Haynes M, Krause L, Sala E, Sandin SA, Thurber RV, Willis BL, Azam F, Knowlton N, Rohwer F. Microbial ecology of four coral atolls in the Northern Line Islands. PLoS One. 2008;3(2):e1584. doi:10.1371/journal.pone.0001584.

Eddy SR. Profile hidden Markov models. Bioinformatics. 1998;14(9):755–763. doi:btb114 [pii]

Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–7. doi:10.1093/nar/gkh340.

Ernst PB, Gold BD. The disease spectrum of Helicobacter pylori: the immunopathogenesis of gastroduodenal ulcer and gastric cancer. Annu Rev Microbiol. 2000;54:615–40. doi:10.1146/annurev.micro.54.1.615.

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. 2009;41(8):648–56.

Hacker J, Hentschel U, Dobrindt U. Prokaryotic chromosomes and disease. Science. 2003;301(5634):790–3. doi:10.1126/science.1086802.

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):14428–33. doi:10.1073/pnas.222375399.

Joachims T. Making large-scale support vector machine learning practical. In: Schölkopf B, editors. Advances in kernel methods: support vector machines. Cambridge, MA: MIT Press; 1999. doi:citeulike-article-id:227265.

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;47(4):959–68. doi:10.1128/JCM.02330-08.

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):235–43.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. Clustal W and Clustal X version 2.0. Bioinformatics. 2007;23(21):2947–8. doi:10.1093/bioinformatics/btm404.

Linz B, Balloux F, Moodley Y, Manica A, Liu H, Roumagnac P, Falush D, Stamer C, Prugnolle F, van der Merwe SW, Yamaoka Y, Graham DY, Perez-Trallero E, Wadstrom T, Suerbaum S, Achtman M. An African origin for the intimate association between humans and Helicobacter pylori. Nature. 2007;445(7130):915–8. doi:10.1038/nature05562.

Lorenz P, Eck J. Metagenomics and industrial applications. Nat Rev Microbiol. 2005;3(6):510–6. doi:10.1038/nrmicro1161.

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):2004–11. doi:10.1111/j.1349-7006.2008.00950.x.

McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. Bioinformatics. 2000;16(4):404–5.

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):419–30. doi:10.1038/nrmicro1901.

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):1181–90. doi:10.1053/j.gastro.2005.12.038.

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):1497–500.

Petrosino JF, Highlander S, Luna RA, Gibbs RA, Versalovic J. Metagenomic pyrosequencing and microbial identification. Clin Chem. 2009;55(5):856–66. doi:10.1373/clinchem.2008.107565.

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):32344–52. doi:10.1074/jbc.M606172200.

Sicinschi LA, Correa P, Peek RM, Camargo MC, Piazuelo MB, Romero-Gallo J, Hobbs SS, Krishna U, Delgado A, Mera R, Bravo LE, Schneider BG. CagA C-terminal variations in Helicobacter pylori strains from Colombian patients with gastric precancerous lesions. Clin Microbiol Infect. 2010;16(4):369–78. doi:10.1111/j.1469-0691.2009.02811.x.

Suerbaum S, Michetti P. Helicobacter pylori infection. N Engl J Med. 2002;347(15):1175–86. doi:10.1056/NEJMra020542.

Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The human microbiome project. Nature. 2007;449(7164):804–10. doi:10.1038/nature06244.

Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, Egholm M, Henrissat B, Heath AC, Knight R, Gordon JI. A core gut microbiome in obese and lean twins. Nature. 2009;457(7228):480–4. doi:10.1038/nature07540.

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):784–9. doi:10.1056/NEJMoa001999.

Ullman TA, Itzkowitz SH. Intestinal inflammation and cancer. Gastroenterology. 2011;140 (6):1807–16. doi:10.1053/j.gastro.2011.01.057.

Whitman WB, Coleman DC, Wiebe WJ. Prokaryotes: the unseen majority. Proc Natl Acad Sci U S A. 1998;95(12):6578–83.

Zhang J, Wang Q, Barz B, He Z, Kosztin I, Shang Y, Xu D. MUFOLD: a new solution for protein 3D structure prediction. Proteins. 2010;78(5):1137–52. doi:10.1002/prot.22634.

Zhang C, Zheng G, Xu S-F, Xu D. Computational challenges in characterization of bacteria and bacteria-host interactions based on genomic data. J Comput Sci Technol. 2012a;27(2):225–39. doi:10.1007/s11390-012-1219-y.

Zhang C, Xu S, Xu D. Risk assessment of gastric cancer caused by Helicobacter pylori using CagA sequence markers. PLoS One. 2012b;7(5):e36844. doi:10.1371/journal.pone.0036844.

Zhang C, Cleveland K, Schnoll-Sussman F, McClure B, Bigg M, Thakkar P, Schultz N, Shah MA, Betel D. Identification of low abundance microbiome in clinical samples using whole genome sequencing. Genome Biol. 2015;16:265. doi:10.1186/s13059-015-0821-z.