BIG DATA MINING AND ANALYTICS
ISSN 2096-0654 01/05 pp145–158
Volume 2, Number 3, September 2019
DOI: 10.26599/BDMA.2019.9020005

Bayesian Analysis of Complex Mutations in HBV, HCV, and HIV Studies

Bing Liu, Shishi Feng, Xuan Guo, and Jing Zhang*

Abstract: In this article, we aim to provide a thorough review of the Bayesian-inference-based methods applied to Hepatitis B Virus (HBV), Hepatitis C Virus (HCV), and Human Immunodeficiency Virus (HIV) studies with a focus on the detection of the viral mutations and various problems which are correlated to these mutations. It is particularly difficult to detect and interpret these interacting mutation patterns, but by using Bayesian statistical modeling, it provides a groundbreaking opportunity to solve these problems. Here we summarize Bayesian-based statistical approaches, including the Bayesian Variable Partition (BVP) model, Bayesian Network (BN), and the Recursive Model Selection (RMS) procedure, which are designed to detect the mutations and to make further inferences to the comprehensive dependence structure among the interactions. BVP, BN, and RMS in which Markov Chain Monte Carlo (MCMC) methods are used have been widely applied in HBV, HCV, and HIV studies in the recent years. We also provide a summary of the Bayesian methods' applications toward these viruses' studies, where several important and useful results have been discovered. We envisage the applications of more modified Bayesian methods to other infectious diseases and cancer cells that will be following with critical medical results before long.

Key words: Bayesian analysis; Hepatitis B Virus (HBV); Hepatitis C Virus (HCV); Human Immunodeficiency Virus (HIV); complex mutations; Markov chain Monte Carlo

1 Introduction

Per historical data, there are up to 30 million people across the world who are infected with Hepatitis B Virus (HBV) and up to 600 thousand die every year[1,2]. According to World Health Organization (WHO), among infected adults, "less than 5% of otherwise healthy persons who are infected as adults will develop chronic infection, and 20% – 30% of adults who are chronically infected will develop cirrhosis and/or liver cancer"; and rate is higher in younger populations: "80% – 90% of infants infected during the first year of life develop chronic infections, and 30% – 50% of children infected before the age of 6 years develop chronic infections."[2] There are about 1/3 of chronic infected subjects will have irreversible outcome as liver damage, and it leads to cirrhosis and hepatocellular carcinoma; and the other 2/3 infected subjects will retain the virus in their body and become highly infectious though asymptomatic[3]. In total, up to 25% of subjects with chronic infected HBV die from the complications due to the disease[3].

HBV is a member of the Hepadnaviridae family, and it comprises an icosahedral protein capsid surrounding the viral DNA, with a lipoprotein viral envelope[4–7]. The virus DNA is organized in 4 Open Reading Frames (ORF): S, which stands for surface and encodes HBsAg; C, which stands for core and encodes HBcAg and HBeAg; P, which stands for polymerase and

• Bing Liu, Shishi Feng, and Jing Zhang are with the Department of Mathematics and Statistics, Georgia State University, Atlanta, GA 30303, USA. E-mail: bing-liu@outlook.com; fengss0105@icloud.com; jzhang47@gsu.edu.
• Xuan Guo is with the Department of Computer Science and Engineering, University of North Texas, Denton, TX 76203, USA. E-mail: xuan.guo@unt.edu.
∗ To whom correspondence should be addressed.
Manuscript received: 2018-09-28; revised: 2019-02-10; accepted: 2019-02-16

encodes DNA polymerase; and X, which encodes an X protein, and currently we are not clear of its exact function[5, 7]. Two highly immunogenic proteins, HBcAg and HBeAg, consist of the nucleocapsid, and a less immunogenic surface antigen HBsAg is in the viral envelop[5–7]. In patients with chronic infection, serum HBV-DNA reflects the disease progression and the transition across the different stages of the disease[8]. "Identifying HBsAg mutations correlated with different levels of serum HBV-DNA in HBV chronically infected patients naive to anti-HBV drugs"[8] is one of the interests of HBV studies. In the meantime, "Occult HBV Infection (OBI) is a threat for the safety of blood-supply, and has been associated with the onset of HBV-related hepatocellular carcinoma and lymphomagenesis."[9] "The genetic markers in HBsAg (particularly in D-genotype, the most common in Europe) significantly associated with OBI in vivo are missing", so the correlation between HBaAg-mutations and OBI and its impact on HBsAg detection are also important[9]. The above problems can be solved by using Bayesian framework.

Hepatitis C Virus (HCV) is a single-strand RNA virus and has been classified into at least six genotypes with several subtypes in each. The response patterns of different genotypes to interferon-based therapy are diverse with them spreading in different regions[10]. In previous clinical experience, Interferon (IFN) and ribavirin combined therapy has a significantly higher rate of sustained response in chronic HCV patients compared with interferon-based therapy which has only less than 20% sustained response[11, 12].

Some variations in the HCV sequences have the ability of interfering the effective functioning of IFN-based therapies. Among all these variations, the ones in the NS5A region[13, 14] are the main subject in our review. NS5A is a nonstructural protein that can lead to IFN therapy resistance by impacting the function of an important mediator of IFN response called dsRNA dependent Protein Kinase (PKR)[15, 16]. NS5A region has 1344 base pairs linking to 448 amino acid and constitutes several regions: "the membrane attachment region (aa 1 – 236), the carboxyl region (aa 237 – 448), and the regions within the carboxyl end, such as PKRbd (aa 237 – 302), Variable region 4 (V4; aa 310 – 330), Variable region 3 (V3; aa 381 – 409), the region between V3 and V4 (aa 331 – 380), and the downstream region of V3 (aa 410 – 448)."[17]

In general, mutations in NS5A region have been proposed to be related to therapy resistance by Enomoto and Sato[18] and other researchers[19, 20]. However, the relation between mutations in NS5A region and IFN resistance remains ambiguous because of contradictory results obtained in studies concerning PKR binding domain in NS5A[21]. Thus, a better and deeper comprehension of the role of NS5A region in antiviral resistance to IFN therapy will contribute greatly to the development of treatment strategies against HCV.

Human Immunodeficiency Virus (HIV) is an enveloped virus with a single-stranded RNA genome and is the cause of the Acquired Immunodeficiency Syndrome (AIDS) which killed more than 20 million people since 1980s (www.who.int/hiv/en/)[22, 23]. The replication cycle of HIV-1 virus consists of 13 important steps[24], beginning with the attachment step and ending with the protease-mediated mutation process. The attachment step marks the entry of virus into host cell by the fusion of membranes of the cell and virus[25, 26]. A trimer of gp120 and gp41 heterodimers forms the only protein envelope on the viral surface. The HIV-1's delivery of genome into the host cell is an extremely intricate process in which a collaborative interaction of the envelope glycoprotein gp120 with the CD4 receptor and with chemokine receptors is required. The chemokine receptors mainly refer to CC chemokine Receptor type 5 (CCR5) and C-X-C chemokine Receptor type 4 (CXCR4)[27].

These receptors can be used to classify HIV-1 virus since the ability of virus to use the CCR5 and CXCR4 co-receptor differs from each other. It has been proposed by previous studies that R5-reopic viruses which can only use the CCR5 co-receptor are the predominant in majority of newly HIV-1 infected patients and are generally responsible for the initial infection. Meanwhile, CXCR4 co-receptor usage is observed more often in advanced stages of disease[27, 28]. And among the domains of HIV-1 gp120, the V3 loop is the primary determinant for HIV-1 co-receptor usage[29]. Thus, in order to provide more valuable information for the development of anti-HIV-1 drugs targeting on inhibiting the entry of CCR5-tropic HIV-1 strains into host cell, we keep our focus on defining the V3 genetic determinants and the structural features underlying the ability of HIV-1 to

use the CCR5 and CXCR4 co-receptors[30]. Moreover, understanding the detailed interaction mutation patterns related to drug-resistance in V3 is also of great importance to develop effective treatment against HIV[24].

Genomic data can be studied using various methods including Bayesian methods and data mining methods. Some encoding schemes may also be very useful in genomic representation and feature learning[31]. Zagordi et al.[32] developed a Bayesian approach to detect minority variants for estimating HIV quasispecies[33]. Zhang et al.[34] proposed an innovative Bayesian method for investigating mutation interactions of HIV after certain drug treatment. This method has been used in detecting genome-wide associations on HBV and HCV as well. The inference of Bayesian Network (BN) has been applied to model the drug resistance of HBV and HIV[35,36]. Thai et al.[35] confirmed that lamivudine resistance is a complex trait encoded by the entire HBV genome using a set of Bayesian networks of polymorphic amino acid sites of pre- and post-treatment from HBV patients. Beerenwinkel et al.[36] used isotonic conjunctive Bayesian networks, a class of BN, to model the evolutionary escape dynamics of HIV-1. Recently, Chaillon et al.[37] used Bayesian-based statistical modeling to access the likelihood of sexual transmission and persistence of drug resistance mutations in HIV infection. In this article, we will provide a thorough review of the Bayesian Variable Partition (BVP) model, BN, the Recursive Model Selection (RMS), and their real applications in HBV, HCV, and HIV studies.

2 Introduction of Bayesian Inference

Bayesian inference is a technique of statistical inference which is specifically based on the use of Bayesian theorem. It has been widely applied to update the probability estimate for a hypothesis as evidence or information becomes available, and it's the formal methods for combining prior beliefs with observed information to answer the questions that researchers are usually interested in. It is not complicated to build a model by using the combination of multiple experiments' information. This natural way can also fit realistic. With all the benefits, Bayesian inference however often comes with a high computational cost and it needs to express subjective prior beliefs into a mathematical prior formula or function.

Bayes' rule (http://en.wikipedia.org/wiki/Bayesian_inference). When θ is a discrete random variable with a probability mass function, the Bayes' rule is

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{\sum\limits_i p(x|\theta_i)p(\theta_i)}.$$

When θ is continuous with a probability density function, the Bayes' rule becomes

$$p(\theta|x) = \frac{p(\theta,x)}{p(x)} = \frac{p(x|\theta)p(\theta)}{p(x)} = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)\mathrm{d}\theta}.$$

Bayes' rule is often written as $p(\theta|x) \propto p(\theta)p(x|\theta)$, when treated as a function of θ for a fixed x, where $p(x|\theta)$ is the likelihood $L(x|\theta)$. Bayes' rule can be considered as

$$\text{Posterior} \propto \text{Prior} \times \text{Likelihood},$$
$$p(\theta|x) \propto p(\theta)p(x|\theta).$$

This is expressed in words as "the posterior is proportional to the product of the prior and the likelihood."

Basics of the Bayesian inference. When we use Bayesian statistics to make inferences, consider

(1) Setting up a probability model;

(2) Applying the probability theory and the Bayes' rule.

For example, let $x_1, x_2, \ldots, x_n$ be an independent sample from Binomial distribution $\mathrm{Bin}(n,\pi)$, where n is the sample size and π is the probability of success. We have $x|\pi \sim \mathrm{Bin}(n,\pi)$. The likelihood can be written as

$$p(x|\pi) = \left(\begin{array}{c} n \\ x \end{array}\right) \pi^x (1-\pi)^{n-x},\ \pi \in [0,1].$$

If we want to make inference on π given x and n, a prior distribution $p(\pi)$ for π is needed. We can use a uniform distribution $\pi \sim U(0,1)$:

$$p(\pi) = \begin{cases} 1, & 0 \leqslant \pi \leqslant 1; \\ 0, & \text{otherwise.} \end{cases}$$

Then by applying the Bayes' rule, we get

$$p(x,\pi) = \left(\begin{array}{c} n \\ x \end{array}\right) \pi^x (1-\pi)^{n-x},$$

$$p(x) = \int_0^1 \left(\begin{array}{c} n \\ x \end{array}\right) \pi^x (1-\pi)^{n-x}\mathrm{d}\pi = \frac{1}{n+1},$$

$$p(\pi|x) = \frac{p(x,\pi)}{p(x)} = (n+1)\left(\begin{array}{c} n \\ x \end{array}\right) \pi^x (1-\pi)^{n-x}.$$

3 Bayesian Methods in HBV, HCV, and HIV Studies

In this section, we will first summarize and generalize the Bayesian statistical models applied to HBV, HCV, and HIV studies in terms of finding the virus sequence mutations and the difference in two (or three) different groups of patients. Then a summary of important and interesting results found by applying these methods will be carried out for HBV, HCV, and HIV studies.

3.1 Bayesian variable partition model

Zhang et al.[34] first developed the BVP model to detect and understand combinatorial mutation patterns responsible for HIV drug resistance. Up to now, this method has been successfully applied in various virus studies[38–41].

Following the notations in Ref. [42], generally, suppose we have two data sets in the form of matrices, say $A = [A_1, \ldots, A_m]$ (of dimension $n_A \times m$) and $B = [B_1, \ldots, B_m]$ (of dimension $n_B \times m$), respectively (each row is a sequence and each column is a position of amino acid sequence). The numbers of sequences in two groups are denoted using n_A and n_B, and m denotes the number of positions. On top of that, we establish the following four assumptions for the distribution of the positions from the two groups:

H1: The identity of the **independent** positions, where group A and group B data share the **same** probability distribution.

H2: The identity of the **independent** positions, where group A and group B data have **different** probability distributions.

H3: The identity of the **dependent** positions, where group A and group B data share the **same** probability distribution.

H4: The identity of the **dependent** positions, where group A and group B data have **different** probability distributions.

From these hypotheses, we are interested in positions from H2 and H4 particularly. Therefore, we will start with the positions from H2. Given that the position i is from H2, and we assume there are c_i possible values (amino acids) at position i, and for every sequence in group A, we have p_1 for the first value, p_2 for the second, $\ldots$, p_{c_i} for the last value, and $\sum_{j=1}^{c_i} p_j = 1$. Then we can calculate the likelihood for data set A at position i is

$$P(A_i|p_1, p_2, \ldots, p_{c_i}, \mathrm{H2}) = \prod_{j=1}^{c_i} p_j^{n_j},$$

where n_j denotes the number of sequence with the j-th value in A_i. At the same time, we have p'_j for the j-th value in group B, and $\sum_{j=1}^{c_i} p'_j = 1$. So the likelihood for group B at position i is

$$P(B_i|p'_1, p'_2, \ldots, p'_{c_i}, \mathrm{H2}) = \prod_{j=1}^{c_i} (p'_j)^{n'_j},$$

where n'_j is the number of sequence with the j-th value in B_i.

Under the assumption of H2, $p_j \neq p'_j$, since we do not know the true values of p_j or p'_j, we assume they are random and a Dirichlet prior is applied on them.

$$p \sim \mathrm{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_{c_i}):$$
$$P(p_1, p_2, \ldots, p_{c_i}|\mathrm{H2}, \alpha_1, \alpha_2, \ldots, \alpha_{c_i}) = \frac{1}{B(\alpha)} \prod_{j=1}^{c_i} p_j^{\alpha_j - 1},$$

where $B(\alpha) = \dfrac{\prod_{j=1}^{c_j} \Gamma(\alpha_j)}{\Gamma\left(\sum_{j=1}^{c_j} \alpha_j\right)}$, $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_{c_i})$, and $\Gamma(x) = \displaystyle\int_0^\infty t^{x-1} \mathrm{e}^{-t} \mathrm{d}t$.

$$p' \sim \mathrm{Dirichlet}(\alpha'_1, \alpha'_2, \ldots, \alpha'_{c_i}):$$
$$P(p'_1, p'_2, \ldots, p'_j|\mathrm{H2}, \alpha'_1, \alpha'_2, \ldots, \alpha'_{c_i}) = \frac{1}{B(\alpha')} \prod_{j=1}^{c_i} (p'_j)^{\alpha'_j - 1},$$

where $B(\alpha') = \dfrac{\prod_{j=1}^{c_j} \Gamma(\alpha'_j)}{\Gamma(\sum_{j=1}^{c_j} \alpha'_j)}$, $\alpha' = (\alpha'_1, \alpha'_2, \ldots, \alpha'_{c_i})$, and $\Gamma(x) = \displaystyle\int_0^\infty t^{x-1} \mathrm{e}^{-t} \mathrm{d}t$.

Then we have

$$P(A_i, p_1, p_2, \ldots, p_{c_i}|\mathrm{H2}) = \prod_{j=1}^{c_i} p_j^{n_j} \times \mathrm{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_{c_i}) = \frac{1}{B(\alpha)} \prod_{j=1}^{c_i} p_j^{n_j + \alpha_j - 1},$$

$$P(B_i, p'_1, p'_2, \ldots, p'_{c_i}|\mathrm{H2}) = \prod_{j=1}^{c_i} {p'_j}^{n'_j} \times \mathrm{Dirichlet}(\alpha'_1, \alpha'_2, \ldots, \alpha'_{c_i}) = \frac{1}{B(\alpha')} \prod_{j=1}^{c_i} {p'_j}^{n'_j + \alpha'_j - 1}.$$

By integrating out p and p', respectively, we get

$$P(A_i|\mathrm{H2}) = \int P(A_i, p_1, p_2, \ldots, p_{c_i}|\mathrm{H2})\mathrm{d}p = \prod_{j=1}^{c_j} \frac{\Gamma(n_j+\alpha_j)}{\Gamma(\alpha_j)} \frac{\Gamma(\sum_{j=1}^{c_j}\alpha_j)}{\Gamma(\sum_{j=1}^{c_j}(n_j+\alpha_j))},$$

$$P(B_i|\mathrm{H2}) = \int P(B_i, p'_1, p'_2, \ldots, p'_{c_i}|\mathrm{H2})\mathrm{d}p' = \prod_{j=1}^{c_j} \frac{\Gamma(n'_j+\alpha'_j)}{\Gamma(\alpha'_j)} \frac{\Gamma(\sum_{j=1}^{c_j}\alpha'_j)}{\Gamma(\sum_{j=1}^{c_j}(n_j+\alpha'_j))}.$$

And then

$$P(A_i, B_i|\mathrm{H2}) = P(A_i|\mathrm{H2})P(B_i|\mathrm{H2}).$$

Now under H1, we have $p_j = p'_j$, so we can obtain

$$P(A_i, B_i|\mathrm{H1}) = \int P(A_i, B_i, p_1, p_2, \ldots, p_{c_i}|\mathrm{H1})\mathrm{d}p = \int \frac{1}{B(\alpha)} \prod_{j=1}^{c_i} p_j^{n_j+n'_j+\alpha_j-1}\mathrm{d}p = \prod_{j=1}^{c_i} \frac{\Gamma(n_j+n'_j+\alpha_j)}{\Gamma(\alpha_j)} \frac{\Gamma(\sum_{j=1}^{c_j}\alpha_j)}{\Gamma\left(\sum_{j=1}^{c_j}(n_j+n'_j+\alpha_j)\right)}.$$

For hypothesis H4, we assume there are c possible value combinations of the dependent positions. Likewise, suppose for every sequence in group A, we have p_1 for the first combination, p_2 for the second combination, ..., p_c for the last combination, and $\sum_{j=1}^{c} p_j = 1$; for every sequence in group B, we have p'_1 for the first combination, p'_2 for the second combination, ..., p'_c for the last combination, and $\sum_{j=1}^{c} p'_j = 1$. Then, we have

$$P(\text{dependent positions in } A|\mathrm{H4}) = \prod_{j=1}^{c} \frac{\Gamma(n_j+\alpha_j)}{\Gamma(\alpha_j)} \frac{\Gamma(\sum_{j=1}^{c}\alpha_j)}{\Gamma(\sum_{j=1}^{c}(n_j+\alpha_j))},$$

$$P(\text{dependent positions in } B|\mathrm{H4}) = \prod_{j=1}^{c} \frac{\Gamma(n'_j+\alpha'_j)}{\Gamma(\alpha'_j)} \frac{\Gamma(\sum_{j=1}^{c}\alpha'_j)}{\Gamma(\sum_{j=1}^{c}(n'_j+\alpha'_j))},$$

where n_j and n'_j are the numbers of the j-th combination in A and B, respectively, and then

$$P(\text{dependent positions in } A, B|\mathrm{H4}) = P(\text{dependent positions in } A)\times P(\text{dependent positions in } B).$$

Now under H3, we have $p_j = p'_j$, so similarly we have

$$P(\text{dependent positions in } A, B|\mathrm{H3}) = \prod_{j=1}^{c} \frac{\Gamma(n_j+n'_j+\alpha_j)}{\Gamma(\alpha_j)} \frac{\Gamma(\sum_{j=1}^{c}\alpha_j)}{\Gamma\left(\sum_{j=1}^{c}(n_j+n'_j+\alpha_j)\right)}.$$

We define an indicator vector $I = [I_1, I_2, \ldots, I_m]$ to indicate the hypothesis group of m different positions belong to, where $I_i = 1$ means position i is from H1, $I_i = 2$ means position i is from H2, $I_i = 3$ means position i is from H3, and at last $I_i = 4$ means that the position i is from H4.

Then, as we are interested in the inference of I, so we want to find the posterior distribution of I, given the data sets A and B, i.e., $P(I|A, B)$. Applying the Bayes' theorem, we obtain

$$P(I|A, B) = \frac{P(I)P(A, B|I)}{\sum_{\text{all possible } I} P(I)P(A, B|I)}.$$

Therefore,

$$P(I|A, B) \propto P(I)P(A, B|I).$$

Based on H1, H2, H3, and H4, we have

$$P(A, B|I) = \prod_{I_i=1,2} P(A_i, B_i|I_i)\times P(\text{dependent positions from H3})\times P(\text{dependent positions from H4}).$$

In practice, we also need to assume the prior for I. For example, we may assume most positions should be in H1 and H3, then we set $P(I_i = 2) = P(I_i = 4) = 0.01$, and $P(I) = \prod_{i=1}^{m} P(I_i)$.

3.2 Dirichlet process mixture

Zagordi et al.[32] developed a probabilistic Bayesian approach to minimize the effect of errors on the detection of minority variants when estimating HIV quasispecies[33]. This approach assumes that sequencing reads tend to cluster around the true haplotypes[43], with a distribution depending on the error process, while these haplotypes can be separated by their true evolutionary distance. Although general-purpose clustering algorithms can be used to do the read clustering, they face the problem of choosing the right number of clusters. To overcome this issue, Zagordi et al.[32] used a Bayesian fashion with a Dirichlet Process Mixture (DPM)[44], which defines a prior distribution on the unknown number of haplotypes. DPM is capable to capture the uncertainty in the number of clusters and the phylogenetic structure of unknown haplotypes. The prior on mixing proportions then leads

to a few dominating classes and is controlled by a hyperparameter α. This prior is expressed in the following equation:

$$p(c_i = c|c_j : j \neq i) = \begin{cases} \dfrac{n_c}{n-1+\alpha}, & \text{read } i \text{ is assigned to class } c; \\ \dfrac{\alpha}{n-1+\alpha}, & \text{read } i \text{ is assigned to a new class,} \end{cases}$$

where n_c is the size of class c, and n is the total number of reads.

In addition to the prior on the assignment of observations, the generation of reads from different haplotypes is modeled by

$$p(R|C,H) = \prod_{k=1}^{K} \omega^{m_k} \left(\frac{1-\omega}{|O|-1}\right)^{m'_k},$$

where $R = \{r_1, \ldots, r_n\}$ denotes a set of the given reads, $C = \{c_1, \ldots, c_n\}$ is the assignments of the given reads, $H = \{h_1, \ldots, h_K\}$ is the set of haplotypes determined eventually by the data, ω is the probability that a base is drawn without error, O is the alphabet of the bases, and m_k and m'_k denote the number of matched and unmatched bases between reads and assigned haplotypes, respectively. When the read i is assigned to a new class, the generation process is formulated the following equation:

$$p(r_i|h_0) = \left[\omega \cdot \gamma + (1-\gamma)\frac{1-\omega}{|O|-1}\right]^{m_{i,0}} \cdot \left[\frac{1}{(|O|-1)^2}(\omega + \gamma + |O|(1-\omega\cdot\gamma) - 2)\right]^{m'_{i,0}},$$

where h_0 is the known reference genome, γ is the mutation probability of a base, and $m_{i,0}$ and $m'_{i,0}$ denote the number of matched and unmatched basesetween the reads i and h_0. The intuition of this equation is that the reads generated by the reference genome are affected by both sequencing error and mutation. Based on the above three models, a Markov chain Monte Carlo algorithm performing Gibbs sampling was used to sample the joint posterior distribution haplotype sequences, assignment of reads to haplotypes, and error rate of the sequencing process to obtain estimates of the local haplotype structure of the population. More details about this Gibbs sampling can be found in Ref. [32].

3.3 Bayesian partition on dual usage of co-receptor model

To detect and understand genetic and structural features in HIV-1 B subtype V3 underlying HIV-1 co-receptor usage, Chen et al.[30] developed a Bayesian Partition on Dual Usage of Co-receptor Model (BPDUCM) to define V3 genetic determinants either independently or interactively associated with the usage CCR5 co-receptor only, CXCR4 co-receptor only, or dual of CCR5/CXCR4 co-receptor.

This method was applied to analyze three datasets — CCR5 only, CXCR4 only, and dual usage. To clearly show the method, the notations from Chen et al.[30] are directly employed here. Suppose there are N_t sequences from CXCR4-using viruses, N_u from CCR5-using viruses, and N_w from dual-using. Each sequence is of q-residues long. Let $X = \{X_1, X_2, \ldots, X_q\}$ be the observation of sequences. X_j is a column vector that contains $N = N_t + N_u + N_w$ observations at the j-th position. Set dataset indicator $Y = \{Y_1, Y_2, \ldots, Y_N\}$ represents the status of co-receptor usage of each sequence: $Y_i = 0$ if i-th sequence is from CCR5, $Y_i = 1$ if CXCR-4, and $Y_i = 2$ if dual-using. The goal is to describe the complicated relationship between the sequence observations (X) and the dataset indicator (Y). Basically, we partitioned the q positions into K groups according to their relationship to Y. Each of the K groups represents one relationship between X and Y. Denote with $I = \{I_1, I_2, \ldots, I_q\}$ as the group indicator, $I_j = k$ ($j = 1, \ldots, q$ and $k = 1, \ldots, K$) means j-th position is partitioned into the k-th group. Given Y, we want to infer I when X is observed and when we have q and K as fixed. The likelihood is $P(X|I,Y)$, and the posterior probability is $P(I|X,Y)$, we have

$$P(I|X,Y) \propto P(I|Y)P(X|I,Y).$$

Assume I is independent from Y, $P(I|Y) = P(I)$.

3.4 Bayesian networks

To further explore the relations between variables and improve the outcome predictive accuracy, recent studies used BNs to model evolutionary escape dynamics of virus during the antiretroviral therapies[35,36]. BNs represent a set of variables, for instance, sequence mutations and drug resistance phenotypes, and their conditional dependencies via a Directed Acyclic Graph (DAG). The learning of the BN structure can be accomplished using many methods, which can be categorized into three groups, i.e., constrain-based, score-based, and tree-based[45,46]. The Peter and Clark (PC) algorithm is one of the most commonly used constrain-based approach to construct causal network, which can be treated as a BN. Once we know the BN structure, the estimation of the conditional probabilities can be obtained using maximum likelihood estimation,

Bayesian estimation, and Expectation-Maximization (EM) algorithm when we have incomplete data. Because a BN is a complete model for the variables and their relationships, it can be used to answer probabilistic queries about them. For example, find out the probability that variable has a specific value, or the most likely explanation for some evidence. In the context of HBV and HIV studies, we can employ BN to investigate epistatic connectivity of drug-resistance mutations or to predict how likely the virus to become resistant to a certain drug. Next, we will briefly introduce the PC algorithm and maximum likelihood estimation for BN inference, and cover some case studies of BN shortly.

PC algorithm. Assume that we have a set of variables $X = (X_1, \ldots, X_n)$ with a global probability distribution about them. a, b, and c represent subsets of variables of X. $\mathrm{Ind}(a, b|c)$ denotes that a and b are conditionally independent given c. PC algorithm assumes faithfulness, which means that a DAG can exactly represent the independence relationships among the variables in X by the d-separation criterion[47]. PC first tries to find the underlying undirected graph (Algorithm 1) and on a posterior step makes the orientation of the edges.

In Algorithm 1, Adj_{X_j} is the set of nodes adjacent to X_j in Graph G'. The intuition is that if there is no link between X_j and Y, S_{X_jY} will contain a set that makes $\mathrm{Ind}(X_j, Y|S)$, and this set will be used in the orientation stage. The orientation step proceeds by checking sets of three variables $\{a, b, c\}$ where only two edges exist among these three. For example, we have edge(a, c) and edge(b, c). If $c \notin S_{ab}$, then it orients the edges from a to c and from b to c as known as a v-structure. Next it tries to orient the rest of the edges similarly but not to create cycles or new v-structures. Note that it is possible that the orientation of some of the edges may be arbitrarily determined.

Algorithm 1 PC algorithm used to find the underlying undirected graph

1: Start with a complete undirected graph G'
2: $i \leftarrow 0$
3: For each $X_j \in X$
4: For each $Y \in \mathrm{Adj}_{X_j}$
5: $S_{X_jY} \leftarrow \varnothing$
6: Test $\exists S \subset \mathrm{Adj}_{X_j}/Y$, and $|S| = i$, and $\mathrm{Ind}(X_j, Y|S)$
7: If S existis
8: $S_{X_jY} \leftarrow S_{X_jY} \cup S$
9: Remove edge between X_j and Y from G'
10: $i \leftarrow i + 1$, repeat Step 3, until $|\mathrm{Adj}_{X_j}| \leqslant i, \forall X_j$

Maximum likelihood estimation. Given a BN structure $G(V, E)$ on a set of variables V and a data set $D \in \mathrm{dom}(V)$ of cases, learning the parameters of the BN means to find vertex potentials $\mathrm{po}(v)_{v \in V}$ subject to some optimality criterion with regard to G and D holds. The simplest criterion is the maximum likelihood criterion, i.e., the probability of the data given the BN is maximal. Instead of the likelihood p, often $\log p$ is used, called log-likelihood. Here, we take a BN with each node corresponding to discrete variables as an example. Given samples $D = \{x_1, \ldots, x_M\}$ from unknown BN that factors over the DAG G, the parameters of a Bayesian model are simply the conditional probabilities that define the factorization. For each node $v_i \in G$, we need to learn $p(v_i|\mathrm{Pa}(v_i))$ which is governed by the parameter $\theta_{v_i|\mathrm{Pa}(v_i)}$, where Pa means the parent and $\mathrm{Pa}(v_i)$ means the parent of node v_i. The log-likelihood we want to maximize is defined by

$$\log l(\theta) = \sum_{m}^{M} \sum_{v \in V} \log \theta_{v(x_m)|\mathrm{Pa}(v(x_m))}.$$

It is easy to prove that $\log l(\theta)$ is maximal if and only if

$$\theta_{v=x|\mathrm{Pa}(v)=y} = \frac{|\{d \in D|d \text{ where } v = x \text{ and } \mathrm{Pa}(v) = y\}|}{|\{d \in D|d \text{ where } \mathrm{Pa}(v) = y\}|}.$$

3.5 Metropolis-hastings algorithm

The Markov Chain Monte Carlo (MCMC) is used to sample from the posterior probability like $P(I|A, B)$ (or $P(I|X, Y)$) via the Metropolis-Hastings (M-H) algorithm to infer which variables are associated with the treatment status, group indicators, etc. The procedure of M-H algorithm is as follows:

(1) Initialization. Randomly assign a starting value $I^{(t)}$ to I, here $t = 0$;

(2) Proposal. Propose a new I as follows: randomly choose one $I_i^{(t)}$ and change it to other values with equal probabilities, set new I as y;

(3) Evaluation. Evaluate the posterior. Since the proposal is symmetric, the acceptance probability is $\alpha(I^{(t)}, y) = \min\{1, P(I = y|A, B)/P(I = I^{(t)}|A, B)\}$;

(4) Update. Generate u from standard uniform distribution $U(0, 1)$ and set

$$I^{(t+1)} = \begin{cases} y, & \text{if } u \leqslant \alpha(I^{(t)}, y); \\ I^{(t)}, & \text{otherwise;} \end{cases}$$

(5) If $t \geqslant N$ (N is the total number of iterations),

stop; otherwise, set $t = t + 1$ and go to Step (2) and repeat this procedure.

3.6 Recursive model selection

RMS procedure is applied to make inferences on the detailed dependence structure among the interacting positions generated by the Bayesian variable partition model[34]. The idea is to select a model from two unrefined models recursively until the data does not support more detailed models. One of the two models is the chain-dependence model and the other is the V-dependence model.

Chain-dependence model[24]. Set of variables X_G follows a chain-dependence model if the index set G can be partitioned into three subsets U, V, and W, such that X_U and X_W are independent given X_V, e.g., $X_U \to X_V \to X_W$ (shown in Fig. 1). The joint distribution of the chain-dependence model is given as

$$p(X_G) = p(X_U)p(X_V|X_U)p(X_W|X_V) = \frac{F(X_V, X_U)F(X_W, X_V)}{F(X_V)},$$

where $F(X_V, X_U, \ldots)$ is the joint probability function of $(X_V, X_U, \ldots)$.

V-dependence model[48]. A set of variables X_G follows a V-dependence model if the index set G can be partitioned into three subsets U, V, and W, such that X_U and X_W are mutually independent, i.e., $X_U \to X_V \leftarrow X_W$ (shown in Fig. 2). The joint distribution of the V-dependence model is

$$p(X_G) = p(X_U)p(X_W)p(X_V|X_U, X_W) = F(X_U)F(X_W)\frac{F(X_U, X_V, X_W)}{F(X_U, X_W)}.$$

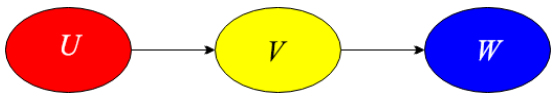


Fig. 1 Chain-dependence model structure.

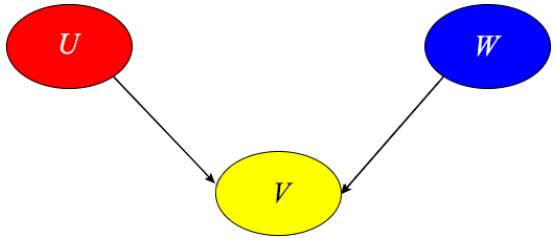


Fig. 2 V-dependence model structure, in which the variables in U are marginally independent of the variables in W.

Note that in these two models, only set W can be empty, in which case these models become the saturated model.

A model indicator $I^{CV} = (I_1^{CV}, I_2^{CV}, \ldots, I_L^{CV})$ can be used to imply the membership of the L positions with $I_j^{CV} = 0$ representing the chain-dependence model and $I_j^{CV} = 1$ indicating the V-dependence model. If we use S to denote the set partition, then the posterior distribution of S and I^{CV} is

$$P(S, I^{CV}|\text{data}) \propto P(\text{data}|S, I^{CV})P(S)P(I^{CV}).$$

One can set equal priors for I^{CV} and S. Then we can use the MCMC algorithm again to sample from the posterior and find the optimal model type and variable selection. The procedure is applied recursively until only single-variable nodes are available.

Then we can apply BVP and RMS sequentially to the data of the different groups to make inferences on the mutations.

3.7 Applications of Bayesian methodology to HBV, HCV, and HIV studies

3.7.1 Applications in HBV studies

The Bayesian methods described have been applied to multiple HBV related studies including detecting correlation between specific mutations in the C-terminus domain of HBV surface antigen and low level of serum HBV-DNA in patients with chronic HBV infection, HBV amino acid sequence mutations in occult infections, and the correlation between HBsAg markers and occult HBV infection and detection. A summary of the results from these studies can be found in Table 1. Note that one of the advantages is that the Bayesian-based method showed the ability of analyzing high-order combinations of positions[48].

The inference of BN has been applied to model the drug resistance of HBV[35]. Thai et al.[35] confirmed that lamivudine resistance is a complex trait encoded by the entire HBV genome rather than by a single mutation based on the investigation of epistatic connectivity using a set of BNs of polymorphic amino acid sites in HBV proteins of pre- and post-treatment viral populations from HBV patients[35]. In most of the patients, drug-resistant HBV variants were evolved from minority subpopulations, the number of sites in BN varied from 76% – 100% of all polymorphic sites.

3.7.2 Applications in HCV studies

By applying BVP model and RMS method to multiple controlled datasets, some interesting findings were

Table 1 A summary of results from applications of Bayesian methods to HBV studies. IQR represents interquartile range.

HBV genotype	Mutation discovered by Bayesian methods	Correlation	Comment
D (and/or A)	M197T, -S204N-Y206C/H-F220L	Serum HBV-DNA<2000 IU/mL[8]	These mutations were localized in the HBsAg C-terminus, known to be involved in virion and/or HBsAg secretion[8].
D (and/or A)	Y206C/H and/or F220L	Lower median (IQR) HBsAg-levels and lower median (IQR) transaminases[8]	
C (HBV and OBI)	RT mutation V173L	Drug resistance in patients receiving antiviral treatments, such as adefovir and lamivudine[49]; HBV vaccine escape[50].	Details results can be found in Ref. [48].
C (HBV and OBI)	H126Q, H126Q+138R	OBI samples[48]	
D	20 HBsAg-mutations	Occult HBV D-genotype infection in vivo[9]	Details results can be found in Ref. [9].

discovered to help understanding the HCV drug response and resistance related mutations.

Fu et al.[17] concentrated on NS5A region particularly for HCV genotype 1a. In NS5A region there are 1344 base pairs, linking to 448 amino acids. The Bayesian methods were applied to the pretreatment sequences of response (47 sequences) and non-response (29 sequences) samples. "The result gives us a reliable idea of the mutation mechanism of positions 49, 349, and 199, 209, 242, 398 which have the highest frequencies."[17] Detailed results can be found in Table 2.

They also found that a lot of positions are not mutating independently. Figure 3 shows the interacting positions detected by BVP in response samples and Fig. 4 shows the interacting positions detected by BVP in non-response samples. Figures 3 and 4 are reproduced with permission from the authors based on their original findings. And some significant discoveries can be found in Table 3.

3.7.3 Applications in HIV studies

The Bayesian methods summarized in previous subsections have also been successfully applied to multiple HIV studies in both single-drug treatments and multiple-drug treatments. A summary of the results of such Bayesian analysis was carried out in Table 4[34,41].

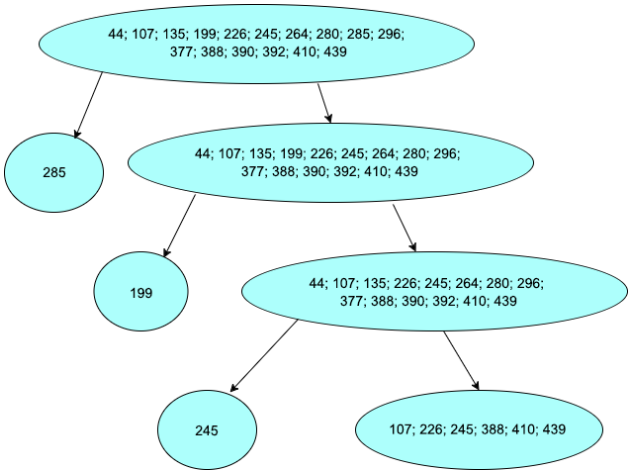


Fig. 3 Flowchart of detected mutation positions and position combinations in the pretreatment sequence of patients who respond to the treatment[42].

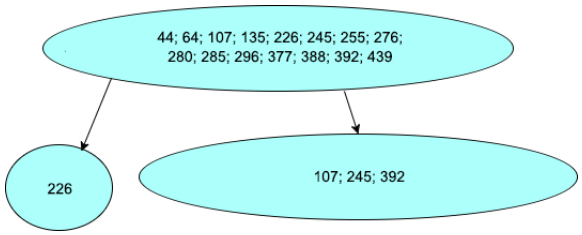


Fig. 4 Flowchart of detected mutation positions and position combinations in the pretreatment sequence of patients who do not respond to the treatment[42].

Table 2 Single positions result summary.

Position	Result	Comment
49, 349	Positions 49 and 349 are statistically different in response and non-response patients and independent of other positions[17].	Position 49 is in membrane attachment region; Position 349 is in the region between V3 and V4; Positions 199 and 209 are in membrane attachment region; Position 242 is in Interferon Sensitivity Determining Region (ISDR); Position 398 is in V3 region. These positions may have some biological influence on drug resistance to IFN and ribavirin[17].
199, 209, 242, 398	Positions 119, 209, 242, and 398 are dependent and demonstrate significant difference in response and non-response patients[17].	

Table 3 Dependence structure inferred by RMS in detail.

Position	Amino acid	Result	Comment
285	E	Frequency is 13.8% in non-response samples and 8.5% in the response samples[42]	
199	L	Frequency decreases from 100% to 87.2%, from non-response samples to response samples[42]	
226	M	Frequency decreases from 20.75% to 14.9%, from non-response samples to response samples[42]	
107, 226, 288, 410, 439	EMIAE	Does not exist in response samples[42]	Those positions combined may be a distinguishing factor for response and non-response patients[42].
107, 226, 288, 410, 439	KEIAG, TMVAG, TLIAE	Only exist in non-response samples[42]	

Table 4 A summary of results from applications of Bayesian methods to HIV drug resistance studies[24].

Drug	Antiretroviral effect	Mutation interaction discovered by Bayesian methods	Comment
Indinavir (IDV)	Protease inhibitor	{24, 47 {32{46 ⊥ 54 | 82}}}{10, 71}{73, 90}	Interesting group {46, 54, 82}①
Nevirapine	Non-nucleoside RT inhibitor	{106}{188}{103 ? 181}{190}	Weak interactions
Zidovudine	Nucleoside analog RT inhibitor	{41, 210, 215}{67, 219}{70}	Further biochemicalinvestigations needed②
IDV, NFV	Protease inhibitors	{24, 54, 82}{30, 88}{73, 90}	6 positions disappeared③
IDV, SQV	Protease inhibitors	{61, 71}{46, 54, 82}{73, 90}	Other details ambiguous
IDV, NFV, SQV	Protease inhibitors	{30, 88}{73, 90}{24, 46, 54, 82}	Ambiguous structure in 3rd group

Notes: Epistatic mutations discovered with BVP approach are partitioned using RMS algorithm. Independence groups are enclosed in brackets. "?" indicates inconclusive result.

① Sequential mutation acquisition in this group leads to conditional independence. The results were confirmed by the Molecular Dynamics (MD) simulations.

② It is not possible to study the structural basis of mutations using MD simulations for Zidovudine.

③ When compared to single-drug treatment profiles.

One can observe that several statistically significant interaction patterns among resistance causing mutations have been discovered using the Bayesian methods. It is important that the molecular basis of multiple interacting mutations found by RMS was analyzed with MD simulations and free energy calculations[41]. "Therefore, this is an example of the statistical study where biological processes underling drug resistance can be extracted from the discovered independence groups."[24]

In addition, Beerenwinkel et al.[36] used isotonic conjunctive Bayesian networks, a class of BN, to model the evolutionary escape dynamics of HIV-1. The partial order constraints among viral resistance mutations were employed to generate to a limited number of mutational pathways, and phenotypic drug resistance was modeled as monotonically increasing along these escape pathways. Using this model, the individualized genetic barrier[51,52] which means the probability of the virus is not acquiring additional mutations that confer resistance to each drug was derived and used to quantify the virus' genetic potential for developing drug resistance under combination therapy. The experimental results showed that this data-derived predictor, individualized genetic barrier of treatment outcome, has the potential to advance the understanding of genotypic drug resistance tests. Recently, Chaillon et al.[37] used Bayesian-based statistical modeling to access the likelihood of sexual transmission and persistence of Drug Resistance Mutations (DRM) regarding HIV infection. Their goal was to assess the rate at which a drug resistance mutation was transmitted from original partners to their receivers and whether the transmission was

affected by the relative frequency or the absolute copy numbers of each mutation[37]. For modeling the between-pair variability and between-site variability, they fitted Bayesian hierarchical Bernoulli logistic regression models with within-pair fixed effect, and two random intercepts for pairs and sites as crossed random effects[53]. They assessed the model convergence with the Gelman-Rubin convergence statistic $\hat{R}$[53]. The Bayes factor, which is a ratio of the probability of obtaining data under null and alternative hypotheses, was used for null-hypothesis significance testing[54]. Bayes Factors (BFs) of 1 to 3, 3 to 10, 10 to 30, 30 to 100, and $\geqslant$100 are considered anecdotal, moderate, strong, very strong, and extreme evidence for against a null hypothesis[55]. One key conclusion from this Bayesian analysis is that the majority of DRM (the relative frequency of DRM $\geqslant 20\%$) were consistently transmitted from source to recipient, the probability of detecting a minority DRM (the relative frequency of DRM $< 20\%$) in the recipient was not increased when the same minority DRM was detected in the source (BF $= 6.37$).

4 Summary and Discussion

In this review article, we presented and summarized important applications of the Bayesian inference paradigm in three types of studies. We reviewed Bayesian-based statistical approaches including BVP, BN, and RMS procedure and their applications in HBV, HCV and HIV studies. Firstly, in HBV studies, the evidence has been provided that there exists some specific HBsAg-mutations which correlate with its replicative potential, particularly, the state of low level serum HBV-DNA and HBsAg[8, 9, 40, 48, 56]. Secondly, several independent HCV-drug-resistance-related mutations and interacting mutation patterns have been detected[17, 42]. Moreover, a detailed understanding of complex interacting mutation patterns and new genetic determinants underlying co-receptor usage in HIV-1 have been revealed[30, 34, 41, 57].

The Bayesian statistical analysis of viral genetic characteristics summarized in the review is an advanced and innovative analytical approach that can connect statistical modeling with molecular dynamic simulations[42], thus to detect interacting mutations. However, certain significant issues should be addressed in more detailed and be paid more attention to, such as the emergence of bias caused by multiple subpopulations in the data and the decreased sensitivity of the BVP caused by the transmitted resistance occurrence[24]. Moreover, many factors that may affect the results of the above studies about the three viruses have been ignored since the summarized BVP method is only designed as a baseline analysis. For instance, further studies might be needed to strengthen the correlation between HBsAg mutations and low serum HBV-DNA due to the overlapping of HBsAg and RT genes[8].

Extensions to the summarized Bayesian methods and different Bayesian approaches have also been developed and applied to related research areas. Guo et al.[58] introduced a "simple, fast and powerful method, named DAM, using Bayesian inference to detect genome-wide multi-locus epistatic interactions in multiple diseases". Wang et al.[59] proposed a Bayesian model for detection of high-order interactions among genetic variants in genome-wide association studies. Random forest and Bayesian prediction have also been used for HBV reactivation analysis[60]. Bayesian analysis has also been applied to cancer research[61, 62] and neuroscience[63–67].

Despite all other ignored possibilities, the Bayesian methods summarized here have given us some valuable information that will contribute to not only further studies in related areas but also the development of antiviral treatment.

References

[1] Hepatitis B foundation, Hepatitis B foundation statistics, http://www.hepb.org/statistics.htm, 2009.

[2] World health organization, Hepatitis B: Fact sheet No. 204, http://www.who.int/mediacentre/factsheets/fs204/en/, 2016.

[3] M. Nettleman, Hepatitis B (HBV, Hep B), http://www.emedicinehealth.com/hepatitis_b/article_em.htm, 2019.

[4] C. R. Bourne, S. P. Katen, M. R. Fulz, C. Packianathan, and A. Zlotnick, A mutant hepatitis B virus core protein mimics inhibitors of icosahedral capsid self-assembly, *Biochemistry*, vol. 48, no. 8, pp. 1736–1742, 2009.

[5] World Health Organization, Hepatitis B: The Hepatitis B virus, https://apps.who.int/iris/bitstream/handle/10665/67746/WHO_CDS_CSR_LYO_2002.2_HEPATITIS_B.pdf, 2002.

[6] J. P. Miguet and D. Dhumeaux, *Progress in Hepatology 93*. John Libbey Eurotext, 1993.

[7] A. F. Voevodin and P. A. Marx, *Simian Virology*. Ames, IA, USA: Wiley-Blackwell, 2009.

[8] C. Mirabelli, M. Surdo, F. Van Hemert, Z. C. Lian, R. Salpini, V. Cento, M. F. Cortese, M. Aragri, M. Pollicita,

C. Alteri, et al., Specific mutations in the C-terminus domain of HBV surface antigen significantly correlate with low level of serum HBV-DNA in patients with chronic HBV infection, *J. Infect.*, vol. 70, no. 3, pp. 288–298, 2015.

[9] V. Svicher, V. Cento, M. Bernassola, M. Neumann-Fraune, F. Van Hemert, M. J. Chen, R. Salpini, C. Liu, R. Longo, M. Visca, et al., Novel HBsAg markers tightly correlate with occult HBV infection and strongly affect HBsAg detection, *Antiviral Res.*, vol. 93, no. 1, pp. 86–93, 2012.

[10] T. Poynard, V. Leroy, M. Cohard, T. Thevenot, P. Mathurin, P. Opolon, and J. P. Zarski, Meta-analysis of interferon randomized trials in the treatment of viral hepatitis C: Effects of dose and duration, *Hepatology*, vol. 24, no. 4, pp. 778–789, 1996.

[11] G. L. Davis, R. Esteban-Mur, V. Rustgi, J. Hoefs, S. C. Gordon, C. Trepo, M. L. Shiffman, S. Zeuzem, A. Craxi, M. H. Ling, et al., Interferon alfa-2b alone or in combination with ribavirin for the treatment of relapse of chronic hepatitis C, *N. Engl. J. Med.*, vol. 339, no. 21, pp. 1493–1499, 1998.

[12] J. G. McHutchison, S. C. Gordon, E. R. Schiff, M. L. Shiffman, W. M. Lee, V. K. Rustgi, Z. D. Goodman, M. H. Ling, S. Cort, and J. K. Albrecht, Interferon alfa-2b alone or in combination with ribavirin as initial treatment for chronic hepatitis C, *N. Engl. J. Med.*, vol. 339, no. 21, pp. 1485–1492, 1998.

[13] A. Macdonald and M. Harris, Hepatitis C virus NS5A: Tales of a promiscuous protein, *J. Gen. Virol.*, vol. 85, no. 9, pp. 2485–2502, 2004.

[14] N. Enomoto, I. Sakuma, Y. Asahina, M. Kurosaki, T. Murakami, C. Yamamoto, Y. Ogura, N. Izumi, F. Marumo, and C. Sato, Mutations in the nonstructural protein 5A gene and response to interferon in patients with chronic hepatitis C virus 1b infection, *N. Engl. J. Med.*, vol. 334, no. 2, pp. 77–82, 1996.

[15] M. J. Jr. Gale, M. J. Korth, N. M. Tang, S. L. Tan, D. A. Hopkins, T. E. Dever, S. J. Polyak, D. R. Gretch, and M. G. Katze, Evidence that hepatitis C virus resistance to interferon is mediated through repression of the PKR protein kinase by the nonstructural 5A protein, *Virology*, vol. 230, no. 2, pp. 217–227, 1997.

[16] M. Jr. Gale, C. M. Blakely, B. Kwieciszewski, S. L. Tan, M. Dossett, N. M. Tang, M. J. Korth, S. J. Polyak, D. R. Gretch, and M. G. Katze, Control of PKR protein kinase by hepatitis C virus nonstructural 5A protein: Molecular mechanisms of kinase regulation, *Mol. Cell Biol.*, vol. 18, no. 9, pp. 5208–5218, 1998.

[17] Y. Fu, G. Chen, X. Guo, J. Zhang, and Y. Pan, Analyzing the effects of pretreatment diversity on HCV drug treatment responsiveness using Bayesian partition methods, *J. Bioinform. Proteom. Rev.*, vol. 1, no. 1, pp. 1–6, 2015.

[18] N. Enomoto and C. Sato, Clinical relevance of hepatitis C virus quasispecies, *J. Viral. Hepat.*, vol. 2, no. 6, pp. 267–272, 1995.

[19] N. Enomoto, I. Sakuma, Y. Asahina, M. Kurosaki, T. Murakami, C. Yamamoto, N. Izumi, F. Marumo, and C. Sato, Comparison of full-length sequences of interferon-sensitive and resistant hepatitis C virus 1b. Sensitivity to interferon is conferred by amino acid substitutions in the NS5A region, *J. Clin. Invest.*, vol. 96, no. 1, pp. 224–230, 1995.

[20] M. Gerotto, F. Dal Pero, D. G. Sullivan, L. Chemello, L. Cavalletto, S. J. Polyak, P. Pontisso, D. R. Gretch, and A. Alberti, Evidence for sequence selection within the non-structural 5A gene of hepatitis C virus type 1b during unsuccessful treatment with interferon-α, *J. Viral. Hepat.*, vol. 6, no. 5, pp. 367–372, 1999.

[21] M. J. Clemens and A. Elia, The double-stranded RNA-dependent protein kinase PKR: Structure and function, *J. Interf. Cytokine Res.*, vol. 17, no. 9, pp. 503–524, 1997.

[22] M. Pirmohamed and D. J. Back, The pharmacogenomics of HIV therapy, *Pharmacogenomics J.,* vol. 1, pp. 243–253, 2001.

[23] T. Lengauer and T. Sing, Bioinformatics-assisted anti-HIV therapy, *Nat. Rev. Microbiol.,* vol. 4, no. 10, pp. 790–797, 2006.

[24] I. Kozyryev and J. Zhang, Bayesian analysis of complex interacting mutations in HIV drug resistance and cross-resistance, in *Advance in Structural Bioinformatics*, D. Q. Wei, Q. Xu, T. Z. Zhao, and H. Dai, eds. Springer, 2015, pp. 367–383.

[25] A. Engelman and P. Cherepanov, The structural biology of HIV-1: Mechanistic and therapeutic insights, *Nat. Rev. Microbiol.*, vol. 10, no. 4, pp. 279–290, 2012.

[26] C. Flexner, HIV drug development: The next 25 years, *Nat. Rev. Drug Discov.*, vol. 6, no. 12, pp. 959–966, 2007.

[27] E. A. Berger, R. W. Doms, E. M. Fenyö, B. T. M. Korber, D. R. Littman, J. P. Moore, Q. J. Sattentau, H. Schuitemaker, J. Sodroski, and R. A. Weiss, A new classification for HIV-1, *Nature*, vol. 391, no. 6664, p. 240, 1998.

[28] R. R. Regoes and S. Bonhoeffer, The HIV coreceptor switch: A population dynamical perspective, *Trends Microbiol.*, vol. 13, no. 6, pp. 269–277, 2005.

[29] T. L. Hoffman and R. W. Doms, HIV-1 envelope determinants for cell tropism and chemokine receptor use, *Mol. Membr. Biol.*, vol. 16, no. 1, pp. 57–65, 1999.

[30] M. J. Chen, V. Svicher, A. Artese, G. Costa, C. Alteri, F. Ortuso, L. Parrotta, Y. Liu, C. Liu, C. F. Perno, et al., Detecting and understanding genetic and structural features in HIV-1 B subtype V3 underlying HIV-1 co-receptor usage, *Bioinformatics*, vol. 29, no. 4, pp. 451–460, 2013.

[31] N. Yu, Z. H. Li, and Z. Yu, Survey on encoding schemes for genomic data representation and feature learning — From signal processing to machine learning, *Big Data Min. Anal.*, vol. 1, no. 3, pp. 191–210, 2018.

[32] O. Zagordi, L. Geyrhofer, V. Roth, and N. Beerenwinkel, Deep sequencing of a genetically heterogeneous sample: Local haplotype reconstruction and read error correction, *J. Comput. Biol.*, vol. 17, no. 3, pp. 417–428, 2010.

[33] O. Zagordi, R. Klein, M. Däumer, and N. Beerenwinkel, Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies, *Nucl. Acids Res.*, vol. 38, no. 21, pp. 7400–7409, 2010.

[34] J. Zhang, T. J. Hou, W. Wang, and J. S. Liu, Detecting and understanding combinatorial mutation patterns responsible for HIV drug resistance, *Proc. Natl. Acad. Sci. USA.*, vol. 107, no. 4, pp. 1321–1326, 2010.

[35] H. Thai, D. S. Campo, J. Lara, Z. Dimitrova, S. Ramachandran, G. L. Xia, L. Ganova-Raeva, C. G. Teo, A. Lok, and Y. Khudyakov, Convergence and coevolution of hepatitis B virus drug resistance, *Nat. Commun.*, vol. 3, p. 789, 2012.

[36] N. Beerenwinkel, H. Montazeri, H. Schuhmacher, P. Knupfer, V. Von Wyl, H. Furrer, M. Battegay, B. Hirschel, M. Cavassini, P. Vernazza, et al., The individualized genetic barrier predicts treatment response in a large cohort of HIV-1 infected patients, *PLoS Comput. Biol.*, vol. 9, no. 8, p. e1003203, 2013.

[37] A. Chaillon, M. Nakazawa, J. O. Wertheim, S. J. Little, D. M. Smith, S. R. Mehta, and S. Gianella, No substantial evidence for sexual transmission of minority HIV drug resistance mutations in men who have sex with men, *J. Virol.*, vol. 91, no. 21, p. e00769-17, 2017.

[38] V. Svicher, C. Alteri, A. Artese, J. M. Zhang, G. Costa, F. Mercurio, R. D'Arrigo, S. Alcaro, G. Palù, M. Clementi, et al., Identification and structural characterization of novel genetic elements in the HIV-1 V3 loop regulating coreceptor usage, *Antivir. Ther.*, vol. 16, no. 7, pp. 1035–1045, 2011.

[39] V. Svicher, M. Chen, C. Alteri, G. Costa, S. Dimonte, L. Chang, L. Parrotta, C. Dimaio, M. Surdo, P. Saccomandi, et al., Key genetic elements in HIV-1 gp120 V1, V2 and C4 domains tightly and differentially modulate gp120 interaction with the CCR5 and CXCR4 N-terminus and HIV-1 antigenic potential, in *Proc. 2^{nd} Int. Workshop on HIV & Hepatitis Virus Drug Resistance and Curatove Strategies*, Los Cabos, Mexico, 2011.

[40] V. Svicher, V. Cento, M. Bernassola, M. Neumann-Fraune, M. Chen, R. Salpini, L. Chang, R. Longo, M. Visca, S. Romano, et al., Specific HBsAg genetic determinants are associated with occult HBV infection *in vivo* and HBsAg detection, in *Proc. 2^{nd} Int. Workshop on HIV & Hepatitis Virus Drug Resistance and Curatove Strategies*, Los Cabos, Mexico, 2011.

[41] J. Zhang, T. J. Hou, Y. Liu, G. Chen, X. Yang, J. S. Liu, and W. Wang, Systematic investigation on interactions for HIV drug resistance and cross-resistance among protease inhibitors, *J. Proteome Sci. Comput. Biol.*, vol. 1, no. 1, p. 2, 2012.

[42] Y. Fu, G. Chen, L. Z. Fu, and J. Zhang, Investigating genotype 1a HCV drug resistance in NS5A region via Bayesian inference, *Tsinghua Sci. Technol.*, vol. 20, no. 5, pp. 484–490, 2015.

[43] J. Fellay, D. L. Ge, K. V. Shianna, S. Colombo, B. Ledergerber, E. T. Cirulli, T. J. Urban, K. L. Zhang, C. E. Gumbs, J. P. Smith, et al., Common genetic variation and the control of HIV-1 in humans, *PLoS Genet.*, vol. 5, no. 12, p. e1000791, 2009.

[44] R. M. Neal, Markov chain sampling methods for dirichlet process mixture models, *J. Comput. Graph. Stat.*, vol. 9, no. 2, pp. 249–265, 2000.

[45] D. Dash and M. J. Druzdzel, Robust independence testing for constraint-based learning of causal structure, in *Proc. 19^{th} Conf. Uncertainty in Artificial Intelligence*, Acapulco, Mexico, 2002, pp. 167–174.

[46] J. Cheng, R. Greiner, J. Kelly, D. Bell, and W. R. Liu, Learning Bayesian networks from data: An information-theory based approach, *Artif. Intell.*, vol. 137, nos. 1&2, pp. 43–90, 2002.

[47] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Elsevier, 2014.

[48] Z. C. Lian, Q. N. Tian, Y. Liu, V. Cento, R. Salpini, C. F. Perno, V. Svicher, G. Chen, C. Li, and J. Zhang, Detecting hepatitis B viral amino acid sequence mutations in occult hepatitis B infections via bayesian partition model, *J. Proteomics Bioinform.*, doi:10.4172/jpb. S6-005.

[49] O. Lada, Y. Benhamou, A. Cahour, C. Katlama, T. Poynard, and V. Thibault, In vitro susceptibility of lamivudine-resistant hepatitis B virus to adefovir and tenofovir, *Antivir. Ther.*, vol. 9, no. 3, pp. 353–363, 2004.

[50] J. Sheldon, B. Ramos, J. Garcia-Samaniego, P. Rios, A. Bartholomeusz, M. Romero, S. Locarnini, F. Zoulim, and V. Soriano, Selection of Hepatitis B Virus (HBV) vaccine escape mutants in HBV-infected and HBV/HIV-coinfected patients failing antiretroviral drugs with anti-HBV activity, *J. Acquir. Immune Defic. Syndr.*, vol. 46, no. 3, pp. 279–282, 2007.

[51] R. Gish, J. D. Jia, S. Locarnini, and F. Zoulim, Selection of chronic hepatitis B therapy with high barrier to resistance, *Lancet Infect. Dis.*, vol. 12, no. 4, pp. 341–353, 2012.

[52] N. Beerenwinkel, M. Däumer, T. Sing, J. Rahnenführer, T. Lengauer, J. Selbig, D. Hoffmann, and R. Kaiser, Estimating HIV evolutionary pathways and the genetic barrier to drug resistance, *J. Infect. Dis.*, vol. 191, no. 11, pp. 1953–1960, 2005.

[53] J. H. Albert and S. Chib, Bayesian analysis of binary and polychotomous response data, *J. Am. Stat. Assoc.*, vol. 88, no. 422, pp. 669–679, 1993.

[54] J. O. Berger and L. R. Pericchi, The intrinsic Bayes factor for model selection and prediction, *J. Am. Stat. Assoc.*, vol. 91, no. 433, pp. 109–122, 1996.

[55] P. M. Lee, *Bayesian Statistics*: *An Introduction, 4th Edition*. John Wiley & Sons, 2012.

[56] M. Surdo, M. F. Cortese, C. Mirabelli, R. Salpini, J. Zhang, F. Van Hemert, V. Cento, M. Pollicita, G. Gubertini, G. M. De Sanctis, et al., Key patterns of HBsAg mutations correlate with mechanisms underlying levels of serum HBVDNA, *J. Hepatol.*, vol. 60, no. S1, p. S299, 2014.

[57] V. Svicher, F. Mercurio, and A. Artese, Signature mutations in V3 and bridging sheet domain of HIV-1 gp120 HIV-1 are specifically associated with dual tropism and modulate the interaction with CCR5 N-terminus, *Infection*, vol. 39, no. 1, pp. S11–S91, 2011.

[58] X. Guo, J. Zhang, Z. P. Cai, D. Z. Du, and Y. Pan, Searching genome-wide multi-locus associations for multiple diseases based on Bayesian inference, *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 14, no. 3, pp. 600–610, 2017.

[59] J. X. Wang, T. Joshi, B. Valliyodan, H. Y. Shi, Y. C. Liang, H. T. Nguyen, J. Zhang, and D. Xu, A Bayesian model for detection of high-order interactions among genetic variants in genome-wide association studies, *BMC Genomics*, vol. 16, p. 1011, 2015.

[60] H. Wang, Y. Liu, and W. Huang, Random forest and Bayesian prediction for Hepatitis B virus reactivation, in *2017 13th Int. Conf. Natural Computation, Fuzzy Systems and Knowledge Discovery*, Guilin, China, 2017.

[61] L. Zapata, H. Susak, O. Drechsel, M. Friedlander, X. Estivill, and O. Stephan, Bayesian inference of cancer driver genes using signatures of positive selection, bioRxiv: 059360, 2017.

[62] L. Xu, Y. B. Zheng, J. Liu, D. Rakheja, S. Singleterry, T. W. Laetsch, J. F. Shern, J. Khan, T. J. Triche, D. S. Hawkins, et al., Integrative Bayesian analysis identifies rhabdomyosarcoma disease genes, *Cell Rep*., vol. 24, no. 1, pp. 238–251, 2018.

[63] X. Guo, B. Liu, L. Chen, G. T. Chen, Y. Pan, and J. Zhang, Bayesian inference for functional dynamics exploring in fMRI data, *Comput. Math. Methods Med.*, vol. 2016, p. 3279050, 2016.

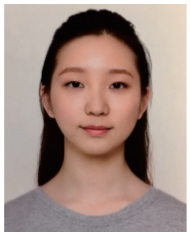
[64] X. C. Xiao, B. Liu, J. Zhang, X. L. Xiao, and Y. Pan, Detecting change points in fMRI data via bayesian inference and genetic algorithm model, in *Bioinformatics Research and Applications*, Z. P. Cai, O. Daescu, and M. Li, eds. Springer, 2017, pp. 314–324.

[65] X. Xiao, B. Liu, J. Zhang, X. Xiao, and Y. Pan, An optimized method for Bayesian connectivity change point model, *J. Comput. Biol.*, vol. 25, no. 3, pp. 337–347, 2018.

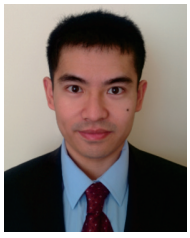
[66] B. Liu, X. Guo, and J. Zhang, Bayesian Bi-cluster change-point model for exploring functional brain dynamics, in *Proc. 2018 Int. Conf. Bioinformatics and Computational Biology*, Las Vegas, NV, USA, 2018.

[67] J. Zhang, T. Liu, and G. Deshpande, Probablistic methods in computational neuroscience, *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 15, no. 2, pp. 535–536, 2018.

Bing Liu received the PhD degree in mathematics and statistics from Georgia State University in 2017. He is currently a research scientist at Eli Lilly and Company. His research interests include Bayesian methods in brain connectivity change point detection, big data analysis, data mining and machine learning, design and analysis of clinical trials, and empirical likelihood.



Shishi Feng received the MS degree from Georgia State University in 2018. She is currently a data solution analyst at AnalyticsIQ, Inc., providing marketing advices based on internal databases, statistical analysis, and ad hoc analysis.



Xuan Guo received the PhD degree in computer science from Georgia State University in 2015. He is currently an assistant professor at University of North Texas. From 2015 to 2017, he was a post-doctoral research associate at Oak Ridge National Laboratory. His research focuses on big data mining and high-performance computing and their applications in environment, food, and health sectors. He serves many premium conferences and journals as editor, chair, or TPC member.

Jing Zhang received the PhD degree from Harvard University in 2009. She is currently an assistant professor at Georgia State University. From 2009 to 2010, she was a postdoc at Harvard University. From 2010 to 2014, she was an assistant professor at Yale University. Her research interests include Bayesian inference on complicated interactions, big data analysis, NeuroImage analysis, brain mapping, functional brain imaging analysis, statistical genetics, epigenetics, computational virology, Bayesian network, graph models, and computational neuroscience. She has authored/co-authored over 40 peer reviewed journal/conference papers. She received Seessel Awards at Yale University from 2010 to 2013, and Brains and Behavior Seed Grant Georgia State University from 2015 to 2016.