The characteristics of hepatitis B virus sequence with virology and clinical pathology

Yi-Chun Lin[a] Koun-Tem Sun[b] Wen-Chun Lin[c] Ting-Tsung Chang[d] Yueh-Min Huang[a]

[a] Department of Engineering Science [c]Institute of Molecular Medicine [d]Institute of Basic Medical Sciences, [acd]National Cheng Kung University ; [b]Department of Information and Learning Technology, Nation University of Tainan, Tainan, Taiwan

ktsun@mail.nutn.edu.tw

Abstract

The hepatitis B virus infection is the invisible killer of pathological changes of liver, and endangered people's health for a long time. Recently, related studies have found that difficult genotypes and HBsAg seroconversion effect the pathological change of liver. In this paper, we explored the variation of hepatitis B virus (HBV) on different genotypes and the HBsAg seroconversion in patients. Then, we develop a technique for association rules to solve this problem.

1. Samples and data preprocessing

To investigate the HBV DNA sequences, we collected analyzing samples form routine diagnostic blood specimens amount to twenty-two patients and obtained the amino acid sequences amount to three hundred and fifty-eight from NCBI database.

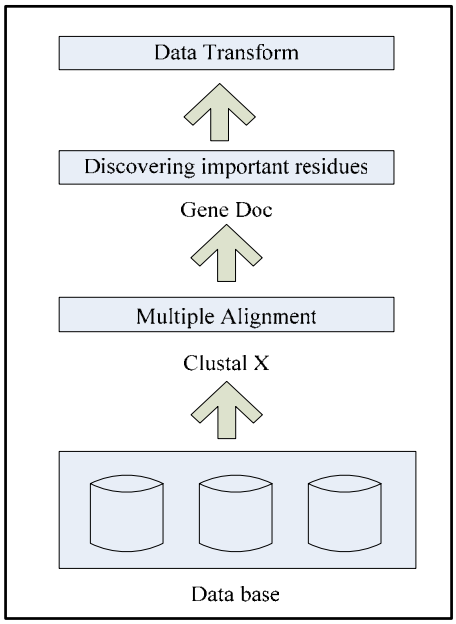

Figure 1. The flow chart of data preprocessing

2. Mining specific positions of HBV sequence

Of particular interest in study is employed association analysis to find the specific positions of HBV sequence and which have relation. Then, we further prune the discovered associations to remove those insignificant associations and find a set of useful rules.

3.1 Association rules of distinguish genotypes

We discovered that the rules related to genotyping appear in p gene, S gene, PreS1 gene, respectively.

Table 1.The specific positions in p gene and s gene

Number	Gene	Association rules
rule1	P	{position_104=N,position_234=G,position_584=H,position_199=P}
rule3	S	{position_8=L,position_47=Vposition_49=L,position_56=Q,position_57=I,position_59=S,position_64=C,position_77=-,position_85=C}
rule7	PreS1	{position_10=K,position_35=K,position_39=E,position_45=L,position_48=H,position_51=N,position_54=D,position_57=K}

Experimentally, the genotype A is classified according to that when position_104 shows “K” of P gene, position_47 shows “L” and position_59 shows N of S gene at the same time. Therefore, we can corollary the following rules.

1. {position_104=K, position_49=L, position_59=N} => Genotype A.
2. {position_199=P, position_47=V, position_59=S}=> Genotype B.
3. {position_104 =R, position_199=Q } => Genotype C.
4. {position_199 =Q, position_47=V, position_49=L} => Genotype D.
5. {position_104 =K, position_199=H, position_584=N} => Genotype E.
6. {position_104 =L, position_234=N} => Genotype F.
7. {position_104=K,position_199=Q,position_234=R, position_47=V,position_49=P } => Genotype G.
8. {position_199=A, position_234=N, position_584=A} => Genotype H.