Saarland University
Center for Bioinformatics
Max Planck Institute for Informatics
Bachelor's Program in Bioinformatics

Bachelor's Thesis

Web Implementation of the HBV Dual Infection Model

submitted by

Matthias Döring

on September 29, 2011

Supervisor

Prof. Dr. Thomas Lengauer, Ph.D.

Advisor

Dipl.-Mathematiker Bastian Beggel

Reviewers

Prof. Dr. Thomas Lengauer, Ph.D.
Dr. Ingolf Sommer

Statement

I hereby confirm that this thesis is my own work and that I have documented all sources used.

Saarbrücken, September 29, 2011

Matthias Döring

Abstract

Hepatitis B is endemic in many parts of the world and each year approximately 600 000 people die as a consequence of infection with hepatitis B virus (HBV). HBV genomes are divided into eight genotypes A to H according to phylogenetic analysis. Genotyping of HBV is an important tool in tracking the evolution of the viral genome as well as in therapy. It is possible for one organism to be infected by HBV viruses of different strains. Such an infection is referred to as a HBV dual infection. The identification of dual infections could improve results in the areas mentioned above. Although several computational methods for genotyping beside phylogenies have been developed in the past, none of them was explicitly concerned with HBV dual infections.

In this thesis, we present the implementation of a web tool that is able to identify the genotype(s) of one or multiple HBV nucleotide sequences, either full or subgenomic. Its main feature is the ability to recognize HBV dual infections.

In this approach, HBV genotypes are predicted according to a probabilistic model that is based on genotype-specific nucleotide distributions. A maximum likelihood classifier is used to determine the most likely genotype.

Our method was able to identify the genotypes of all 2258 annotated single infection sequences in a data set retrieved from NCBI correctly. Recombinant genotypes were correctly identified in 46 (86.8%) of 53 annotated recombinant sequences. To determine the performance for dual infections, an artificial data set of 6482 sequences was generated by pairwise mixing of sequences. For this data set, dual infections of HBV with different genotypes could be correctly identified in 5206 (99.9%) of 5211 such sequences.

This tool enables the detection of HBV dual infections, single infections, and recombinants in a reliable and fast manner. It may be of use to basic research and may help to further therapy and tracking of viral evolution. It will be available on the geno2pheno[hbv] web server, which is located at http://hbv.bioinf.mpi-inf.mpg.de/index.php.

Acknowledgments

First of all, I want to thank my advisor Bastian Beggel for providing me the topic of this thesis and always giving helpful advice, while guaranteeing a positive working atmosphere. I am greatly indebted to Prof. Dr. Lengauer, Ph.D. for giving me the opportunity to write my thesis in his group at the Max Planck Institute for Informatics in such a nice atmosphere.

I also want to thank my fellow students from the MPI and CBI for the relaxing lunch and coffee breaks, and especially Alexander Junge and Thorsten Will for proofreading my thesis.

Last but not least I would like to use this opportunity to thank my family for all their support throughout the years.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation for Studying HBV

HBV is endemic in many parts of the world. According to an estimation by the World Health Organization (WHO) about 2 billion people worldwide have been infected with the hepatitis B virus (HBV) and about 350 million live with chronic infection. Each year approximately 600 000 people die as a consequence of acute or chronic HBV infection [73].

1.2 Biological Background

1.2.1 Structure of HBV

HBV is an orthohepadnavirus of the hepadnaviridae family, whose members are characterized by DNA genomes and infection of liver cells. HBV has a partially double stranded DNA genome (about 3.2 kb) and infects the cells of the human liver, causing hepatitis, which is an inflammation of the liver.

The virion has an outer lipid envelope and an icosahedral nucleocapsid core that is composed of protein. Within the nucleocapsid core reside the viral DNA and a DNA polymerase with reverse transcriptase activity. The outer lipid layer also contains proteins, which play a role in viral cell binding and entry. HBV is one of the smallest enveloped animal viruses with a diameter of only 42 nm.

The genome of HBV contains four open reading frames: hepatitis B surface antigen (HBsAg), hepatitis B core antigen (HBcAg), hepatitis B virus DNA polymerase (HBpol), and hepatitis B protein X (HBx). In the following we will discuss the most important proteins encoded by the four genes.

HBsAg is present in the surface of the protein, i.e. in its lipoprotein coat. It is present in the sera of patients infected with HBV. Patients with developed antibodies against HBsAg are usually considered non-infectious.

HBcAg is the major component of the icosahedral nucleocapsid and is an indicator of active viral replication.

HBeAg (Hepatitis B early antigen) is a secretory variant of HBcAg, although a portion of HBeAg is known to remain cell-associated. It is produced by proteolytic processing of the pre-core protein. Patients can be either HBeAg-positive or HBeAg-negative, that is, show presence of HBeAg in blood serum or not. HBeAg-negativity is the consequence of a mutations in the pre-core and core promoters [69]. Chronic patients with early HBeAg seroconversion have a more favourable outcome than those with late HBeAg seroconversion, which may be associated with a faster progression to cirrhosis [34].

Hepatitis B virus DNA polymerase has reverse transcriptase activity and is integral to the replication cycle of HBV.

The function of HBx is still not fully elucidated.

1.2.2 Life Cycle

HBV gains access to the cell by binding to a receptor on the surface of the cell and entering via endocytosis. Then the viral membrane fuses with the membrane of the host cell and viral DNA and core proteins are released. HBV requires the DNA polymerase of the host cell for transcription. Therefore the virion is transported to the cell nucleus via chaperones of the host cell. In the nucleus, the partially double stranded viral DNA is transformed into covalently closed circular DNA, which is a fully double-stranded DNA. This transformation is necessary for the synthesis of the four viral mRNAs.

The largest mRNA is used to make new copies of the genome and generate new capsid core proteins as well as viral DNA polymerases. After processing of the four viral mRNAs, progeny virions are formed. They are released from the cell or returned to the nucleus where their nucleic acids are re-cycled to produce even more copies [37].

The replication cycle of HBV in chronic carriers seems to deter from the one outlined above since viral DNA has been found to be integrated into host DNA and being involved in carcinogenesis [46].

1.2.3 Transmission of HBV

HBV is transmitted via three routes: by parenteral exposure to blood or bodily fluids, by horizontal transmission, or by perinatal transmission.

In parenteral transmission, the skin or mucous membrane is pierced, allowing entry of HBV pathogens. An example for this route is the transmission of HBV by sharing of HBV-contaminated injection needles.

Horizontal transmission refers to transmission of HBV from an infected person to another person. An example for horizontal transmission is sexual intercourse.

When a mother infected with HBV transmits the infection to the neonate shortly before or after the time of birth, this is called perinatal transmission [4].

Perinatal transmission is the major mode of HBV transmission in most areas with high HBV incidence [18], for example Southeast Asia or Africa. This is because in such regions people are either unaware about infection or HBV vaccines are unavailable or too expensive. Perinatal infection presents a particular threat because it almost always leads to chronic infections. This is contingent on the fact that the infant immune system usually does not spontaneously clear the infection [4].

In areas with low HBV endemicity, for example in most parts of Europe, North America, temperate parts of Central and South America, and Australia, HBV is often contracted via other routes, such as sharing of contaminated needles during drug use and sexual contact. Perinatal transmission is less common in these countries due to infant vaccination. If HBV antibodies (hepatitis B immune globulin) are administered within twelve hours of birth, infection fails to appear in 90% of the cases [33].

1.2.4 Acute and Chronic Infection

There are two types of infection regarding hepatitis B: acute and chronic. An infection with hepatitis B is considered to be acute during the first six months after initial contraction. Acute infection is associated with symptoms such as general ill-health, loss of appetite, nausea, vomiting, body aches, mild fever, dark urine, and jaundice. Acute infections with HBV in adults are self-limiting in 95% of all cases, that is, they are cleared by the body itself. However, this is different for younger children and newborns, for which the clearance rates are 30% and 5%, respectively [4].

Because infection with HBV in adults is usually self-limiting, antiviral therapy of acute infection is necessary only for those patients that are subject to very aggressive acute infections (fulminant hepatitis) and those who are immunocompromised.

An infection with hepatitis B is considered chronic if HBsAg is still present after six months after being exposed to the virus [36]. A patient with chronic hepatitis B infection is also called a hepatitis B carrier. Chronic infections can last a lifetime. In an estimated amount of 70% to 90% of chronic infections, infection with hepatitis B is asymptomatic for several decades [11], that is, the patient is HBsAg-positive but without impaired liver function.

Asymptomatic carriers are likely to develop liver cirrhosis or primary liver cancer [75]. The asymptomatic course of the disease is especially problematic, since infected persons may not be aware about their infection, thereby facilitating the further propagation of the virus. General symptoms of chronic infection with HBV include malaise, tiredness, weakness, and, in cases of advanced liver damage, jaundice.

Chronic HBV infection can lead to cirrhosis, a state in which the liver tissue is scarred. In addition, chronic HBV infection increases the risk for acquiring hepatocellular carcinoma (HCC), the most common type of liver cancer. Extensive liver damage leads to weight loss, easy bruising

and bleeding tendencies, peripheral edema (swelling of the legs), and ascites (accumulation of fluid in the abdominal cavity). Cirrhosis can lead to various complications, such as esophageal varices (extremely dilated veins in the esophagus that can cause life-threatening bleeding) hepatic encephalopathy (confusion and coma) and hepatorenal syndrome (kidney dysfunction).

Chronic patients may be treated with antiviral drugs in order to reduce the risk for cirrhosis and HCC. Patients with chronically raised levels of alanine aminotransferase, a marker of liver damage, and evidence of high HBV DNA levels in blood serum are to be considered for therapy [32].

1.2.5 Treatment

Hepatitis B is usually treated only in the following cases: in patients with fulminant hepatitis, in immunocompromised patients with acute infection, and in the presence of chronic infection with abnormal levels of aminotransferase. Such patients are treated with drugs from two classes: nucleot(s)ide analog reverse transcriptase inhibitors and immune system modulators.

Nucleotide analog reverse transcriptase inhibitors are chemically similar to the naturally occurring nucleotides. However, they do not contain a 3'-hydroxyl group on the deoxyribose moiety in contrary to the naturally occurring nucleotides. If such a nucleotide analog is integrated into the DNA chain during the elongation process, the next nucleotide will not be able to form the 5'-3' phosphodiester bond that is needed to extend the chain. This halts viral DNA synthesis.

The only difference between nucleotide analogs and nucleoside analogs is that the latter need first be activated by addition of three phosphate groups before they can be incorporated into DNA. Examples for drugs of this class that are used to treat HBV infections include lamivudine, adefovir dipivoxil, and tenofovir. They usually have to be taken once-daily orally.

Immune system modulators on the other hand directly act upon the immune system. An example for a drug of this class that is used in HBV therapy is pegylated interferon alpha 2-a. Interferon is a protein that is made and released by host cells in the presence of pathogens. It is crucial for cell communication triggering the defensive mechanisms of the immune system. The name *interferon* comes from its ability to interfere with viral replication.

Pegylated interferon alpha 2-a has an enhanced half-life compared to its native form. It requires injections only once a week in contrast to daily or thrice a week. It was shown that interferon is effective in terminating viral replication and eradicating the carrier state in patients that are HBeAg positive [74], thereby improving the outcome [56]. Another study revealed that interferon treatment also reduces the chance of acquiring hepatocellular carcinoma in patients with liver cirrhosis [23].

1.2.6 HBV Serotypes

There are four major serotypes that are assigned to HBV according to antigenic epitopes on its surface protein HBsAg, namely adr, adw, ayr, and ayw. In comparison to genotypes, the significance of serotypes is only minor. In the past, HBV serotypes were used for tasks such

as tracking the route of HBV transmission or determining the geographical migration of HBV carriers [1, 10].

However, the use of serotypes has been superseded by genotypes with the emergence of DNA sequencing in the mid-seventies due to their greater clinical impact and usefulness in epidemiological studies.

1.2.7 HBV Genotypes

Genotypes, Subgenotypes, and Clades

The genetic constitutions of organisms are described by their genotypes, which are used to classify the hepatitis B virus. HBV genotypes are subdivided into eight genotypes A to H according to phylogenetic analysis based on a threshold of 8% nucleotide divergence between genotypes across the entire genome [51].

Moreover, there are also two proposed genotypes I and J [52, 68]. Both genotypes are not universally accepted. This is because genotype I exhibits a genetic divergence from genotype C of less than 8% [31] and genotype J has not yet been confirmed by other groups of researchers.

A further subdivision of HBV genotypes is provided by subgenotypes. They are defined on the basis of an inter-group nucleotide difference of at least 4% across the complete genome [57]. Subgenotypes are typically named by appending Arabic numbers to the genotype of the virus, starting at 1.

There is also the notion of clades, which compartmentalize subgenotypes. Members of a clade exhibit an inter-group divergence of only less than 4% [57].

Importance of Genotypes

Genotypes show a distinct geographical distribution [50]. For example, genotype E is confined to Western Africa, while genotype B is only present in Asia, with subgenotype B1 being specific to Japan and subgenotype B2 being specific to Asia [30]. This property of genotypes is the reason why genotyping is a crucial tool for the analysis of molecular evolution and patterns of HBV spread.

Genotypes are also of clinical importance due to their influence on therapy and disease progression [30]. Knowledge about the HBV genotype of a patient could be used in order to predict the risk of adverse outcomes such as fulminant hepatitis, cirrhosis, or HCC or to guide treatment decisions, such as the choice of drugs [16].

Impact of Genotypes on Disease Progression

Several studies have been undertaken to investigate the clinical significance of HBV genotypes. Common criteria for measuring the impact of HBV genotype on the course of disease include HBeAg seroconversion, activity of liver disease, and rate of progression to HCC or cirrhosis.

A common deficit of most undertaken studies is the fact that they are only regional studies, that is, are only concerned with patients from a restricted geographical area. In consequence, due to the geographical distribution of HBV, these studies are only able to analyze limited spectra of genotypes. However, in order be able to reasonably compare the implications of individual genotypes with each other, global cross-sectional studies would be more appropriate.

When considering the results of the following studies, one should keep in mind that other factors such as age of disease onset, ethnic background of patients, and exposure to certain substances such as alcohol or other environmental toxins may also have had an impact.

The majority of information on the association of HBV genotypes on the outcome of chronic infections is based on studies that were conducted in Asia [16]. Since genotypes B and C are predominant in Asia, all of the studies conducted deal mainly with these two genotypes. It could be shown that genotype B was associated with spontaneous HBeAg seroconversion at a younger age, less active liver disease, and a slower progression to cirrhosis [65, 35, 25, 67, 53, 7]. Patients that are infected with HBV of genotype B are also less likely to have hepatitis flares and are more likely to remain in remission after HBeAg seroconversion [7].

Data on HCC is contradictory regarding genotypes B and C: most studies showed that HCC is less frequent and occurs at an older age in patients infected with genotype B [13, 67, 53]. Another study, however, showed that HCC was more frequent in patients younger than 50 years who were infected with HBV of genotype B [27].

There is a paucity of data on genotypes other than B and C. A study from Spain revealed that genotypes A and D possess similar HBeAg seroconversion rates, but that remission was more common in patients with genotype A. The need for liver transplantation and the deaths related to liver disease were comparable between genotypes A and D. Patients with genotype F were shown to be more likely to die from liver disease than those with genotypes A and D [60].

A cross-sectional study conducted in the USA, featuring a broader spectrum of genotypes ranging from A to G and 694 patients, showed that genotypes B and D were associated with lower levels of HBeAg than genotype A. In addition, it was observed that genotype B was associated with lower rates of hepatic functional deterioration compared to genotype A, C, and D [8].

Role of Genotypes in HBV Treatment

HBV genotypes also play a role in therapy, in particular regarding interferon treatment. In a study that dealt with pegylated interferon it was shown that HBeAg seroconversion occurred more often in patients with genotype A (47%) and B (44%), but more seldom in patients with genotype C (28%) and genotype D (25%) [39]. Another study showed that genotype B has a greater response rate to interferon-alpha treatment than genotype A, with response rates of 41% and 15%, respectively [27]. Similar results were observed in another study, in which antiviral response was initiated by interferon-alpha in 39% of patients with genotype B, but only in 17% of patients with genotype C [72]. It has not yet been determined if HBV genotypes play a role in interferon treatment in patients with HBeAg-negative chronic hepatitis.

A conclusion which can be drawn from these studies is that it is reasonable to stratify populations in studies that evaluate the efficacy of interferon therapy according to genotypes.

For lamivudine and adefovir dipivoxil no correlation between therapy and genotype could be found. However, in one study, lamivudine resistant mutants emerged faster in patients with genotype A than in patients with genotype D [76].

Regarding the impact of HBV genotype on the outcome of acute infections very little is known. A study conducted in Switzerland observed that 80% of patients with acute hepatitis B had genotype D, while 80% of patients with chronic hepatitis B had genotype A [40]. These results were confirmed by a study from Japan, which showed that 12% of patients with chronic infection had genotype B, while 39% of patients with acute infection had genotype B [24].

There is not much information on the association between HBV genotype and fulminant hepatitis. It is, however, suggested that cofactors causing liver injury may be more important than viral factors in the fulminant course of the disease [17].

HBV Dual Infections

A dual infection with HBV is present when a person is infected with two different strains of HBV simultaneously. One can distinguish two types of dual infections: superinfection and coinfection. In a superinfection, the second infection occurs after the organism's immune reaction to the first infection, while in a coinfection the infection with the two strains occurs at roughly the same time, that is, before any immune reaction.

The distinction between these two types of infections is made because of the clinical impact of superinfections, which may affect the course of disease greatly. For example, superinfection with a more virulent HBV strain will exacerbate the health status of patients with previous HBV infections. In addition, superinfection with another strain can impact the efficiency of antiviral treatment, in particular when a drug resistant strain is introduced into the host.

Dual infections are the prerequisite for viral recombination events in which two virions exchange segments of genetic material. The resulting viral genomes consist of a mixture of both genomes and may therefore contain a combination of different genotypes. Recombination provides a mechanism of viral genetic variation.

Recombinants seem to arise predominantly in areas of high viral incidence. For example, Tibet has a HBV carrier rate of 26.2% and 96% of sequenced isolates were determined to be C/D recombinants. The clinical impact of HBV recombinants is not known yet.

Prevalence of Dual Infections The data on the prevalence of HBV dual infections is inconclusive. In a study conducted in Western Japan, four (17.4%) of the 23 patients with genotype D were co-infected with genotype C [42]. In another study concerned with patients with chronic hepatitis B infection of genotype A, dual infections were found in 20 (67%) of 30 HBeAg-positive patients treated with interferon [20]. A further study among chronically infected HBV patients revealed that only 0.8% (2 of 244 Taiwanese patients) harbored a superinfection [26].

A study in which the patient genotype distribution approximated that of the US revealed that only 15 (1.6%) of 946 samples exhibited infections with multiple genotypes [38]. Further studies showed dual infection prevalences of 4.4% (8 out of 183 Chinese patients) [12], 10.9% (28 out of

256 patients from all around the world) [28], 10.1% (19 out of 118 Canadian patients) [54], and 14% (Kirschberg, Schaefer, and Erhardt, unpublished data).

A reason for the divergence in the number of detected dual infections in the above mentioned studies may be due to the experimental methods employed. Some methods are better suited for the detection of infections with two viral strains than other methods. For example, genotype-specific probes assays (GSPA) are more sensitive to dual infections than enzyme-linked immunosorbent assays (ELISA) [28]. Another factor is the the limited statistical validity of the analyzed populations, that is, the limited population sizes, the restriction of populations to certain regions, and the limited number of genotypes.

Although these data do not make a specific statement about the prevalence of HBV dual infections, one can still see that they are a quite frequent phenomenon among some populations.

Clinical Significance of Dual Infections There exists only little information on the clinical significance of HBV dual infections. In one study no evidence of clinical differences between mono- and co-infected patients with genotype D were detected [42]. Another study observed that there was no significant impact of dual infections on interferon treatment.

An event referred to as genotypic shift, in which the predominant HBV genotype changes, was common after interferon treatment of patients infected with two strains of HBV. Such a genotypic shift is the result of the minor viral strain becoming more fit as a consequence of treatment. After discontinuation of therapy, however, the previously disappeared genotypes returned again [20]. Another study [26] concluded that superinfection in HBV carriers may be associated with acute exacerbations in chronically infected patients.

1.3 Related Work

There exist several approaches to genotyping HBV. We will first discuss a selection of general methods for genotyping and deal with web-based computational tools later on.

1.3.1 General Methods for Genotyping

Sequencing Followed by Phylogenetic Analysis

Sequencing followed by phylogenetic analysis is the gold standard for HBV genotyping [3]. It is based on PCR amplification and sequencing of the entire HBV genome followed by phylogenetic analysis [49, 51]. The method of phylogenetic analysis allows the delineation of genotypes by determining the relative and evolutionary relatedness of sequences to each other [41].

This method is very reliable and it is the only method that is suitable for the analysis of new genotypes. Phylogenetic analysis, however, is time-consuming and calls for experience in complex computer analysis programs. Another limitation of this approach is its inability to detect mixed viral populations. This is because PCR cloning techniques are able to detect only the predominant

species. In order to detect mixed populations, one would need to screen a large number of clones, e.g. 100. This, however, would make analysis time-consuming and expensive.

Genotyping HBV Using Type-Specific Primers

This approach to genotyping HBV is based on the use of genotype-specific primers in PCR [45].

To amplify HBV DNA, two rounds of PCR were performed in this study. In the first round, genotype-unspecific primers were used to amplify HBV DNA. In the second round, however, type-specific primers were employed. These type-specific primers were subdivided into two sets of primers, primer mixes A and B. Mix A was used to detect genotypes A, B, and C, while mix B was used to detect genotypes D, E, and F. For mix A, a single sense primer was used, but three different antisense primers, one for each of the tree genotypes. For mix B, a single antisense primer was used in combination with three genotype-specific sense primers.

To determine the genotypes of HBV sequences, the amplicons resulting from the PCR runs with each of the primer mixes were analyzed using gel electrophoresis. The amplicon genotypes could be distinguished in the gel according to differences in their sequence sizes, which were effected by the used primer combinations.

This method is convenient and allows the rapid and sensitive genotyping of HBV.

Clonal Analysis of PCR products

Clonal analysis is a method that can be used to identify the genotypes involved in a HBV dual infection [54].

In this approach, specific amplicons of a specimen are purified and cloned into a plasmid vector. The resulting ligated products are then transformed into *Escherichia coli*. After that, individual clone colonies are picked and the inserted DNA is sequenced.

To identify the genotypes, the sequenced data are aligned to sequences, which are representative for their genotype. If only a single viral strain is present, all picked clones will align reasonably well to reference sequences of only a single genotype. For dual infections on the other hand, different clones will align reasonably to various genotypes. The explicit genotypes in these cases can be determined by the reference strains with the highest scores in the alignment.

In this approach the number of clonal colonies used is crucial: if the number is too low, underrepresented genotypes can be missed. On the other hand, picking a large number of clones is both time-consuming and expensive.

Genotyping HBV by GSPA

Genotype-specific probes assay (GSPA) is a wet laboratory method developed by Kato et al. [28] that is able to detect HBV dual infections.

This approach is based on PCR amplification of query HBV nucleotide sequences using a primer labeled with biotin. After PCR, the amplicons are delivered into wells on a plate that each contain complementary, immobilized sequences, which are specific to one genotype each. After hybridization, viral DNA labeled with biotin can be detected using colorimetry. The genotype of a specimen is assigned according to the well in the microplate that supersedes the optical density threshold. If two wells on the plate supersede the threshold value for the optical density, this indicates that an infection of two different HBV genotypes is present.

In the evaluation of this assay, 28 of 256 sera showed a dual infection with HBV strains of distinct genotypes, although ELISA detected the presence of only a single genotype in these samples. To validate the findings, clonal analysis was performed for five representative sera, which confirmed the results of GSPA for these sera.

Genotyping with ELISA

Enzyme-linked immunosorbent assay (ELISA) is a method in which the presence of an antibody or antigen is detected.

There are three steps in ELISA. First, a certain amount of antigen is affixed to a surface. Second, a specific antibody that is linked to an enzyme is applied over this surface in order to facilitate reaction with the antigen. Third, a substance, which the enzyme attached to the antibody can convert into a detectable signal, is added to the assay. In this way antigen-antibody binding is detected.

Here [70], monoclonal antibodies were raised against genotype-specific epitopes in HBsAg and labeled with horseradish peroxidase. HBsAg antigens in sera were captured by immobilizing antibodies against the common determinant and evaluated for reactivity with the enzyme-labeled antibodies.

ELISA is well-suited for large-scale surveys because it allows the serological detection of genotypes without sequencing nucleotides.

INNO-LiPA HBV Genotyping Assay

INNO-LiPA (Innogenetics line probe assay) is an assay that is used for direct molecular genotyping of HBV genomes [54]. The method is based on DNA hybridization.

First, DNA of the HBsAg region is amplified via PCR, generating biotinylated amplicons. These amplicons are then denatured and hybridized to specific oligonucleotide probes that were immobilized as parallel lines on membrane-based strips. After hybridization, partially bound DNA is removed in a washing step. Thereon, the strips are incubated with streptavidin conjugate in order to be able to detect hybridization according to biotin color development.

The performance of this assay was evaluated via sequencing followed by phylogenetic analysis. When LiPA and sequencing detected different genotypes, the findings were confirmed using cloning, picking of individual colonies, and sequencing of the plasmid-inserted HBV DNA. Subsequently,

genotypes were determined by aligning the sequenced DNA to GenBank sequences with known genotypes.

LiPA was able to detect 19 mixed genotypes among 188 tested specimens. All except six of the identified dual infection genotypes were in agreement with the results from clonal analysis. This might have been contingent on sequencing only up to 12 clones per specimen. For single genotypes, a concordance with phylogenetic analysis was observed in 81% of cases.

The INNO-LiPA genotyping assay is an easy-to-use and fast method for detecting HBV genotypes and sensitively identifying infections of mixed genotypes.

1.3.2 Web Servers for Genotyping

NCBI HBV Prediction Server

The NCBI HBV prediction server was developed by Rozanov et al. [59] in 2004 and is based on aligning the input sequence against several reference sequences using BLAST. The reference sequence database contains three sequences for each genotype. The similarity of the input sequence to each of the reference sequences is given by the corresponding BLAST score, i.e. by the score of the alignment. The genotype of the input sequence is determined according to the genotype of the reference sequence that attained the highest scores in the alignment with the input sequence.

In order to determine recombinant sequences, a sliding-window approach is employed. In this approach, the input sequence is partitioned into a number of subsequences, called windows. BLAST queries are effected on each of these windows rather than on the full input sequence. The usage of multiple layers of such windows enables the pinpointing of genotypic regions in recombinant sequences. The results of genotyping are illustrated by a graph that shows the BLAST scores of each genotype and the identified genotypes of each window.

The NCBI genotyping tool allows the almost instant identification of HBV genotypes and is especially useful when dealing with recombinants. A disadvantage of this tool is that it is not explicitly concerned with HBV dual infections and is therefore not able to identify them reliably.

HBV STAR

HBV STAR is a web server for the identification of HBV genotypes that was developed by Myers et al. [43] in 2006. It is based on genotype- and position-specific scoring matrices that are given on amino acid level.

The input sequence is scored according to the same scoring function that was already used for HIV STAR [44]. It involves the computation of discriminant odds based on scoring matrices. For each amino acid in the translated input sequence, the odds ratio of this position belonging to a certain genotype vs. this position belonging to another genotype, is computed. Odd scores greater than 1 indicate positive discriminants, while odd scores smaller than 1 indicate negative discriminants. The use of discriminants prevents overprediction due to positions that are highly indicative of individual genotypes.

In the end, the total number of positive and negative discriminants are tallied, respectively. This allows the computation of the discriminant odds ratio, which is given as the ratio of the number of positive discriminants against the number of negative discriminants for each genotype.

The genotype with the greatest ratio is the genotype that resembles the genotype of the input sequence with the greatest likelihood. This ratio is converted to a Z-score in order to make it independent of length and easier to interpret. The genotype that achieves the highest Z-score is the predicted genotype of the query sequence. To minimize false positives, predictions are only made for input sequences whose predicted genotype supersedes the Z-score threshold of 2.0.

Recombinant sequences are predicted in the same manner as presented for the NCBI prediction server, i.e. with a sliding-window approach.

This method also does not deal with HBV dual infections explicitly and the resulting graphical output is of only limited usefulness for interpretation.

1.4 Purpose

At the current point in time there are no computational methods available that are specifically concerned with the detection of HBV dual infections from a mixture sequence. Therefore, the objective of this thesis is to develop and validate a web implementation of the HBV dual infection model, which is able to identify the genotypes of infections with two viral strains of HBV.

Chapter 2

Materials and Methods

Mallory et al. already used the notion of ambiguous positions in electropherograms resulting from Sanger sequencing in order to detect infections by subpopulations of more than one viral genotype [38]. However, they were not able to determine the genotypes of these subpopulations from the sequencing data. The approach presented in this chapter can exactly do that. It integrates the concept of ambiguous positions present in a mixture sequence into a scoring scheme that allows the detection of HBV dual infections.

2.1 Prediction Method

The prediction approach presented in this section is based on supervised statistical learning on genotype-specific nucleotide distributions and genotypic classification of input sequences according to a maximum likelihood classifier.

2.1.1 The Input Sequence

The input sequence of this tool is a partial or full HBV DNA sequence, which was generated by Sanger sequencing. We will first describe the procedure used to generate such a sequence and then discuss the formal definition of the input sequence.

Sanger Sequencing

Sanger sequencing is a method for sequencing the DNA of an organism using a chain-termination mechanism [61]. This method requires a primer that is complementary to the template strand and a DNA polymerase that extends this primer. However, instead of providing only the conventional four deoxynucleotides for DNA elongation, one augments these nucleotides with di-deoxynucleotides, which are marked using fluorescent dyes. The incorporation of a di-deoxynucleotide

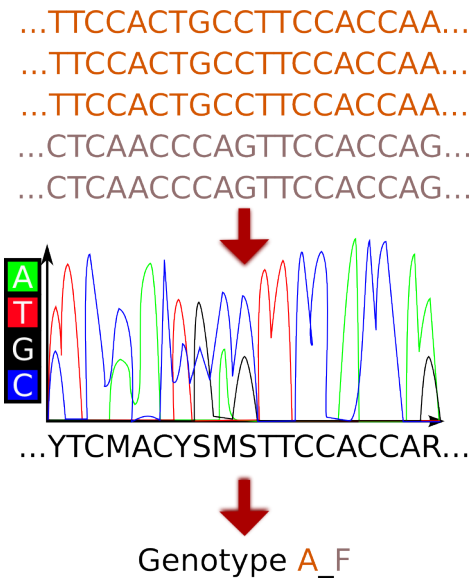


Figure 2.1: Ambiguities in chromatograms hint at dual infections. The top of the image shows two viral subpopulations of different genotypes (orange for genotype A and gray for genotype F), which are represented by their corresponding DNA sequences. The chromatogram resulting from Sanger sequencing is shown in the middle. It contains ambiguous reads of cardinality 2 at each position for which the genomes of the two viral subpopulations differ from each other. Such ambiguous positions are visible as double peaks in the chromatogram. Each color in the chromatogram represents one nucleotide. The colors identifying the nucleotides in the chromatogram are given on the left. The bottom of the chromatogram shows the resulting sequence of nucleotides, which indicates ambiguous reads using IUPAC nucleotide ambiguity codes. Such a sequence is the input to this HBV genotyping approach.

into the sequence of nucleotides results in termination of the elongation process. In consequence, this method produces a set of DNA fragments that are each terminated by a di-deoxynucleotide.

The presence of a certain di-deoxynucleotides in a fragment can be detected according to the specific wavelength that is emitted by its marker dye. Since the lengths of the individual fragments differ, they can be put into the correct order by chromatographic methods, such as high pressure liquid chromatography. This is possible because short fragments pass through the chromatographic channel more quickly than longer fragments, which are delayed. The sequence of nucleotides is determined by photometric measurement of the wavelengths emitted by fragments reaching the end of the chromatographic channel.

The result of Sanger sequencing is a chromatogram, which shows the nucleotidic signals at each position in the sequence as color-coded peaks. Computational analysis of such chromatograms leads to the final nucleotide sequence. It is possible for two or more peaks to occur at the same position in the chromatogram. Such positions are called ambiguous or mixed positions. They are

represented by IUPAC nucleotide ambiguity codes [9] in the final DNA sequence. Table 1 gives an overview of the available ambiguity codes. The mechanism of converting a chromatogram to a sequence of nucleotides in string representation is referred to as base calling. Results of Sanger sequencing may vary depending on the specific base calling procedure employed.

The ambiguous positions in a HBV sequence are of exceptional concern for the approach presented in this chapter because ambiguous reads are indicative of viral subpopulations that differ in their genomic makeup. A small number of ambiguous positions most likely indicates a viral quasispecies, which is a population of viruses with slightly different genomes that arose in the process of Darwinian evolution. A large number of ambiguous positions on the other hand rather indicates subpopulations of different genotypes. This observation is the cornerstone of our approach. The implications of ambiguous positions and their usage is illustrated in figure 2.1.

Formal Definition of the Input Sequence

The input of the program is a sequence $R = (r_i, \ldots, r_{i+n})$ with $r_j \in B = B_1 \cup B_2 = \{A, C, G, T\} \cup \{M, R, W, S, Y, K, V, H, D, B, X, N\}$, the alphabet of sequencing reads. Here, the set B_1 represents the conventional DNA bases, while B_2 represents the IUPAC ambiguity codes. Internally, reads are represented as tuples (b_1, b_2) with $b_1, b_2 \in B_1$ for ambiguous reads of cardinality 2 and as $b \in B_1$ for unambiguous reads. Therefore, the cardinality of a read r_j is given by the number of tuple elements. This means that $|r_j| = 1$ for A, C, G, and T, while $|r_j| = 2$ for M (A or C), R (A or G), W (A or T), S (C or G), Y (C or T), and K (G or T).

Only reads with a cardinality $|r_j| \leq 2$ are scored in our method. The reason for this is that we are only concerned with single and dual infections in this approach and not, for example, with triple infections. In addition, reads with $|r_j| > 2$ are very rare and might be due to errors in sequencing. However, if there were an incentive to consider infections with more than two viral strains, it would not be difficult to extend the already existing probabilistic framework.

The input sequence can be either a complete HBV genome sequence or a partial genome sequence. The subgenomic regions that are supported are listed in table 2.1.

Subgenomic region	Start position	End position
RT	130	1161
SHB	155	835
X	1374	1838
CORE	1901	2458
FULL	1	3221

Table 2.1: Supported subgenomic regions for the input sequence. The start and end positions of the intervals are given with respect to the reference sequence AM282986 [55].

		Bases							
		Pos 1				Pos 2			
		G	A	T	C	G	A	T	C
Genotypes	A	0.06	0.0004	0.97	0.014	0.002	0.0004	0.98	0.012
	B	0.0004	0.002	0.043	0.95	0.002	0.002	0.99	0.002
	C	0.002	0.001	0.2	0.78	0.001	0.002	0.99	0.005
	D	0.002	0.006	0.14	0.85	0.0004	0.006	0.99	0.0004
	E	0.008	0.008	0.98	0.01	0.0008	0.004	0.99	0.0008
	F	0.003	0.003	0.01	0.97	0.003	0.003	0.98	0.003
	G	0.025	0.025	0.125	0.825	0.025	0.025	0.925	0.025
	H	0.0125	0.0125	0.0125	0.96	0.0125	0.0125	0.962	0.0125
	I	0.02	0.02	0.85	0.1	0.02	0.02	0.93	0.02

Figure 2.2: Detail of genotype-specific nucleotide distributions. Each row in this table represents the nucleotide profile of the genotype given in that row. The columns list, for each position, the four DNA bases. Entries in the table represent the probability of the base in the current column, given the model in the current row.

2.1.2 Genotype-specific Nucleotide Distributions

In order to judge what genotype gt a nucleotide at a certain position in the input sequence R belongs to, genotype-specific nucleotide distributions are used. A genotype-specific nucleotide distribution measures for each position (in relation to a reference sequence) the probability of seeing each nucleotide in a sequence of this genotype. Such a distribution therefore represents how we believe typical sequences of this genotype are composed. This is why these distributions are also referred to as nucleotide profiles. Figure 2.2 shows a detail of all these profiles.

In our approach, we made use of nine nucleotide profiles, one for each of the considered genotypes A to I. A nucleotide profile for genotype J was not used because no sequences of that genotype were available at the time of profile generation. The manner in which these profiles were generated by my advisor is detailed in the next section.

Generation of Nucleotide Profiles

A set of HBV DNA sequences was downloaded from NCBI GenBank in August 2010. Genotypes were labeled by parsing the annotations from NCBI. Thereafter, unlabeled sequences (1297 complete genome sequences) and known recombinants (81 recombinants of which 53 were labeled with genotypes) were removed from the data set. This resulted in a data set consisting of 2258 labeled complete genome and 2467 labeled gene S sequences.

In the next step, the sequences in the data set were aligned pairwise to the well-annotated 3221 bp genotype A reference strain AM282986 [55]. This strain was chosen because the intervals of all its subgenomic regions are annotated. To ensure that all aligned sequences shared the length of the reference sequence, all positions introducing a gap into the reference sequence were removed.

Entries in the Nucleotide Profiles

The nucleotide counts from the previously generated alignment were used to determine the probabilities for each nucleotide at each position, giving rise to a multinomial distribution $p_{gt,j}$ over the four bases A, C, G, and T, where $gt \in GT_s = \{A, B, C, D, E, F, G, H, I\}$ refers to genotype and $j \in \{1, 2, \ldots, 3221\}$ gives the position in the reference sequence. The probability for base $b \in B_1$ to appear at position j in a sequence of genotype gt is given by $p_{gt,j}[b]$ with $\sum_{b \in B_1} p_{gt,j}[b] = 1$.

2.1.3 Adjusting the Input Sequence

By aligning the input sequence R to the reference sequence AM282986 and truncating it, the adjusted input sequence R' is yielded. It differs from the original input sequence R in two regards. First, it may contain gap characters due to the alignment, i.e. $r'_j \in B' = B \cup \{-\}$, the gapped read alphabet. Gaps are not scored in our method. Second, R' is of the same length as the reference strain, that is, 3221 bp. In this way each position r'_j in the aligned input sequence $R' = r'_1, \ldots, r'_{3221}$ with genotype gt is distributed according to $p_{gt,j}$.

This gives rise to a method of genotyping that is based on computing the likelihood of the aligned input sequence R' for each prediction model $m \in M = GT_s \cup GT_d$, where GT_s is the set of all single infection genotypes and $GT_d = \{(A, A), (A, B), (A, C), \ldots, (A, H), (B, B), (B, C), \ldots, (B, H), \ldots, (I, I)\}$ is the set of all possible genotype combinations for dual infections. The cardinality of M is given by $|M| = |GT_s| + |GT_s| \cdot \frac{|GT_s|+1}{2}$, which is 54 in this case.

2.1.4 Likelihood Computation

The prediction model is based on the following conditional distribution assumption:

$$r'_j \mid gt \sim p_{gt,j}$$

It states that each positional read, given a certain genotype, is distributed according to its entry in the corresponding nucleotide profile. Therefore, the conditional positional probabilities of each read are independent of each other. This allows us to compute the likelihood of the adjusted input sequence as

$$\Pr(R' \mid m) = \prod_{j=1}^{N} \Pr(r'_j \mid m)$$

where N is the length of the reference sequence.

The positional likelihood $\Pr(r'_j \mid m)$ depends on two parameters: the current read r'_j and the model m under consideration. Therefore, four cases have to be distinguished in order to compute the positional likelihood for a single model.

Single Infection and Unambiguous Read

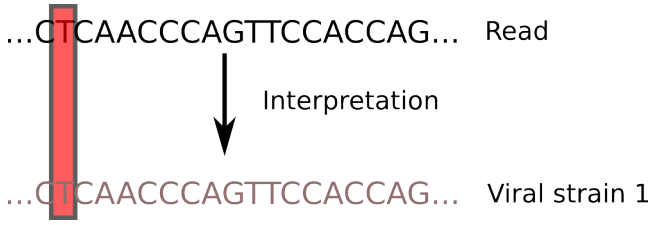


Figure 2.3: Biological interpretation of an unambiguous read for a single infection model. Only a single viral strain of some genotype is present and the current position, which is indicated by the red rectangle, is not ambiguous.

Let $m = gt$ be a single infection model and $|r'_j| = 1$ with $r'_j = b$. The likelihood to have base b at position j for genotype gt is given by the corresponding nucleotide distribution:

$$\Pr(r'_j \mid m) = p_{gt,j}[b]$$

Figure 2.3 illustrates this case.

Single Infection and Ambiguous Read

Let $m = gt$ be a single infection model and $r'_j = (b_1, b_2)$ be an ambiguous read. In this case, both nucleotides b_1 and b_2 have to appear at position j in a HBV sequence of genotype gt simultaneously, as illustrated in figure 2.4. Therefore the likelihood for the current position is given by the product of the entries in the nucleotide profiles:

$$\Pr(r'_j \mid m) = p_{gt,j}[b_1] \cdot p_{gt,j}[b_2]$$

The observation of an ambiguous position for a single infection model can be explained by the concept of quasispecies, which describes a population of genetically similar viruses that exhibit mutations at certain positions, but all originate from the same initial strand. Another possible explanation for encountering ambiguous reads for single infection models are sequencing errors.

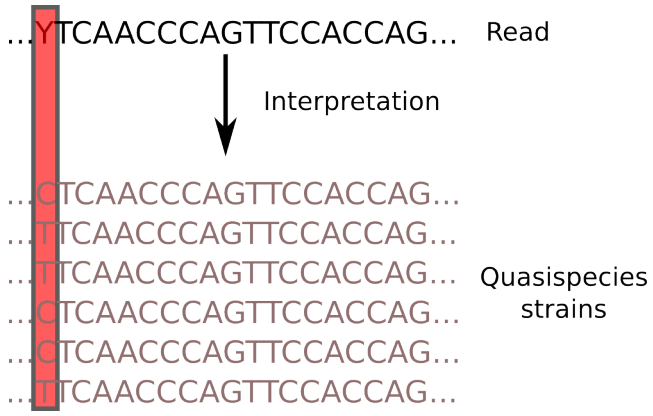


Figure 2.4: Biological interpretation of an ambiguous read for a single infection model. As indicated by the red rectangle, there is an ambiguity at the current position. This ambiguity is possibly the result of a point mutation in the present viral quasispecies.

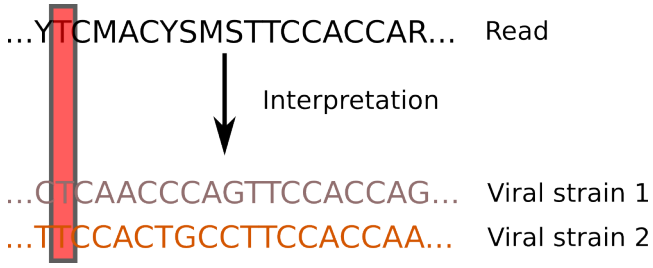


Figure 2.5: Biological interpretation of an unambiguous read for a dual infection model. There have to be two viral strains because a dual infection is presumed. As indicated by the rectangle, there is no ambiguity though. This means that both strains have to carry the same nucleotide at this position.

Dual Infection and Unambiguous Read

Let $m = (gt_1, gt_2)$ be a dual infection model and $r'_j = b$. In this case the base b has to appear at position j in sequences of both genotypes. The biological interpretation of this case is the following: although there exists a dual infection involving two genotypes, the bases of the two strains do not differ at position r'_j, see figure 2.5. This case is not unusual since the overall inter-group divergence of genotypes need only be 8% or greater. We can model the positional likelihood as

$$\Pr(r'_j \mid m) = p_{gt_1,j} \cdot p_{gt_2,j}$$

Dual Infection and Ambiguous Read

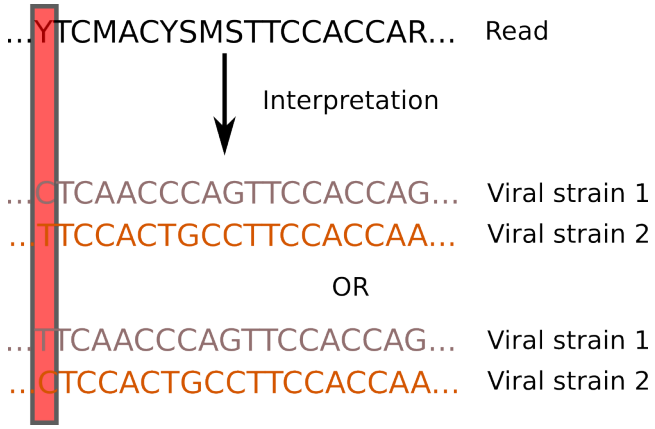


Figure 2.6: Biological interpretation of an ambiguous read for a dual infection model. Here, two viral strains are present and an ambiguous read is present, as highlighted by the rectangle. It is not known to which strain the bases of this ambiguity belong. Hence, there are two possible events, which are shown at the bottom.

Let $m = (gt_1, gt_2)$ be a dual infection model and $r'_j = (b_1, b_2)$ be an ambiguous read of cardinality 2. This case is difficult because the phase information of the two bases is lost in the sequencing process, that is, we do not know to which genotype each of the two bases b_1 and b_2 belong to. This means that b_1 could belong to genotype gt_1 or to genotype gt_2. The same is true for b_2. This corresponds to two biological events: in the first, b_1 is present in a strain of genotype gt_1 and $b2$ is present in a strain of gt_2, in the second, b_1 is present in a strain of genotype gt_2 and b_2 is present in a strain of genotype gt_1, as shown in figure 2.6.

The probability of the first event is $p_{gt_1,j}[b_1] \cdot p_{gt_2,j}[b_2]$, while the probability of the second event is $p_{gt_1,j}[b_2] \cdot p_{gt_2,j}[b_1]$. Because both events are mutually exclusive, there is no intersection between the two of them. As a consequence, we can compute the probability of the first or the second event occurring by simply adding the event probabilities:

$$\Pr(r'_j \mid m) = p_{gt_1,j}[b_1] \cdot p_{gt_2,j}[b_2] + p_{gt_1,j}[b_2] \cdot p_{gt_2,j}[b_1]$$

2.1.5 Improving the Likelihood Computation

The likelihoods of each model can be modeled more accurately by including an error term and excluding certain positions from scoring. These techniques are delineated in the next two sections.

Error Model

The error model is used in order to curb the impact of sequencing errors on the fidelity of predictions. If we did not deal with this problem, sequencing errors at positions that are predictive of the actual genotype or predictive of another genotype could possibly lead to misclassifications.

There are basically two types of sequencing errors: indel (insertion or deletion) errors and substitution errors [21]. In the first class of errors, one or multiple nucleotides are inserted or deleted from the actual sequence. A substitution error on the other hand is present when an incorrect nucleotide takes the place of the actual nucleotide.

Although one might think that aligning the input sequence to the reference sequence reduces the impact of indel errors in sequencing, this is only partially true. First, it is much harder to detect short indels, for example indels of length one, than longer indels using alignments. Second, detection of indels is also dependent on the position of their occurrence. Therefore, we have to consider the overall reported position-specific error rates for Sanger sequencing, which range from approximately 0.001% [48] up to less than 1% [29].

Because of the high sensitivity of the scoring system regarding individual positions, we are pessimistic about the error and set the position-specific error rate p_{err} to 1%. This provides a reasonable trade-off between added noise and robustness of predictions.

To include the probability of a sequencing error into the likelihood computation we make use of the law of total probability. This law states that the probability of an event A can be computed on the basis of n events B_i that partition the sample space:

$$\Pr(A) = \sum_{i}^{n} \Pr(A \mid B_i) \cdot \Pr(B_i)$$

For us, the partition contains two events, err (the presence of a sequencing error) and $\neg err$ (the absence of a sequencing error). Therefore we have:

$$\begin{aligned} \overset{*}{\Pr}(r'_j \mid m) &= \Pr(r'_j \mid m, err) \cdot \Pr(err) + \Pr(r'_j \mid m, \neg err) \cdot (\Pr(\neg err)) \\ &= 1 \cdot p_{\text{err}} + \Pr(r'_j \mid m) \cdot (1 - p_{\text{err}}) \end{aligned}$$

In this calculation, we substituted $\Pr(r'_j \mid m, \neg err)$ with $\Pr(r'_j \mid m)$, the positional likelihood for which we did not assume any errors. Additionally, we used the fact that $\Pr(err)$ and $\Pr(\neg err)$ are complementary to each other, allowing us to set $\Pr(\neg err) = 1 - p_{\text{err}}$. The conditional probability $Pr(r'_j \mid m, err)$ was set to 1.

In this way, the likelihood of improbable reads is at least increased to the probability of an error p_{err}, while the likelihood of probable reads is at least decreased to $1 - p_{\text{err}}$. Thereby, the impact of such positions is reduced, making the scoring system more robust regarding sequencing errors.

RT AA position	Converted nucleotide position	Relevance
80	367, 368, 369	Adefovir resistance
169	634, 635, 636	Lamivudine/entecavir resistance
173	646, 647, 648	Lamivudine resistance
180	667, 668, 669	Lamivudine resistance
181	670, 671, 672	Adefovir resistance
184	679, 680, 681	Lamivudine/entecavir resistance
194	709, 710, 711	Tenofovir resistance
202	733, 734, 735	Entecavir resistance
204	739, 740, 741	Entecavir resistance
236	835, 836, 837	Adefovir resistance
250	877, 878, 879	Entecavir resistance

Table 2.2: Positions associated with resistance to antiviral drugs [63, 77]. The converted nucleotide positions represent the three nucleotide positions in the full HBV genome that correspond to the amino acid positions in the reverse transcriptase region in the first column. These positions were excluded from scoring.

Exclusion of Positions

Several positions in the HBV genome are associated with escape mutants. Escape mutants are given rise to by an increase in selective pressure, for example through vaccination [5] or antiviral treatment [64]. It also known that HBV therapy is first and foremost practiced in highly-developed regions, such as in Europe. Therefore, our nucleotide profiles are biased towards genotypes that predominantly occur in these regions, for example towards genotypes A and D, which predominate in Europe [62]. To reduce this bias, two sets of positions were excluded from scoring.

The first set contains positions that are associated with mutations leading to HBeAg negativity, which provides a means of escaping the response of cytotoxic T-lymphocytes [15] and often arises as a consequence of interferon therapy [47]. This set of positions is based on a publication by Erhardt et al. [14] from 2000. It consists of nucleotide positions 1762, 1764, 1896, and 1899.

The second set contains positions that are associated with resistance against antiviral drugs [63,77]. Because these positions were given in terms of the reverse transcriptase subgenomic region on the amino acid level level, they had to be converted to the nucleotide level first. Here, we chose to ignore the complete codons corresponding to the amino acid positions. Table 2.2 shows a listing of these positions.

The excluded positions of both sets discussed are united in the exclusion set X.

Adjusted Likelihood Computation

The adjusted likelihood that includes the error term and excludes certain positions from scoring is given by:

$$\overset{*}{\Pr}(R' \mid m) = \prod_{\substack{j=1 \\ j \notin X}}^{N} \overset{*}{\Pr}(r'_j \mid m)$$

2.1.6 Maximum Likelihood Model Selection

The predicted model is determined by maximum likelihood classification. It is chosen in the following way:

$$m = \arg\max_{m \in M} \overset{*}{\Pr}(R \mid m)$$

2.1.7 Posterior Computation

The posterior probability of a model m given an aligned input sequence R' is yielded using Bayes' theorem and the law of total probability:

$$\begin{aligned} \Pr(m \mid R') &= \frac{\Pr^*(R' \mid m) \cdot \Pr(m)}{\Pr(R')} \\ &= \frac{\Pr^*(R' \mid m) \cdot \Pr(m)}{\sum_{m_i \in M} \Pr^*(R' \mid m_i) \cdot \Pr(m_i)} \end{aligned}$$

Here, the prior probabilities $\Pr(m_i)$ for $m_i \in M$ are assumed to be uniformly distributed, i.e. $\Pr(m_i) = \frac{1}{|M|}$.

The posterior probability is used as a confidence measure for the predicted model. A high posterior probability for a model means that the its likelihood is considerably larger than that of all other models. A low posterior probability on the other hand indicates the contrary, namely that there are models with similar likelihoods. If the posterior probability of the predicted model is low, then this indicates a certain extent of uncertainty about the model. Such uncertainty could be the consequence of an input sequence that is not typical for its genotype, e.g. a recombinant or erroneous sequence.

2.1.8 Recombinant Analysis by a Sliding-Window Approach

Basic Procedure

To determine the genotypes that are present in a recombinant sequence, a sliding-window approach is employed. In this method the input sequence is divided into so-called windows, which are subsequences of the original sequence. Then, for each window, a genotype is predicted using maximum likelihood classification.

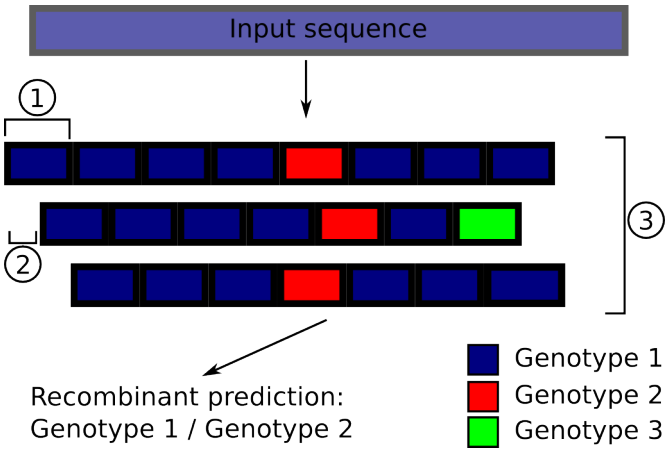


Figure 2.7: The sliding-window approach to recombinant analysis. Each rectangle represents a window with color-coded genotype. The circled numbers indicate one of the three parameters to this approach, each: 1. window size, 2. window offset, 3. maximal depth. Genotype 3 is not included in the recombinant prediction because there are less than three windows of that genotype.

In order to be able to pinpoint the interval in which the recombination occurred, multiple layers of windows are generated. For each layer, the starting positions of the windows is shifted by some amount such as to maximize window overlap. Each layer is assigned a number (starting at 1), which is called the depth. For example, for a maximal depth of 3, three layers are created. Figure 2.7 shows the sliding-window approach and its parameters.

Formal Definition

Let $k, l \in \mathbb{N}$ be the start and end positions of the current window, respectively. Then the likelihood of that window is given by

$$\overset{*}{\Pr}(r'_k, r'_{k+1}, \ldots, r'_l \mid m) = \prod_{\substack{j=k \\ j \notin X}}^{l} \overset{*}{\Pr}(r'_j \mid m)$$

The window interval (k, l) depends on the current depth and the window count in this depth, as well as the preset offset value and window size. Let $m \in \mathbb{N}$ be the current depth and $n \in \mathbb{N}$ be the window count for this depth, with an offset $o \in \mathbb{N}$ and a window size $w \in \mathbb{N}$. Then we have $k = (m-1) \cdot o + (n-1) \cdot w$ and $l = (m-1) \cdot o + n \cdot w$.

The most likely model for this window is defined as

$$m = \arg\max_{m \in M} \overset{*}{\Pr}(r'_k, r'_{k+1}, \ldots, r'_l \mid m)$$

Parameters

Window size	Window offset	Maximal window depth
300	−1 (auto)	3

Table 2.3: Default values for the sliding-window approach.

There are three parameters to the sliding-window approach: window size, window offset, and maximal window depth, which are shown in figure 2.7.

The window size is the number of nucleotides encompassed by each window. This value defaults to 300 since it has been successfully used by other approaches [59, 43]. This parameter is key to recombinant prediction: if the window size is too large, small recombinant segments cannot be recognized. On the other hand, a too small window size would cause individual positions in a window to carry too much weight, resulting in incorrect predictions.

The window offset is the number of nucleotides each successive layer of windows is shifted to the right. At its default value of −1 the offset is computed such that it maximizes the overlap of windows between depths. In this way, each window provides maximal new information.

The maximal window depth is the number of window layers used. The greater the maximal window depth, the more accurately one can determine the positions at which recombination events occurred, albeit at a cost in running time. The default value for this parameter is 3 because it provides a good trade-off between accuracy of recombinant detection and performance. It is also used by the NCBI HBV genotyping server [59].

In general, the length of each window is given by the window size parameter. It defines the number of nucleotides that are analyzed as a single unit. The only exception to this rule is made for the last layer of windows. Here, the last window is extended to the the end of the sequence. This is done in order to also cover the last sequence segment.

Threshold for Recombinant Genotypes

Only those models that are predicted in at least three windows are considered to take part in recombination. If at least two such models are found, the corresponding sequence is considered a recombinant sequence.

2.1.9 Numerical Considerations

The positional likelihoods $\Pr(r'_j \mid m)$ are often very small and would require a considerable number of bits to be stored. Due to the limitations of floating point data structures (64 bits for double precision values) these computed data cannot be stored directly without imprecision due to rounding. Therefore, the entries in the nucleotide profiles were logarithmized and all computations were performed using natural logarithms. To compute the likelihoods, the following two logarithmic equalities were employed:

$$
\begin{aligned}
\log(a \cdot b) &= \log(a) + \log(b) \\
\log(a + b) &= \log\left(a \cdot \left(1 + \frac{b}{a}\right)\right) \\
&= \log(a) + \log\left(1 + \frac{b}{a}\right) \\
&= \log(a) + \log\left(1 + \exp\left(\log\left(\frac{b}{a}\right)\right)\right) \\
&= \log(a) + \log\left(1 + \exp\left(\log(b) - \log(a)\right)\right)
\end{aligned}
$$

In the second equation we have $a > b$, otherwise we would have to switch the two variables in order to restrict the number of decimal values of the term $\exp(\log(b) - \log(a))$.

2.1.10 Implementation

The prediction model was implemented in Python 2.6.6 [71]. To increase the performance of the implementation, the Python code was converted to Cython. The web front-end was implemented with PHP 5.3.6 [2]. To generate a graph for representing genotyping results, the JpGraph plotting library was used.

Communication between the PHP front-end and the Cython back-end is facilitated by a MySQL database. Data exchange between individual PHP scripts is performed using the internal PHP session variable.

2.2 Model Validation

To validate the prediction model, two different types of data sets were used: actual sequences from NCBI and a data set consisting of artificially generated sequences. All computations were done using the default settings for the sliding-window approach described in table 2.3 and a sequencing error probability p_{err} of 1%. Results were analyzed using R 2.13.1 [58]. We will first introduce the NCBI data sets.

2.2.1 NCBI Data Sets

To validate the prediction model on single infections and recombinants, we used different subsets from the initially downloaded NCBI data set, which was discussed in section 2.1.2.

Data Set For Recombinant Analysis

This data set contains a total of 82 recombinant sequences, of which the genotypic composition of the sequence is annotated for a subset of 53 sequences.

Complete Genome Data Set Without Recombinants

This data set contains 2258 complete genome HBV sequences with annotated genotypes and no sequences that are labeled as recombinant. Since this data set was used to train the model, that is, to generate the genotype-specific nucleotide distributions, we had to employ cross-validation in order to use it for validation purposes.

2.2.2 Cross-Validation

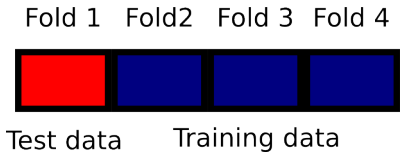


Figure 2.8: Data separation in cross-validation. Here, using $k = 4$, the data set is split into 4 folds. One then iterates over all folds and uses the current fold as the test data set (shown in red), while the elements of all other folds are used to form a new training data set. A new model is trained on this training data set and tested on the data contained in the current fold.

The method of k-fold cross-validation is utilized in supervised statistical learning to approximate the test error of a data set in sparse data scenarios, in which one wants to measure the test error using the training set. In this approach, the data set is separated into k folds, such that each of the folds has the same number of elements. One then iterates over all of the folds. The elements contained in the current fold are used as a test data set, while the elements contained in all other folds are used to train the model. In this process, k models are yielded. Each of these models is validated using a test set whose members are pairwise distinct from the elements of the training set that was used to train the model. In this way the overall bias of the error is reduced.

In the cross-validation we performed, the recombinant sequences were excluded from the NCBI data set. In addition, each of the folds was stratified such that its distribution of genotypes approximated that of the full data set. Stratification was necessary to provide some statistical robustness. Otherwise it could have been possible to obtain training data sets containing very few sequences of certain genotypes, thereby reducing the statistical representativeness of that data set and increasing the misclassification rate.

We performed 4-fold cross-validation. The parameter k was set to 4 for two reasons. It allowed us to keep all folds stratified and reduced the running time needed for cross-validation in comparison to larger values of k.

2.2.3 The Artificial Data Set

There is a paucity of data on identified dual infections with known genotypes. To alleviate this problem, mixture sequences that are representative of dual infections were generated in the manner described hereafter.

Two adjusted sequences R'_1, R'_2 with genotypes gt_1 and gt_2 are chosen from the basis sequence set. Then, a new, artificial HBV sequence R_{new} is generated on the basis of these two sequences by combining the reads at each position. For each pair of reads (r'_{1i}, r'_{2i}) the corresponding tuple of bases $(b \mid b \in r'_{1i} \vee b \in r'_{1i})$ is generated. The read in the new sequence $r'_{\text{new}\,i}$ is then given by the IUPAC ambiguity code that corresponds to this tuple of bases.

To generate the artificial sequence data set we used the previously created cross-validation folds of the NCBI data set. The sequences in each fold were subdivided into model-specific sequence sets S_i, with one S_i for each model. For each cross-validation fold we then iterated over all pairs of genotype-specific sequence sets (S_i, S_j). Note that S_i and S_j needed not be inequal.

Then, we performed the subsequent operations at most n times. Two sequences were randomly chosen, one from S_i, the other from S_j. Sequences were excluded from mixing if they either did not represent complete HBV genomes (sequences shorter than 3 kb) or if they were identical. In these cases, a new pair of sequences was randomly chosen if more than a single sequence of that genotype was available. Sequence labels were assigned to the mixed sequences by combining the labels of the sequences from NCBI according to their genotypes.

The rationale behind excluding sequences that are identical is that they do not match the definition of HBV dual infections, which states that two different viral strains should be present. Subgenomic sequences were also not included because otherwise sequences of differing sizes would be mixed, resulting in erroneous mixture sequences.

The final artificial data set contains 6482 sequences in total, with each genotype-combination occurring at most 50 times per cross-validation fold.

2.3 Result Representation

2.3.1 Nomenclature

Recombinant Sequences

Genotypes that are present in a recombinant sequence are separated by a slash. There is no consensus on how to order the genotypes that are present in a recombinant. We list recombinant genotypes lexicographically.

Dual Infections

The viral genotypes that are involved in a dual infection with two strains of HBV are separated by underscores. The order of the genotypes does not play a role. This means that an A_B dual

infection is essentially the same as a B_A dual infection. Dual infection genotypes are also listed in lexicographic order.

2.3.2 Result Graph

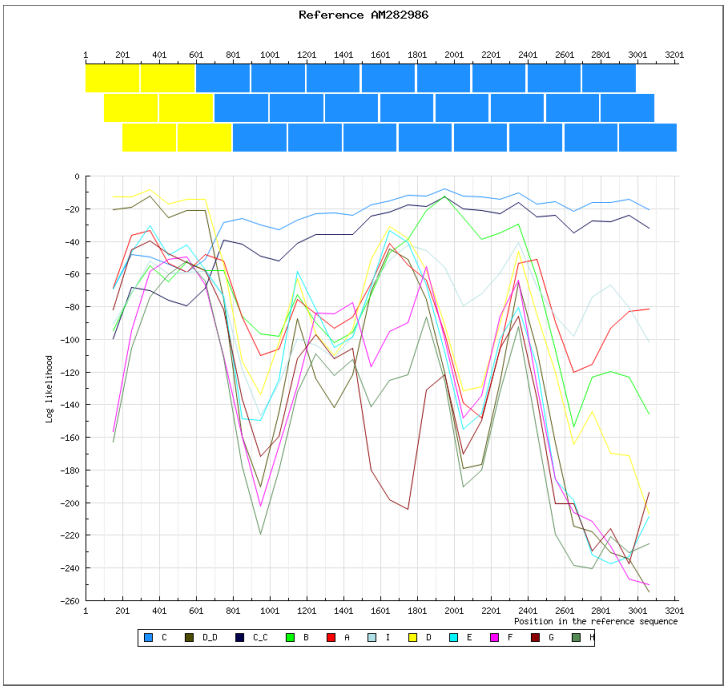


Figure 2.9: Graphical representation of genotyping results. The graph consists of two individual plots. The upper plot shows the windows generated in the sliding-window approach as well as their corresponding identified genotypes by color. The lower plot shows the likelihoods of all single infection genotypes and other high-scoring models. It uses the model-associated likelihoods of all windows as data points. This particular graph was generated on the basis of an annotated C/D recombinant sequence with GenBank accession number DQ478895.

The genotypes of a HBV sequence are depicted using a combined graph generated by JpGraph, which is shown in figure 2.9. The graph, which is based on that of the NCBI genotyping server, consists of two parts.

The upper part shows the windows generated by the sliding-window approach for recombinant detection. Each of the windows is colored according to the model with the greatest likelihood in that sequence interval. Hovering over a window with the mouse reveals additional information, for example the posterior probability of the predicted genotype.

The lower part of the graph shows a graph that contains one plot for each selected model, with the likelihood given on the y-axis and the position in the reference sequence given on the x-axis.

The individual plot data points are given by the positions of the windows in the sequence and their associated likelihoods.

This graph shows a maximal number of 14 models. Generally, all single infection genotypes and predicted window models are shown in this graph. Additional models are added to the graph on the basis of a posterior probability threshold for window models, which defaults to 0.1%.

The models in the legend are ordered from left to right in descending order according their posteriors over the whole input sequence. This means that the predominant genotype is always shown on the far left of the legend.

Chapter 3

Results

3.1 Analysis on Actual Data

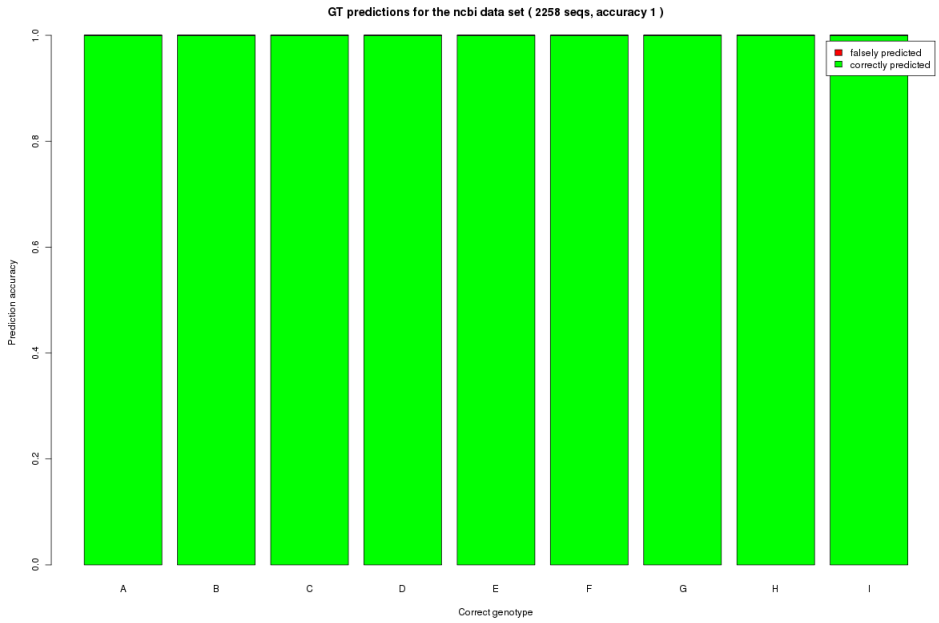


Figure 3.1: Genotype prediction results for the NCBI data set. The y-axis shows the prediction accuracy, i.e. the ratio of correct (green) and incorrect (red) predictions for all sequences of the genotypes listed on the x-axis.

Using 4-fold cross-validation on the complete genome NCBI data set without recombinants, we

correctly identified 2258 (100%) of 2258 sequences, as shown in figure 3.1.

The specificity of recombinant detection for the cross-validation data set was 98.7% (2228 of 2258), while the sensitivity of recombinant detection for an independent set of recombinnt sequences was 96.3% (79 of 82 recombinants), resulting in an overall recombinant detection accuracy of 98.6%.

For the set of recombinant sequences with explicitly labeled genotypes, we correctly identified 46 (86.8%) of 53 cases, as illustrated in figure 3.2. For this analysis, we did not consider the order of recombinant genotypes to play a role. This means that a sequence with genotype A/C/G was regarded as correctly predicted although it was labeled as G/C/A. The reason for this is that there is no consensus on the nomenclature of recombinants.

Our predictions on the sequences of the complete genome NCBI data set suggest no evidence of dual infections. In addition, the sequences in this data set had very few ambiguities, no dual infection labels, and were almost exclusively generated by clonal methods. Therefore, we could not use this data set to evaluate the performance of our tool for HBV dual infections, but rather had to test it on an artificial data set.

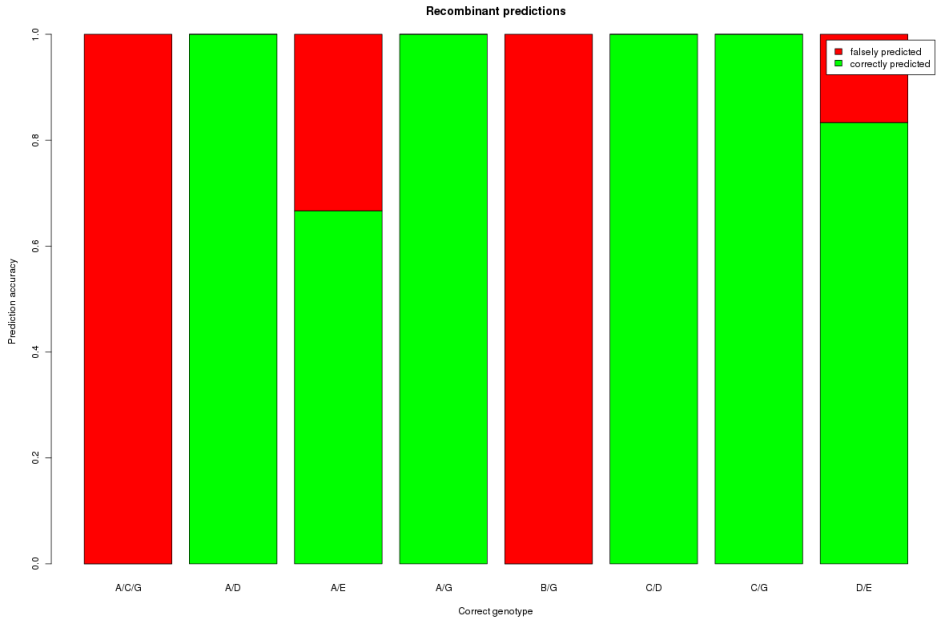


Figure 3.2: Recombinant prediction results. The y-axis shows the prediction accuracy, i.e. the ratio of correct (green) and incorrect (red) predictions against the total number of sequences for each of the recombinant genotypes listed on the x-axis.

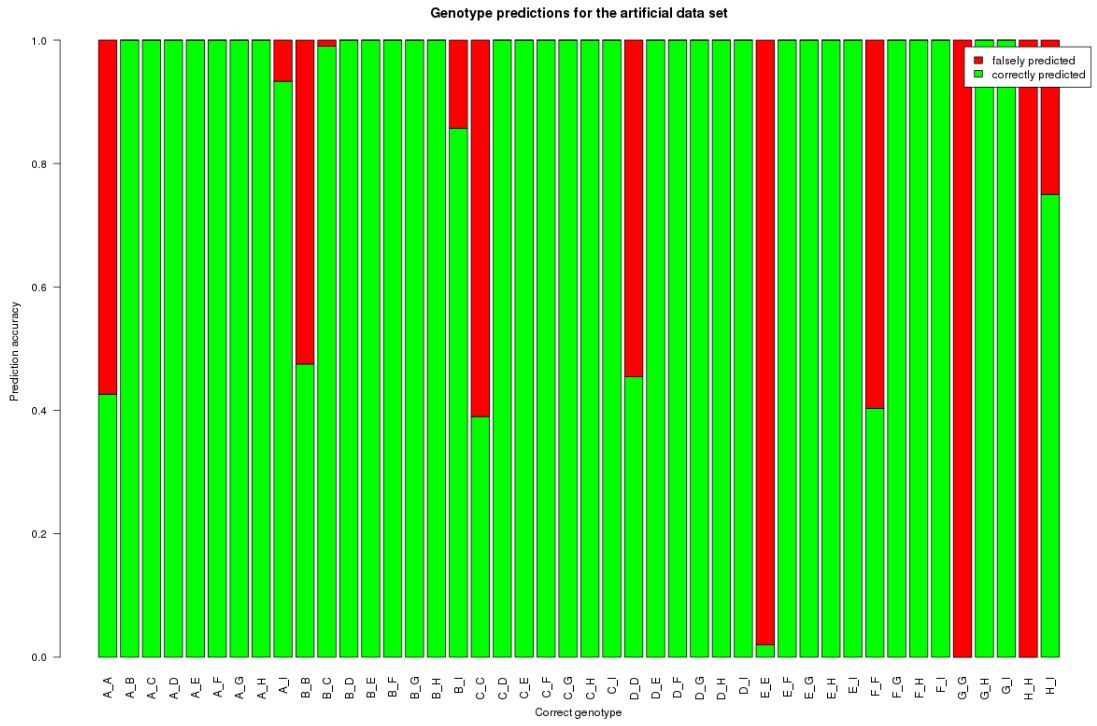


Figure 3.3: Results of genotype prediction for the artificial data set. The x-axis lists the actual genotypes, while the y-axis shows the percentual prediction accuracy normalized on the corresponding actual genotype. Green-colored bars indicate the ratio of correct predictions, while red-colored bars indicate the ratio of incorrect predictions. Evidently, the overall prediction accuracy for dual infections of the same genotype is much lower than that for dual infections with different genotypes. This bar plot does not list any I_I dual infections because there is only one genotype I sequence present in the NCBI cross-validation data set.

3.2 Analysis on Artificial Data

To determine the performance of our tool for mixture sequences, we performed 4-fold cross-validation on the artificial data set described in section 2.2.3. The results of genotype prediction for this data set are listed in table 3.1 and plotted in figure 3.3. The genotypes of 5622 (86.7%) of 6482 mixed sequences were correctly identified, with accuracies of 32.7% for dual infections of the same genotype and 99.9% for dual infections with different genotypes.

To analyze the performance of our tool on sequences labeled as dual infections of the same genotype in more detail, we used a more lax notion of correctness. Here, predictions for dual infections of the same genotype were also considered correct if a single genotype, which was present in the dual infection label, was predicted. In this analysis, 1164 (91.6%) of 1271 dual infections of the same

genotype were correctly identified.

	Same genotypes	Different genotypes
correct	416 (32.7%)	5206 (99.9%)
incorrect	855 (67.2%)	5 (0.1%)

Table 3.1: Dual prediction results for the artificial data set. In total, 6482 mixed sequences representing dual infections were analyzed. Shown are the number of correctly and incorrectly predicted sequences in the artificial data set for both categories (dual infections with strains of the same genotype and dual infections with strains of different genotypes), with the ratios for each category given in brackets.

Chapter 4

Discussion

In this chapter we will discuss the results obtained for the NCBI and the artificial data set.

4.1 Discussing the Results for the NCBI Data Sets

Here, we discuss the results for the NCBI data sets. This includes the identification of single infection genotypes, as well as the detection and identification of recombinant sequences.

4.1.1 Prediction of Single Infection Genotypes

The prediction of single infections worked very well (100% accuracy), as can be seen in figure 3.1. This shows that our prediction model not only has a high sensitivity regarding the detection of single infections, but also a high specificity regarding dual infections.

4.1.2 Prediction of Recombinants

Recombinant sequences were confidently detected by our approach, with an overall detection accuracy of 98.6%. When considering this result, however, one should keep in mind that the set of non-recombinant sequences was much larger (2258 sequences) than that of known recombinant sequences (82 sequences). As a consequence, the overall accuracy is influenced to a much greater extent by specificity of recombinant detection than by sensitivity. Unfortunately, this problem could not be corrected due to the sparseness of available data on recombinant sequences.

The explicit identification of genotypes for a set of genotype-labeled recombinant sequences was successful for the majority (86.7%) of recombinant sequences in this data set. The overall good results regarding the detection and identification of recombinants were to be expected due to the high accuracy of our tool for single infection sequences and the known reliability of the sliding-window approach [59, 43].

In the following we will discuss the seven recombinant sequences, whose genotypic composition we mispredicted.

Impact of Genotype I

Three (accession numbers EU835242, EU835241, and EU835240) of the seven mispredicted recombinant sequences were A/C/G recombinants. This type of recombination is typical for genotype I [22], which indicates their possible affiliation with this genotype. However, this led to problems with correctly identifying the genotypes of these sequences due to the inclusion of genotype I in the scoring scheme.

Genotype I was detected only in sequences that were also recombinant with other genotypes, such as genotypes G and C. This was possibly caused by the small number of positions that are indicative of genotype I [31]. The inclusion of genotype I therefore resulted in the misclassification of recombinant A/C/G sequences, which probably would have been correctly predicted otherwise.

Examination of Sample Sequences The sequence EU835242 is exemplary for the problem identified above. It was annotated as a recombinant of genotypes A, C, and G in NCBI. Our tool on the other hand identified the sequence to be predominantly of genotype I, more precisely as a C/G/I recombinant (figure 4.1a).

The results for the original genotype I sequence AB231908 were similar: it was identified as a C/I recombinant with evidence of genotypes A and G, see figure 4.1b. However, it should have been identified as a sequence purely of genotype I, presuming that this is the actual genotype of the sequence.

These examples underline the difficulty of differentiating between sequences of genotype I and A/C/G recombinants and also show that the genotypic signal of genotype I is not as strong as that of the other, undisputed genotypes. Our findings suggest that genotype I actually identifies only a set of complex recombinant sequences, but does not define a new genotype itself. Therefore, we decided to leave the choice of including or excluding genotype I from scoring up to the user in the final web service.

Analysis of Undetected Recombinants

In two of seven recombinant sequences our tool did not detect any recombination. The results for these two sequences are shown in figure 4.2 and discussed hereafter.

GQ161753 The sequence GQ161753 was predicted as genotype A, with sub-threshold evidence of recombination for genotypes E and D, see figure 4.2a. It was labeled as an A/E recombinant. To validate our finding we checked the sequence using the NCBI genotyping server, which identified the sequence as an A/E recombinant with evidence of genotypes G and C. This shows that, for this particular sequence, our tool was less sensitive than the NCBI genotyping service. On the

other hand one might argue that our prediction was quite close to the label, especially considering the seemingly complex recombinant nature of this sequence.

AB555499 The sequence AB555499 was labeled as a B/G recombinant, but was predicted to be of genotype B by our tool (figure 4.2b). This was confirmed by the NCBI HBV genotyping tool. Therefore, this sequence might actually have been mislabeled.

Analysis of the Remaining Misclassified Recombinants

In two of the misclassified recombinant sequences, additional recombinant genotypes were identified. The results for these sequences are shown in figure 4.3 and evaluated in the following.

GQ161775 The sequence with the GenBank accession number GQ161775 (figure 4.3a) was predicted as an A/E/E_E recombinant by our tool, although it was labeled as an A/E recombinant. Unfortunately it was not possible for us to further validate the finding of a dual infection in this case.

FN594768 The sequence FN594768 (figure 4.3b) was classified as an A/D/E recombinant instead of a D/E recombinant. This result could be partially confirmed by the NCBI prediction server, which identified this sequence as a D/E recombinant with evidence of genotypes A and G. In this case, the threshold for predicting recombinant genotypes might have been too low or this sequence might actually be a more complex recombinant than previously thought.

4.2 Discussing the Results for the Artificial Data Set

The artificial data set of mixed sequences representing dual infections was subdivided into two categories for validation: dual infections of different genotypes and dual infections with strains of the same genotype. In the following we will discuss the results for each of the categories separately, starting with the identification of dual infections with different genotypes. At the end of this section, we will give a remark about the drawbacks of using artificial data for this kind of analysis and comment on the used error value.

4.2.1 Prediction of Dual Infections of Different Genotypes

The low misclassification rate (0.1%) for predicted dual infections of different genotypes shows that we are able to confidently predict dual infections of HBV strains with different genotypes. The performance of our tool for dual infections belonging to this category is much better than that for the other two categories, i.e. dual infections with recombinants and dual infections with same genotypes. This is because the genotypic signals are much clearer if the genotypes of the two strains are different, as this equates to a nucleotide divergence of at least 8%.

All of the five misclassified mixture sequences of different genotypes, were given rise to by a mixture involving one recombinant sequence each.

Dual Infections of Different Genotypes with Recombinant Sequences

Dual infections of different genotypes in which at least one of the two strains is recombinant are more difficult to predict than conventional dual infections due to the fact that the signal of the predominant genotype in recombinants is much weaker than that in sequences with only a single genotype.

Example Mixture Sequence of EU939622 and AY800389

The mixture of sequences EU939622 and AY800389 was predicted as a single infection of genotype B, although one would have expected it to be a dual infection B_C since EU939622 is labeled as genotype C and AY800389 is labeled as genotype B. However, EU939622 is actually a B/C recombinant.

The graphs shown in figure 4.4 illustrate this problem well. The first recombinant segment is correctly identified as a dual infection B_C. The second recombinant segment on the other hand was predicted as genotype B.

This is because the segment of EU939622, which is recombinant with genotype B, leads to a deficiency of positions that are predictive of genotype C in the corresponding region in the mixture sequence. In consequence, the score of the correct dual infection model drops considerably in this region, making genotype B the highest-scoring model. So, if EU939622 were not recombinant with genotype B in this region, this sequence probably would not have been misclassified.

On a side note, we would like to mention he second recombinant region was probably not identified as a dual infection B_B due to the nucleotidic similarity of sequences of the same genotype.

Interpretation of Dual Infections with Recombinant Sequences

Dual infections with recombinant sequences are difficult to interpret regarding their recombinant genotype due to the fact that the phase information is lost in the sequencing process. This means, as a result of the recombination, one cannot say to which strains the individual genotypes belong. Therefore, such sequences are required to be interpreted by experts.

Let us consider the following example: a recombinant of two dual infections of genotypes A_B and A_C is predicted. There are two ways for this result to arise, which are illustrated in figure 4.5. In the first scenario, one strain has genotype A and the other strain is a B/C recombinant. In the second scenario on the other hand, one strain is an A/C recombinant and another strain is an A/B recombinant.

4.2.2 Dual Infections with Two Strains of the Same Genotype

The low rate (32.7%) of correctly predicted dual infections in which the viral strains involved share the same genotype demonstrates the inherent difficulty of identifying the genotypes of such sequences.

This is caused by the fact that the intra-genotypic variance of HBV sequences is much lower (less than 8%) than their inter-genotypic variance (over 8%). This means that strains of the same genotype are more similar than strains of different genotypes. Hence, the mixture sequence of sequences of the same genotype contains a relatively small amount of ambiguous positions, which are crucial for correctly detecting dual infections.

In consequence, the accuracy with which we can detect dual infections of the same genotype depends on the similarity of the two strains. For highly similar strains, usually only a single infection will be detected. If the two viral strains belong to different subgenotypes on the other hand, this increases the chance for identifying the dual infection correctly.

In order to show that the predictions for dual infections of the same genotype were not that far-off from the actual genotype, we performed another analysis on this data set in which we also considered predicted single infections to be correct if they were contained in the corresponding labels. Using this notion of truth, 91.6% of dual infections of the same genotype were correctly identified.

This indicates that the basis single infection of misclassified dual infections of the same genotype is usually correctly identified.

4.2.3 Drawbacks of Analysis on Artificial Data

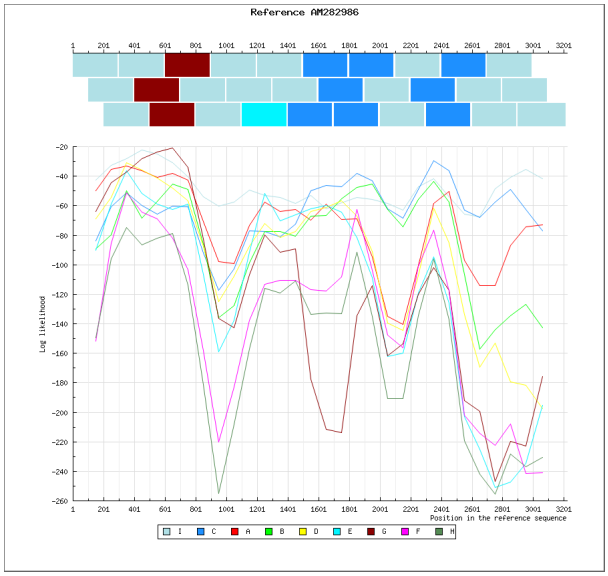
The performance of our tool in identifying dual infections was only tested on artificially generated data. This might have resulted in overly optimistic results due to the fact that the artificially generated dual infection sequences were not subjected to sequencing. In this way, the challenges of sequencing HBV dual infections were circumvented.

In particular, no basecalling needed to be performed for the artificial sequences, although this is especially challenging for dual infections due to the great number of ambiguous positions. As a consequence, the fidelity of the artificially generated sequences might be higher than that of actual data.

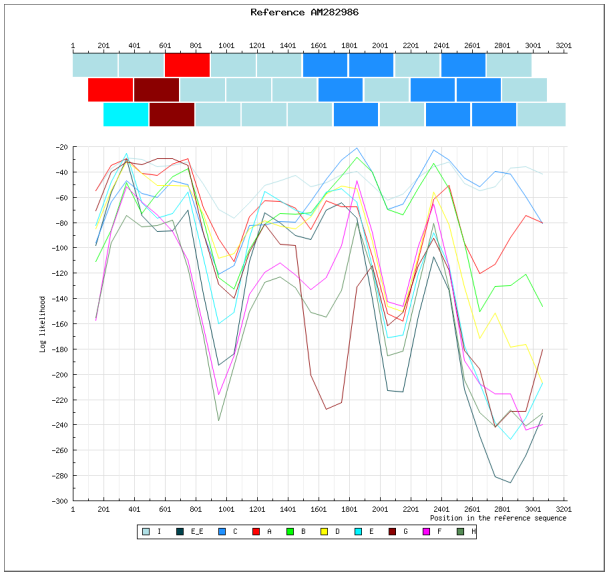
4.2.4 Commentary on the Error Value

The default position-specific probability of a sequencing error in an input sequence p_{err} was set to 1% according to values from publications and used in all computations. A more accurate approximation of the actual error value could have been yielded by more methodological approaches, such as testing the prediction performances of multiple models with different error values and then generating ROC curves for each of them.

However, this was not possible due to the sparseness of data on HBV dual infections.

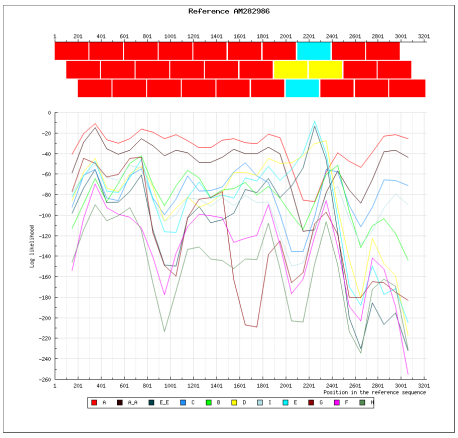


(a) The A/G/C recombinant sequence EU835242G was identified as a C/G/I recombinant due to the inclusion of genotype I in the scoring scheme.

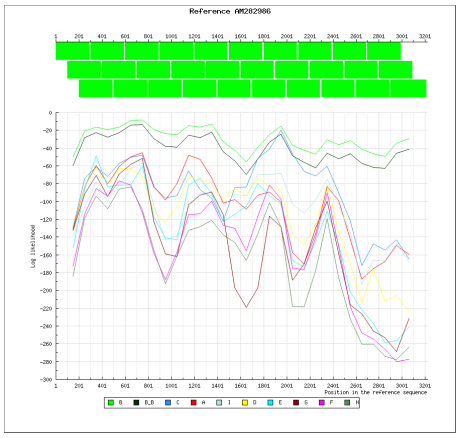


(b) Even the original genotype I sequence AB23190 was identified as a C/I recombinant showing evidence of genotypes A and G.

Figure 4.1: Impact of genotype I on the identification of recombinant HBV sequences. The recombinant genotypes of both shown sequences were not correctly identified.

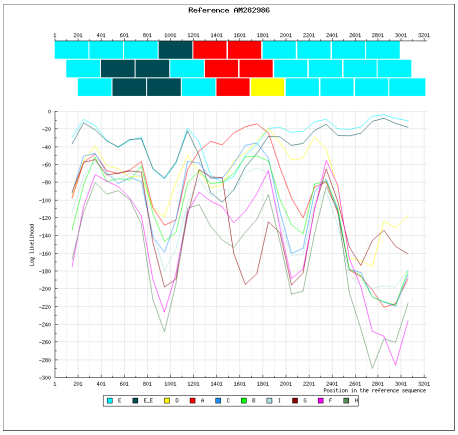


(a) GQ161753 was predicted as genotype A, although it was listed as an A/E recombinant.

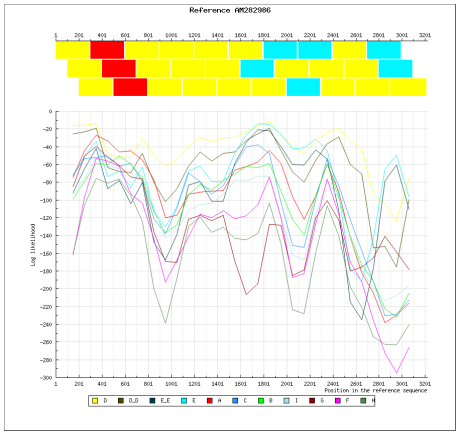


(b) AB555499 was identified as a genotype B sequence, even though it was labeled as a B/G recombinant.

Figure 4.2: Results for the undetected recombinant sequences. These sequences, although labeled as recombinant, were not detected as such in our analysis.

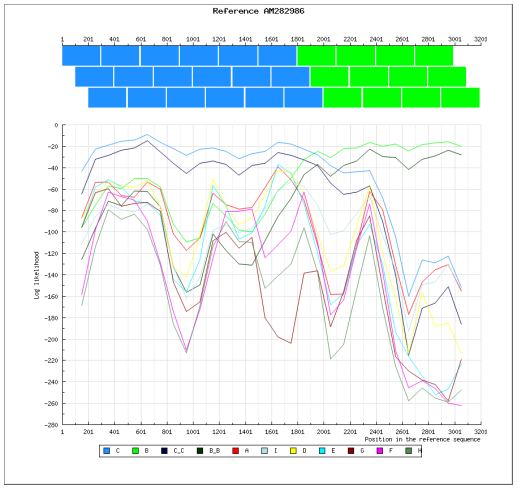


(a) GQ161775 was predicted as an A/E/E_E recombinant but was just labeled as an A/E recombinant.

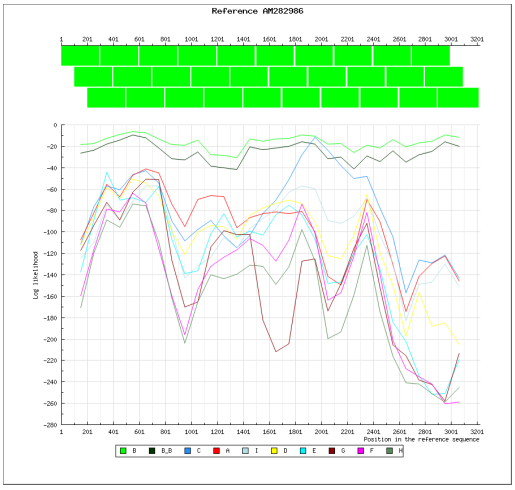


(b) The sequence FN594768 was predicted as an A/D/E recombinant even though it was labeled as an A/E recombinant.

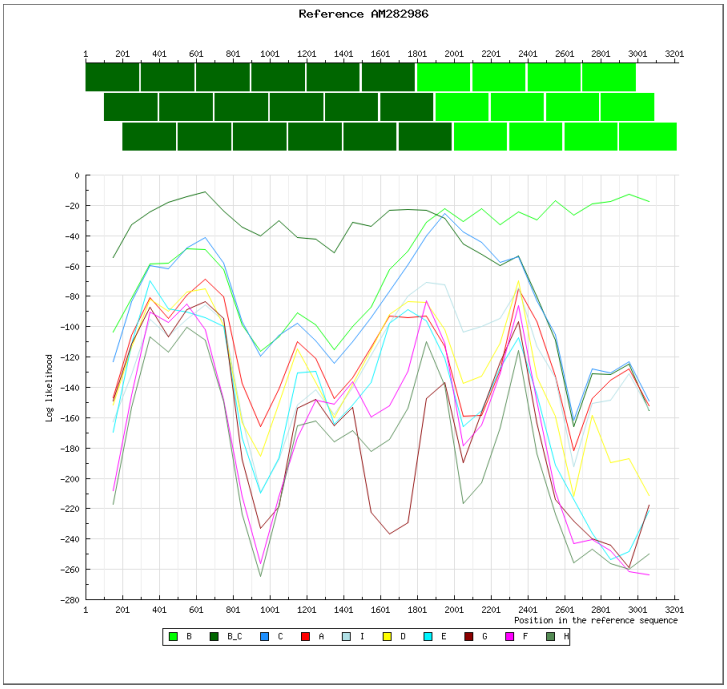
Figure 4.3: Results for the remaining misclassified recombinants. In both sequences, the presence of additional genotypes was predicted.



(a) B/C recombinant sequence EU939622 with predominant genotype C



(b) Genotype B sequence AY800389



(c) The mixture sequence of EU939622 and AY800389 was identified to predominantly be of genotype B, although it should have been identified as a B_C dual infection according to the labels of the two sequences.

Figure 4.4: Prediction of dual infections with recombinants. The upper two graphs show the results for the two sequences that constitute the artificial mixture sequence, of which the results are shown in the lower graph. The genotype of this mixture sequence was misclassified due to the recombinant nature of EU939622.

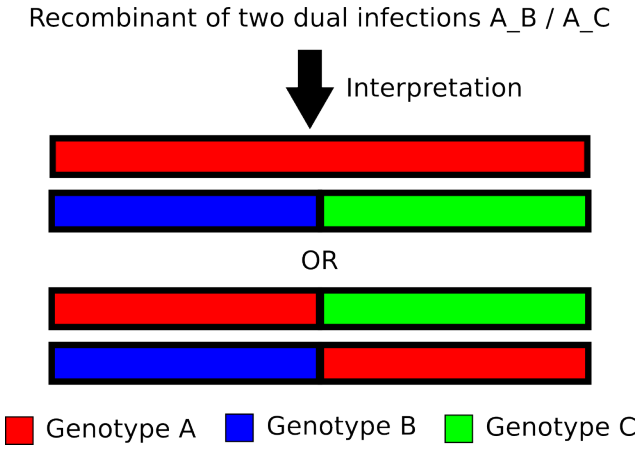


Figure 4.5: Scenario of a dual infection with a recombinant sequence. When the model predicts a recombinant of dual infections it is up to interpretation which kind of dual infection is actually present in the patient. In this case two different types of dual infections are possible. Viral strains are shown as rectangles with color-coded genotypes.

Chapter 5

Conclusion and Outlook

5.1 Usage and Areas of Application

The tool presented in this thesis is able to confidently identify the genotypes of patients infected with HBV strains of different genotypes. In contrary to many experimental methods for genotyping, our approach is specifically aimed at the sensitive detection of mixed HBV infections. Therefore, we recommend the usage of our method for the fast and reliable detection of HBV dual infections.

Only little information is available on the clinical impact of HBV dual infections due to the small number of undertaken studies that investigated this topic. It is crucial that the implications of HBV dual infections are studied more comprehensively in order to allow for improved therapy. We hope that this tool will prove to be an expedient resource in doing so.

5.2 Further Areas of Application

Currently this method supports only DNA sequences of hepatitis B virus. However, it could be easily extended to include hepatitis C virus (HCV), a single-stranded RNA virus that also causes hepatitis. Although HCV is not as prevalent as HBV with approximately 170 million infected individuals worldwide [19], it still poses a major threat: acute infection has a more variable and lower rate of spontaneous clearance (10% to 60%) [6]. This increases the likelihood of chronic infections, whose carriers are known to be at risk for developing cirrhosis and HCC.

There are eleven HCV genotypes, numbered from 1 to 11. The identification of HCC dual infections is motivated by the usage of genotypes for epidemiological studies and their impact on the efficacy of antiviral treatment [66].

5.3 Outlook on Genotyping HBV

We believe that the importance of genotyping HBV and other viruses will increase in the future. Recent developments in sequencing technology have reduced the costs of sequencing drastically, making sequencing more affordable. This could result in an increase in large-scale studies investigating the impact of HBV genotypes. These efforts might be able to identify previously unknown implications of genotypes, including those of HBV dual infections. These findings in conjunction with the reduced costs of sequencing might stimulate the usage of genotyping HBV in clinical praxis, thereby implementing the principles of personalized medicine.

Appendices

IUPAC Nucleotide Ambiguity Codes

IUPAC Code	Meaning
A	A
C	C
G	G
T	T
M	A or C
R	A or G
W	A or T
S	C or G
Y	C or T
K	G or T
V	A or C or G
H	A or C or T
D	A or G or T
B	C or G or T
N	G or A or T or C

Table 1: IUPAC nucleotide ambiguity codes [9].

Bibliography

[1] H.J. Alter, L.B. Seeff, P.M. Kaplan, V.J. McAuliffe, E.C. Wright, J.L. Gerin, R.H. Purcell, P.V. Holland, and H.J. Zimmerman. Type B hepatitis: the infectivity of blood positive for e antigen and DNA polymerase after accidental needlestick exposure. *New England Journal of Medicine*, 295(17):909–913, 1976.

[2] S.S. Bakken, A. Aulbach, E. Schmid, J. Winstead, L.T. Wilson, R. Lerdorf, and Z. Suraski. PHP manual. *URL: http://www. php. net/manual/en*, 15(05):03, 2003.

[3] A. Bartholomeusz and S. Schaefer. Hepatitis B virus genotypes: comparison of genotyping methods. *Reviews in Medical Virology*, 14(1):3–16, 2004.

[4] S. J. Bell and T. Nguyen. The management of hepatitis B. *Australian Prescriber*, 32(4):99–104, 2009.

[5] W.F. Carman, P. Karayiannis, J. Waters, HC Thomas, AR Zanetti, G. Manzillo, and A.J. Zuckerman. Vaccine-induced escape mutant of hepatitis B virus. *The Lancet*, 336(8711):325–329, 1990.

[6] F.A. Căruntu and L. Benea. Acute hepatitis C virus infection: Diagnosis, pathogenesis, treatment. *Journal of Gastrointestinal and Liver Diseases*, 15(3):249, 2006.

[7] C.J. Chu, M. Hussain, and A.S.F. Lok. Hepatitis B virus genotype B is associated with earlier HBeAg seroconversion compared with hepatitis B virus genotype C. *Gastroenterology*, 122(7):1756–1762, 2002.

[8] C.J. Chu, E.B. Keeffe, S.H. Han, R.P. Perrillo, A.D. Min, C. Soldevila-Pico, W. Carey, R.S. Brown, et al. Hepatitis B virus genotypes in the United States: results of a nationwide study. *Gastroenterology*, 125(2):444–451, 2003.

[9] A. Cornish-Bowden. IUPAC-IUB symbols for nucleotide nomenclature. *Nucleic Acids Res*, 13:3021–3030, 1985.

[10] A.M. Couroucé-Pauty, A. Plançon, and J.P. Soulier. Distribution of HBsAg subtypes in the world. *Vox sanguinis*, 44(4):197–211, 1983.

[11] R. de Franchis, G. Meucci, M. Vecchi, M. Tatarella, M. Colombo, E. Del Ninno, M.G. Rumi, M.F. Donato, and G. Ronchi. The natural history of asymptomatic hepatitis B surface antigen carriers. *Annals of Internal Medicine*, 118(3):191, 1993.

[12] X. Ding, H. Gu, Z.H. Zhong, X. Zilong, HT Tran, Y. Iwaki, T.C. Li, T. Sata, and K. Abe. Molecular epidemiology of hepatitis viruses and genotypic distribution of hepatitis B and C viruses in Harbin, China. *Jpn. J. Infect. Dis.*, 56(1):19–22, 2003.

[13] X. Ding, M. Mizokami, G. Yao, B. Xu, E. Orito, R. Ueda, and M. Nakanishi. Hepatitis B virus genotype distribution among chronic hepatitis B virus carriers in Shanghai, China. *Intervirology*, 44(1):43–47, 2000.

[14] A. Erhardt, U. Reineke, D. Blondin, W.H. Gerlich, O. Adams, T. Heintges, C. Niederau, and D. Häussinger. Mutations of the core promoter and response to interferon treatment in chronic replicative hepatitis B. *Hepatology*, 31(3):716–725, 2000.

[15] L. Frelin, T. Wahlstrom, A.E. Tucker, J. Jones, J. Hughes, B.O. Lee, J.N. Billaud, C. Peters, D. Whitacre, D. Peterson, et al. A mechanism to explain the selection of the hepatitis e antigen-negative mutant during chronic hepatitis B virus infection. *Journal of Virology*, 83(3):1379, 2009.

[16] S.K. Fung and A.S.F. Lok. Hepatitis B virus genotypes: do they play a role in the outcome of HBV infection? *Hepatology*, 40(4):790–792, 2004.

[17] R.S. Garfein, W.A. Bower, C.M. Loney, Y.J.F. Hutin, G.L. Xia, J. Jawanda, A.V. Groom, O.V. Nainan, J.S. Murphy, and B.P. Bell. Factors associated with fulminant liver failure during an outbreak among injection drug users with acute hepatitis B. *Hepatology*, 40(4):865–873, 2004.

[18] Y. Ghendon. Perinatal transmission of hepatitis B virus in high-incidence countries. *Journal of Virological Methods*, 17(1-2):69–79, 1987.

[19] A. Grakoui, H.L. Hanson, and C.M. Rice. Bad time for Bonzo? Experimental models of hepatitis C virus infection, replication, and pathogenesis. *Hepatology*, 33(3):489–495, 2001.

[20] C. Hannoun, K. Krogsgaard, P. Horal, and M. Lindh. Genotype mixtures of hepatitis B virus in patients treated with interferon. *Journal of Infectious Diseases*, 186(6):752, 2002.

[21] K. Hoff. The effect of sequencing errors on metagenomic gene prediction. *BMC Genomics*, 10(1):520, 2009.

[22] T.T.T. Huy, T.T. Ngoc, and K. Abe. New complex recombinant genotype of hepatitis B virus identified in Vietnam. *Journal of Virology*, 82(11):5657, 2008.

[23] K. Ikeda, S. Saitoh, Y. Suzuki, M. Kobayashi, A. Tsubota, M. Fukuda, I. Koida, Y. Arase, K. Chayama, N. Murashima, et al. Interferon decreases hepatocellular carcinogenesis in patients with cirrhosis caused by the hepatitis B virus. *Cancer*, 82(5):827–835, 1998.

[24] T. Imamura, O. Yokosuka, T. Kurihara, T. Kanda, K. Fukai, F. Imazeki, and H. Saisho. Distribution of hepatitis B viral genotypes and mutations in the core promoter and precore regions in acute forms of liver disease in patients from Chiba, Japan. *Gut*, 52(11):1630, 2003.

[25] J.H. Kao, P.J. Chen, M.Y. Lai, and D.S. Chen. Hepatitis B genotypes correlate with clinical outcomes in patients with chronic hepatitis B. *Gastroenterology*, 118(3):554–559, 2000.

[26] J.H. Kao, P.J. Chen, M.Y. Lai, and D.S. Chen. Acute exacerbations of chronic hepatitis B are rarely associated with superinfection of hepatitis B virus. *Hepatology*, 34(4):817–823, 2001.

[27] J.H. Kao, N.H. Wu, P.J. Chen, M.Y. Lai, and D.S. Chen. Hepatitis B genotypes and the response to interferon therapy. *Journal of Hepatology*, 33(6):998–1002, 2000.

[28] H. Kato, E. Orito, F. Sugauchi, R. Ueda, T. Koshizaka, S. Yanaka, R.G. Gish, F. Kurbanov, R. Ruzibakiev, A. Kramvis, et al. Frequent coinfection with hepatitis B virus strains of distinct genotypes detected by hybridization with type-specific probes immobilized on a solid-phase support. *Journal of Virological Methods*, 110(1):29–35, 2003.

[29] C.S. Keith, D.O. Hoang, B.M. Barrett, B. Feigelman, M.C. Nelson, H. Thai, and C. Baysdorfer. Partial sequence analysis of 130 randomly selected maize cDNA clones. *Plant Physiology*, 101(1):329, 1993.

[30] A. Kramvis, M. Kew, and G. Francois. Hepatitis B virus genotypes. *Vaccine*, 23(19):2409–2423, 2005.

[31] F. Kurbanov, Y. Tanaka, A. Kramvis, P. Simmonds, and M. Mizokami. When should "I" consider a new hepatitis B virus genotype? *Journal of Virology*, 82(16):8241, 2008.

[32] C.L. Lai and M.F. Yuen. The natural history and treatment of chronic hepatitis B: a critical evaluation of standard treatment criteria and end points. *Annals of Internal Medicine*, 147(1):58, 2007.

[33] M.K. Libbus and L.M. Phillips. Public health management of perinatal hepatitis B virus. *Public Health Nursing*, 26(4):353–361, 2009.

[34] C.L. Lin and J.H. Kao. The clinical implications of hepatitis B virus genotype: Recent advances. *Journal of Gastroenterology and Hepatology*, 26:123–130, 2011.

[35] M. Lindh, C. Hannoun, A.P. Dhillon, G. Norkrans, and P. Horal. Core promoter mutations and genotypes in relation to viral replication and liver damage in East Asian hepatitis B virus carriers. *Journal of Infectious Diseases*, 179(4):775, 1999.

[36] A.S.F. Lok and B.J. McMahon. Chronic hepatitis B. *Hepatology*, 45(2):507–539, 2007.

[37] X. Lu and T. Block. Study of the early steps of the Hepatitis B Virus life cycle. *International Journal of Medical Sciences*, 1(1):21, 2004.

[38] M.A. Mallory, S.R. Page, and D.R. Hillyard. Development and validation of a hepatitis B virus DNA sequencing assay for assessment of antiviral resistance, viral genotype and surface antigen mutation status. *Journal of Virological Methods*, 2011.

[39] P. Marcellin, G.K.K. Lau, F. Bonino, P. Farci, S. Hadziyannis, R. Jin, Z.M. Lu, T. Piratvisuth, G. Germanidis, C. Yurdaydin, et al. Peginterferon alfa-2a alone, lamivudine alone, and the two in combination in patients with HBeAg-negative chronic hepatitis B. *New England Journal of Medicine*, 351(12):1206–1217, 2004.

[40] C. Mayerat, A. Mantegani, and PC Frei. Does hepatitis B virus (HBV) genotype influence the clinical outcome of HBV infection? *Journal of Viral Hepatitis*, 6(4):299–304, 1999.

[41] G.P. McCormack and J.P. Clewley. The application of molecular phylogenetics to the analysis of viral genome diversity and evolution. *Reviews in Medical Virology*, 12(4):221–238, 2002.

[42] K. Michitaka, N. Horiike, Y. Chen, TN Duong, K. Matsuura, Y. Tokumoto, Y. Hiasa, FSM Akbar, and M. Onji. Co-infection with hepatitis B virus genotype D and other genotypes in western Japan. *Intervirology*, 48(4):262–267, 2005.

[43] R. Myers, C. Clark, A. Khan, P. Kellam, and R. Tedder. Genotyping Hepatitis B virus from whole- and sub-genomic fragments using position-specific scoring matrices in HBV STAR. *Journal Of General Virology*, 87(6):1459, 2006.

[44] R.E. Myers, C.V. Gale, A. Harrison, Y. Takeuchi, and P. Kellam. A statistical model for HIV-1 sequence classification using the subtype analyser (STAR). *Bioinformatics*, 21(17):3535, 2005.

[45] H. Naito, S. Hayashi, and K. Abe. Rapid and specific genotyping system for hepatitis B virus corresponding to six major genotypes by PCR using type-specific primers. *Journal of Clinical Microbiology*, 39(1):362, 2001.

[46] M. Nassal and H. Schaller. Hepatitis B virus replication. *Trends in Microbiology*, 1(6):221–228, 1993.

[47] C. Niederau, T. Heintges, S. Lange, G. Goldmann, C.M. Niederau, L. Mohr, and D. H "aussinger. Long-term follow-up of HBeAg-positive patients treated with interferon alfa for chronic hepatitis B. *New England Journal of Medicine*, 334(22):1422–1427, 1996.

[48] H. Noguchi, J. Park, and T. Takagi. Metagene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Research*, 34(19):5623, 2006.

[49] H. Norder, A.M. Couroucé, and L.O. Magnius. Complete genomes, phylogenetic relatedness, and structural proteins of six strains of the hepatitis B virus, four of which represent two new genotypes. *Virology*, 198(2):489–503, 1994.

[50] H. Norder, B. Hammas, S.D. Lee, K. Bile, A.M. Couroucé, I.K. Mushahwar, and L.O. Magnius. Genetic relatedness of hepatitis B viral strains of diverse geographical origin and natural variations in the primary structure of the surface antigen. *Journal of General Virology*, 74:1341, 1993.

[51] H. Okamoto, F. Tsuda, H. Sakugawa, R.I. Sastrosoewignjo, M. Imai, Y. Miyakawa, and M. Mayumi. Typing hepatitis B virus by homology in nucleotide sequence: comparison of surface antigen subtypes. *Journal of General Virology*, 69:2575, 1988.

[52] C.M. Olinger, P. Jutavijittum, J.M. Hübschen, A. Yousukh, B. Samountry, T. Thammavong, K. Toriyama, and C.P. Muller. Possible New Hepatitis B Virus Genotype, Southeast Asia. *Emerging Infectious Diseases*, 14(11):1777, 2008.

[53] E. Orito, M. Mizokami, H. Sakugawa, K. Michitaka, K. Ishikawa, T. Ichida, T. Okanoue, H. Yotsuyanagi, and S. Iino. A case-control study for clinical and molecular biological differences between hepatitis B viruses of genotypes B and C. *Hepatology*, 33(1):218–223, 2001.

[54] C. Osiowy and E. Giles. Evaluation of the INNO-LiPA HBV genotyping assay for determination of hepatitis B virus genotype. *Journal of Clinical Microbiology*, 41(12):5473, 2003.

[55] N. Panjaworayan, S.K. Roessner, A.E. Firth, and C.M. Brown. HBVRegDB: annotation, comparison, detection and visualization of regulatory elements in hepatitis B virus sequences. *Virology journal*, 4(1):136, 2007.

[56] G.V. Papatheodoridis, E. Manesis, and S.J. Hadziyannis. The long-term outcome of interferon-alpha treated and untreated patients with HBeAg-negative chronic hepatitis B. *Journal of Hepatology*, 34(2):306–313, 2001.

[57] M.R. Pourkarim, S. Amini-Bavil-Olyaee, P. Lemey, P. Maes, and M. Van Ranst. Are hepatitis B virus "subgenotypes" defined accurately? *Journal of Clinical Virology*, 47(4):356–360, 2010.

[58] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.

[59] M. Rozanov, U. Plikat, C. Chappey, A. Kochergin, and T. Tatusova. A web-based genotyping resource for viral sequences. *Nucleic Acids Research*, 32(suppl 2):W654, 2004.

[60] J.M. Sánchez-Tapias, J. Costa, A. Mas, M. Bruguera, and J. Rodés. Influence of hepatitis B virus genotype on the long-term outcome of chronic hepatitis B in western patients. *Gastroenterology*, 123(6):1848–1856, 2002.

[61] F. Sanger, S. Nicklen, and A.R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463, 1977.

[62] S. Schaefer. Hepatitis B virus genotypes in Europe. *Hepatology Research*, 37:S20–S26, 2007.

[63] T. Shaw, A. Bartholomeusz, and S. Locarnini. HBV drug resistance: mechanisms, detection and interpretation. *Journal of Hepatology*, 44(3):593–606, 2006.

[64] J. Sheldon and V. Soriano. Hepatitis B virus escape mutants induced by antiviral therapy. *Journal of Antimicrobial Chemotherapy*, 61(4):766, 2008.

[65] S. Shiina, H. Fujino, Y. Uta, K. Tagawa, T. Unuma, M. Yoneyama, T. Ohmori, S. Suzuki, M. Kurita, and Y. Ohashi. Relationship of HBsAg subtypes with HBeAg/anti-HBe status and chronic liver disease. Part I: Analysis of 1744 HBsAg carriers. *The American Journal of Gastroenterology*, 86(7):866, 1991.

[66] P. Simmonds, EC Holmes, TA Cha, SW Chan, F. McOmish, B. Irvine, E. Beall, PL Yap, J. Kolberg, and MS Urdea. Classification of hepatitis C virus into six major genotypes and a series of subtypes by phylogenetic analysis of the NS-5 region. *Journal of General Virology*, 74:2391–2391, 1993.

[67] H. Sumi, O. Yokosuka, N. Seki, M. Arai, F. Imazeki, T. Kurihara, T. Kanda, K. Fukai, M. Kato, and H. Saisho. Influence of hepatitis B virus genotypes on the progression of chronic type B liver disease. *Hepatology*, 37(1):19–26, 2003.

[68] K. Tatematsu, Y. Tanaka, F. Kurbanov, F. Sugauchi, S. Mano, T. Maeshiro, T. Nakayoshi, M. Wakuta, Y. Miyakawa, and M. Mizokami. A genetic variant of hepatitis B virus divergent from known human and ape genotypes isolated from a Japanese patient and provisionally assigned to new genotype J. *Journal of Virology*, 83(20):10538, 2009.

[69] S. Tong, K.H. Kim, C. Chante, J. Wands, and J. Li. Hepatitis B virus e antigen variants. *International Journal of Medical Sciences*, 2(1):2, 2005.

[70] S. Usuda, H. Okamoto, H. Iwanari, K. Baba, F. Tsuda, Y. Miyakawa, and M. Mayumi. Serological detection of hepatitis B virus genotypes by ELISA with monoclonal antibodies to

type-specific epitopes in the preS2-region product. *Journal of Virological Methods*, 80(1):97–112, 1999.

[71] G. Van Rossum and Centrum voor Wiskunde en Informatica. *Python reference manual*. Centrum voor Wiskunde en Informatica, 1995.

[72] C.T. Wai, C.J. Chu, M. Hussain, and A.S.F. Lok. HBV genotype B is associated with better response to interferon therapy in HBeAg (+) chronic hepatitis than genotype C. *Hepatology*, 36(6):1425–1430, 2002.

[73] World Health Organization (WHO). Hepatitis B (fact sheet 204), August 2008. Available at: http://www.who.int/mediacentre/factsheets/fs204/en/index.html.

[74] D.K.H. Wong, A.M. Cheung, K. O'Rourke, C.D. Naylor, A.S. Detsky, and J. Heathcote. Effect of alpha-interferon treatment in patients with hepatitis B e antigen-positive chronic hepatitis B. *Annals of Internal Medicine*, 119(4):312, 1993.

[75] World Health Organization, Regional Office for South-East Asia. Prevention of Hepatitis B in India, 2002.

[76] B. Zöllner, J. Petersen, E. Puchhammer-Stöckl, J. Kletzmayr, M. Sterneck, L. Fischer, M. Schröter, R. Laufs, and H.H. Feucht. Viral features of lamivudine resistant hepatitis B genotypes A and D. *Hepatology*, 39(1):42–50, 2004.

[77] F. Zoulim and S. Locarnini. Hepatitis B virus resistance to nucleos(t)ide analogues. *Gastroenterology*, 137(5):1593–1608, 2009.