FEBS Open Bio 2 (2012) 98–102

journal homepage: www.elsevier.com/locate/febsopenbio

Using decision tree learning to predict the responsiveness of hepatitis C patients to drug treatment

Yoshihiro Kawamura [a,1], Shigeru Takasaki [b,*,1], Masashi Mizokami [a,*]

[a] The Research Center for Hepatitis and Immunology, National Center for Global Health and Medicine, 1-7-1 Konodai, Ichikawa, Chiba 272-8516, Japan
[b] Toyo University, Izumino 1-1-1, Ora-gun Itakuracho, Gunma 374-0193, Japan

ARTICLE INFO

Article history:
Received 16 March 2012
Accepted 20 April 2012

Keywords:
Chronic hepatitis C genotype 1
Pegylated interferon α
Rebavirin
Sustained virologic response
Null virologic response
Single nucleotide polymorphism
Genome-wide association study
Decision tree learning

ABSTRACT

The recommended treatment for patients with chronic hepatitis C, pegylated interferon α (PEG-IFN-α) plus rebavirin (RBV), does not provide a sustained virologic response in all patients, especially those with hepatitis C virus (HCV) genotype 1. It is therefore important to predict whether or not a new patient with HCV genotype 1 will be cured by the recommended treatment. We propose a prediction method for a new patient using a decision tree learning model based on SNPs evaluated in a genome-wide association study. By the decision tree learning for 142 Japanese patients with HCV genotype 1 (78 with null virologic response and 64 with virologic response), we can predict with high probability (93%) whether or not a new patient with HCV will be helped by the recommended treatment.

1. Introduction

Chronic infection with hepatitis C virus (HCV) is a global health problem affecting a significant proportion of world's population. The World Health Organization estimated that in 1999 there were 170 million HCV carriers worldwide, with 3–4 million new cases appearing each year [1,2]. A 48-week course of PEG-IFN-α with RBV is the recommended treatment for patients with HCV involves, but many patients will not be cured by it [3–5]. It also has side effects that prevent some patients from completing therapy [6]. For these reasons, identification of the determinants of responsiveness to the PEG-IFN-α with RBV treatment is a matter of high priority.

Recent genome-wide association studies performed in order to identify human genetic contributions to anti-HCV treatment response have indicated that genetic polymorphisms near the IL28B gene are associated with responses to HCV treatment [7–11]. Tanaka et al. reported that, within a Japanese population of patients with HCV genotype 1, those with minor alleles (TG and GG) of SNP rs8099917 were more strongly associated with null virological response (NVR) than were those with major alleles (TT) ($P = 2.65 \times 10^{-32}$) [7]. They also reported a logistic regression model based on SNP, age, gender, re-treatment, platelet count, aminotransferase level, fibrosis stage, and HCV-RNA level indicated that rs8099917 is the most significant factor for NVR [7]. Ge and his colleagues reported not only that a genetic polymorphism near the IL28B gene, encoding interferon-λ-3 (IFN-λ-3), is associated with an approximately twofold change in responsiveness to treatment, both among African-American patients ($P = 2.06 \times 10^{-3}$) and among patients of European ancestry ($P = 1.06 \times 10^{-26}$) but also that a polymorphism on chromosome 19, rs12979860, is strongly associated with sustained virologic (SVR) in all patient groups ($P = 1.37 \times 10^{-28}$) [8–10]. They also noted that their regression model showed that the CC genotype is associated with a more substantial difference in responsiveness rate than was any of the other known baseline predictors included in the model. Suppiah et al. reported an association to SVR within the gene region encoding interleukin IL28B (rs8099917 combined $P = 9.25 \times 10^{-9}$, OR = 1.98, 95% CI = 1.57–2.52) and indicated that host genetics may be useful for predicting drug responsiveness [11].

The recent reports made clear that genetic polymorphisms near the IL28B gene are associated with the responses of patients with

Abbreviations: GDI, Gini diversity index; GWAS, genome-wide association study; HCV, hepatitis C virus; Het, one major and one minor genotype; MM, both major genotypes; mm, both minor genotypes; NVR, null virologic response; OR, Odds ratio; PEG-IFN-α, pegylated interferon α; RBV, ribavirin; SNPs, single nucleotide polymorphisms; SVR, sustained virologic response

[*] Corresponding authors.

E-mail addresses: s_takasaki@toyo.jp (S. Takasaki), mmizokami@imcjk2.hosp.go.jp (M. Mizokami).

[1] These authors contributed equally to this work.

2211-5463/$36.00 © 2012 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.fob.2012.04.007

HCV genotype 1 to the recommended drug treatment and indicated that predicting responsiveness to the treatment is an urgent necessity. We therefore developed a new method for predicting this responsiveness.

2. Materials and methods

Although the virological responses of patients with HCV genotype 1 to PEG-IFN-α with RBV have been reported to be strongly associated with genetic polymorphisms, there is no report of responsiveness being predicted from the polymorphisms. It is quite important to know whether or not a new patient with HCV genotype 1 will be cured by PEG-IFN-α with RBV before beginning the treatment. Predicting patient's responsiveness from the related SNPs would help to reduce side effects and treatment costs.

The method we propose for predicting responsiveness uses decision tree learning based on the genome-wide SNPs. Decision tree learning is a method that uses inductive inference to approximate a target function that will produce discrete values. It is generally best suited to problems in which instances are represented by attribute-value pairs and the target function has discrete output values. A decision tree classifies each example into a class corresponding to one of the output values [12].

2.1. Model for decision tree learning

Individual SNPs and their alleles (major, hetero, and minor genotypes) in the genome-wide association study (GWAS) were used as attributes. 142 Japanese HCV genotype 1 patients (64 with virologic responses (VRs) and 78 with null virologic responses (NVRs)) were used as training instances [7]. To carry out the supervised learning for their classification, we partitioned the training instances into two data sets: training data for growing the decision tree, and testing data for pruning the decision tree. The classification processes were carried out in two phases: one for growing it and the other for pruning it. We used the SLIQ/SPRINT algorithm in the decision tree learning [13].

The SLIQ/SPRINT algorithm uses a two-branch (yes/no) approach, so for two combinations we selected the three types of branches listed in Table 1. The previous analyses of genome-wide drug responses for HCV genotype 1 patients indicated that specific SNPs are closely associated with VRs/NVRs [7–11]. To analyze SNP contributions to the drug effects, one first needs to calculate an evaluation function (such as the Gini diversity index (GDI)) for three types of branches of individual SNPs and determine the maximum value of that function. Each SNP has three GDIs for individual alleles, and each GDI can be computed in the following way [13]:

$$\begin{aligned} GDI &= 1 - \sum_{l=1}^{K} P(C_l)^2 - \sum_{m=1}^{J} a_m \Big(1 - \sum_{l=1}^{K} P(C_{ml})^2\Big) \\ a_m &= \tfrac{n_m}{N} \quad (m = 1, 2, \ldots, J), \quad \sum_{m=1}^{J} a_m = 1 \end{aligned} \tag{1}$$

Table 1
The three types of branches in the decision tree.

Type	Combination 1	Combination 2
1	MM	Het + mm
2	Het	MM + mm
3	mm	MM + Het

MM: both nucleotides are major genotypes (e.g., CC, C: major genotype).
Het: one nucleotide is a major genotype and the other is a minor genotype (e.g., TC, T: minor genotype).
mm: both nucleotides are minor genotypes (e.g., TT).

where K is the number of classes (in this case $K = 2$, VR and NVR), $P(C_l)$ is the probability of l class for each SNP in the instances, J is a number of branches ($J = 2$ in the SLIQ/SPRINT algorithm), $P(C_{ml})$ is the probability of l class in the branch m, n_m is a number of the instances for the branch m, and N is the total number of the SNP instances (in the first case all instances are used, that is, $N = 142$).

2.2. Growth of the decision tree

A node is introduced for partitioning the instances in the decision tree, and in this article an SNP in the instances is used as a node. The processes for growing the decision tree are, in outline, as follows:

(1) The decision tree starts as a single node representing the HCV patient instances.
(2) If the instances are all of the same class, the node becomes a leaf and is labeled with that class.
(3) Otherwise, the attribute that will best separate the instances into individual classes is selected by calculating the GDI for each attribute. The node is labeled with this attribute.
(4) Two branches are created for the node attribute, and the instances are partitioned accordingly.
(5) The same processes are carried out recursively to form a decision tree for the instances at each partition.
(6) The recursive partitioning stops only when one of the following conditions is met:
   - all instances for a given node belong to the same class, or
   - there are no remaining attributes on which the instances may be further partitioned.

2.3. Pruning of the decision tree

Working backward from the bottom of the tree, the subtree starting at each non-terminal node is examined. If removing a subtree improves the error (misclassification) rate on the testing data, the subtree is removed. This process continues until no further improvement is made.

2.4. Prediction by decision tree learning

After completing the growth and pruning of the decision tree, majority voting is carried out for individual leaves. That is, the individual leaves are labeled with the most common classes (VR and NVR) in the instances. In addition, VR and NVR classes are assembled as the total VRs and NVRs for predicting VR and NVR ratios.

The VR ratio predicted from the decision tree learning based on the SNP information is given by

$$P_{VR} = \frac{D_{VR}}{I_{VR}} \tag{2}$$

where D_{VR} is the total number of VRs in the predicted VR class and I_{VR} is the number of VR instances.

The NVR ratio predicted from the decision tree learning is given by

$$P_{NVR} = \frac{D_{NVR}}{I_{NVR}} \tag{3}$$

where D_{NVR} is the total number of NVRs in the predicted NVR class and I_{NVR} is the number of NVR instances.

The total number of VRs and NVRs predicted from the decision tree learning is based on Eqs. (3) and (4) and is given by

$$P_{VR+NVR} = \frac{D_{VR} + D_{NVR}}{AI} \tag{4}$$

where *AI* is the total number of instances (all the HCV patients). That is, $AI = I_{VR} + I_{NVR}$.

3. Results and discussion

One hundred and forty-two Japanese patients with HCV (78 NVR and 64 VR) receiving PEG-IFN-α/RBV treatment were analyzed by using the SNPs evaluated in a previous GWAS study [7]. Although a total of 621,220 SNPs were used for the genome-wide association analysis, approximately 500 of those with the lowest *P*-values calculated by using a χ^2 test for allele frequencies were selected for the decision tree learning. Then GDIs were calculated for three types of branches of individual SNPs selected. The node with the largest GDI was selected as the root node of the decision tree. In the first case, GDI from rs8099917 was selected as the maximum branch node. That is, 142 Japanese HCV patients were divided into two branches based on rs8099917. The growth of the decision tree was carried out shown in Fig. 1, where one sees that the "yes" and "no" branches from the root node B1 were, respectively, Het + mm and MM. The numbers of instances in the two branches divided by rs8099917 were, respectively, 65 in the "yes" branch and 77 in the "no" branch. As the most instances in the "yes" branch node were NVRs (59 of 65, or 90.8%), this node was labeled as the leaf node. From the descendent node on the "no" branch, a new descendant of the node was obtained using the processes described in Section 2.2. After the decision tree was grown, it was pruned as described in Section 2.3. The final form of the decision tree diagram is shown in Fig. 1.

The decision tree learning was carried out for the 142 Japanese HCV patients with the 500 lowest-P SNPs, and the predicted VR and NVR ratios were calculated using Eqs. (2) and (3):

$$P_{VR} = \frac{48+7}{64} = \frac{55}{64} = 0.859$$

$$P_{NVR} = \frac{59+14}{78} = \frac{73}{78} = 0.936$$

The predicted total number of VRs and NVRs was also calculated using Eq. (4):

$$P_{VR+NVR} = \frac{55+73}{142} = 0.901$$

Before the decision tree learning by the SNP information, we knew only that 64 of the Japanese HCV patients were VR for the drug response and 78 were NVR. Therefore, all we can calculate is that the VR and NVR response rates for those patients are respectively 0.451 (64/142) and 0.549 (78/142). After the decision tree learning, however, we can predict that the VR and NVR drug responses are respectively 0.859 and 0.936 and can predict that the drug response rate for the total 142 samples is 0.901 ((55 + 73)/142). This means that if a new HCV genotype 1 patient wants to know whether or not he/she is VR/NVR for the drug treatment, he/she can predict his/her VR/NVR ratio by checking his/her SNP information in the decision tree shown in Fig. 1.

For example, suppose a new male HCV patient having rs8099917, rs4906195, and rs3816768 alleles that are, respectively, "MM", "Het", and "mm". In this case, it is predicted that he will be VR for the recommended HCV drug treatment. On the other hand, suppose a new female HCV patient having rs8099917 with "mm". In this case, it is predicted that she will mostly likely be NVR for the recommended HCV drug treatment.

It is important to increase the prediction ratio in the decision tree learning. The higher the prediction ratio, the better the drug treatment response for a new HCV patient is predicted. In the decision tree model, the root node plays an important role in increasing the prediction ratio. We therefore also generated decision tree models by using root nodes based on the SNPs having the 30 highest GDIs. As a result, we got two decision trees that can predict ratios higher than that of the first decision tree. They are shown in Figs. 2 and 3. The results of prediction ratios (probabilities predicted) for these decision tree models and the individual SNP attributes used in the models are listed in Table 2. We got a 92.9%

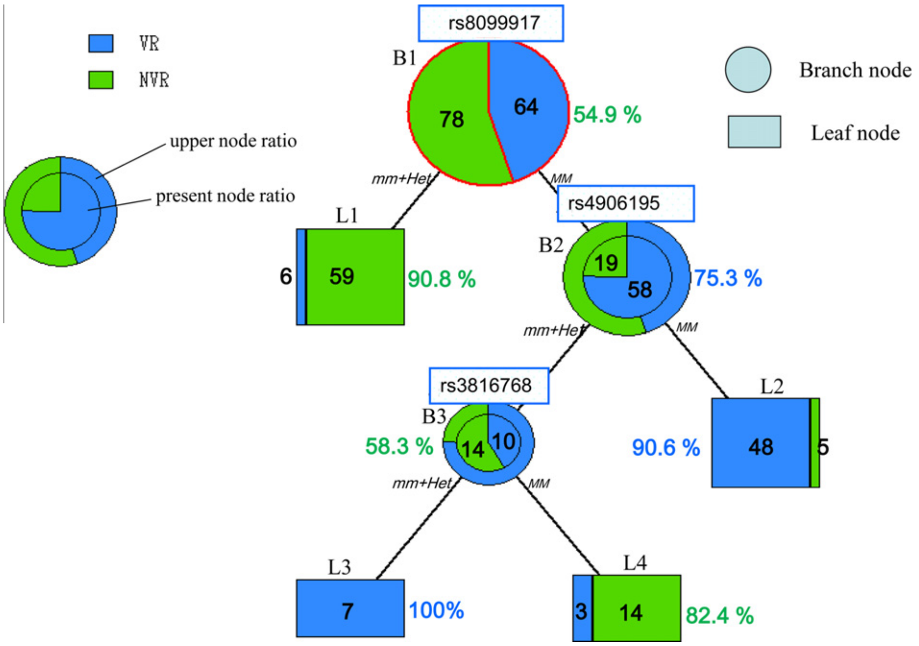

Fig. 1. Decision tree diagram for the treatment of Japanese HCV patients (78 NVR and 64 VR samples). The box at the top of a branch node indicates the SNP ID (e.g., rs8099917). The bottom of the branch node shows the branch condition: left for "yes" and right for "no". In the case of rs8099917, for example, "mm + Het" is "yes" and "MM" is "no". B1, B2, and B3 are branch nodes and B1 is the root node. The numbers in the green parts of the nodes are the numbers of NVR samples, and those in the blue parts are the numbers of VR samples. The percentage next to each node shows the percentage of the greater number of samples. In B1, for example, 78 is greater than 64 and 78/(78 + 64) is 54.9%. The two circles in B2 and B3 indicate two ratios; the external one is the ratio of the upper branch node and the internal one is the ratio of the present branch node. In B2, for example, the external ratio shows that of B1 and the internal ratio indicates that of node B2, i.e., 75.3% (58/(19 + 58)). L1, L2, L3, and L4 are leaf nodes.

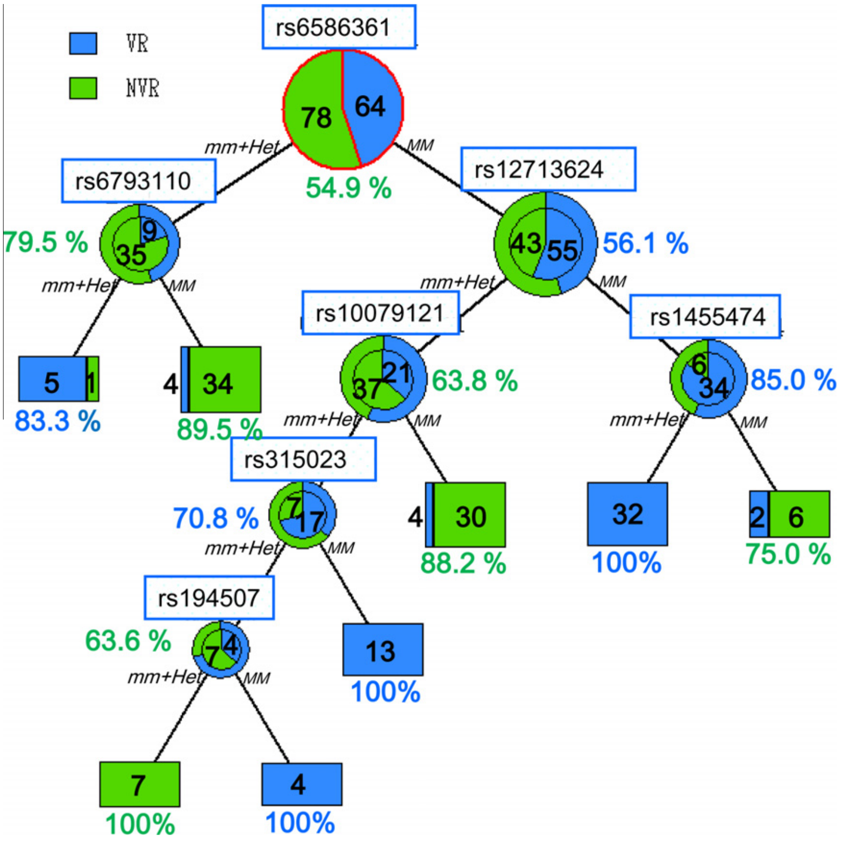

Fig. 2. Diagram of model 2 decision tree. See legend to Fig. 1 for an explanation of symbols used.

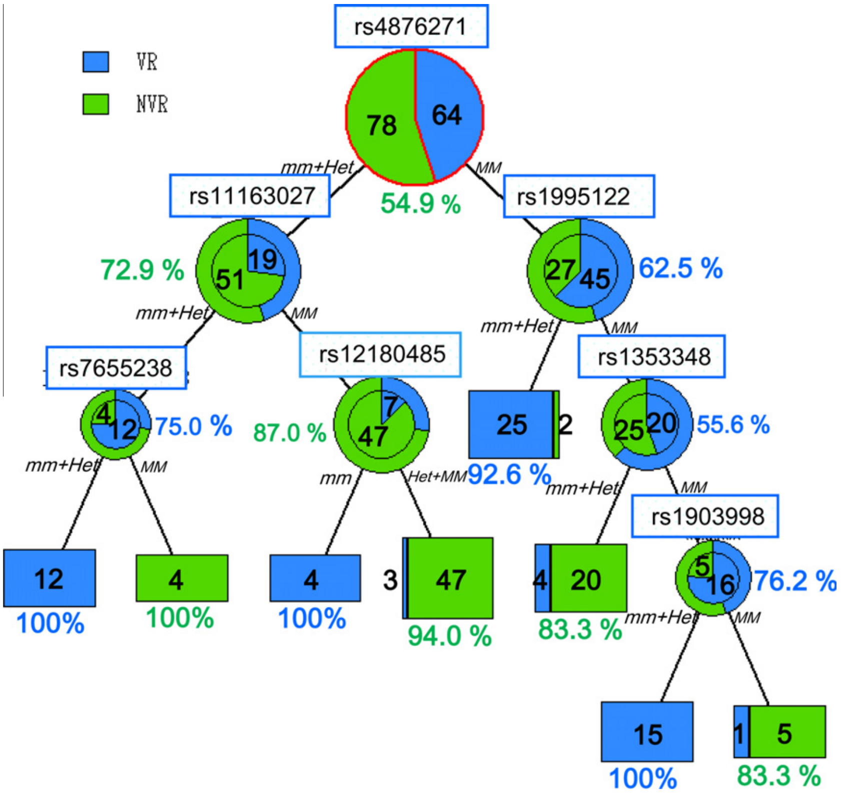

Fig. 3. Diagram of model 3 decision tree. See legend to Fig. 1 for an explanation of symbols used.

Table 2
Attribute SNPs in the three decision trees.

Model	SNP used					Probability predicted (%)
	SNP	*P*-value	OR	Chromosome	Allele ratio (%)	
1	rs8099917	3.11×10^{-15}	30	19	23.9	90.1
	rs4906195	4.52×10^{-4}	3.9	14	18	
	rs3816768	2.85×10^{-4}	4	15	16.4	
2	rs6586361	7.81×10^{-5}	5	1	17.3	92.3
	rs6793110	2.82×10^{-4}	10.7	3	5.7	
	rs12713624	1.56×10^{-4}	3.8	2	33.1	
	rs10079121	1.11×10^{-3}	3.1	5	25.7	
	rs1455474	2.96×10^{-4}	4.6	8	47.9	
	rs315023	2.67×10^{-3}	2.8	1	29.6	
	rs194507	3.19×10^{-4}	3.5	7	34.4	
3	rs4876271	2.3×10^{-5}	4.5	8	27.8	92.9
	rs11163027	1.51×10^{-4}	4.7	1	12.7	
	rs1995122	1.98×10^{-4}	4.1	6	14.8	
	rs7655238	3.21×10^{-3}	3.4	4	46.1	
	rs12180485	5.16×10^{-3}	2.8	6	17.6	
	rs1353348	1.36×10^{-4}	3.9	15	35.1	
	rs1903998	2.58×10^{-3}	3.1	10	40.5	

P-value: χ^2 test for allele frequencies, OR: Odds ratio.

prediction probability with model 3. This is 2.8 percentage points higher than that of model 1.

As the root node plays an important role, we examined what chromosome the root node belongs to in each model. The chromosomes with individual root nodes in the models 1, 2, and 3 were, respectively, 19, 1, and 8 (Table 2). Although SNPs near the IL28B gene on chromosome 19 were recently reported to be the most significant factors, the root nodes in models 2 and 3 are on other chromosomes. Other SNPs used in the models 2 and 3 are also on other chromosomes (Table 2). These results therefore imply that there may be significant SNPs other than those near the IL28B gene on chromosome 19.

Although that the virological responses of HCV patients treated with PEG-IFN-α and RBV have been reported to be strongly associated with genetic polymorphisms, there is no report of what responses can be predicted from the polymorphisms [14]. If the responsiveness of HCV patients to a drug treatment could be predicted from information about related SNPs, side effects and treatment cost could be greatly reduced. Furthermore, because the results of the proposed method implied that SNPs other than those in the IL28B region are strongly related to the prediction of the drug response, the relations between those SNPs and the drug response should be examined experimentally.

Acknowledgment

We thank Mr. T. Yamauchi for his technical support in the data mining.

References

[1] Ray Kim, W. (2002) Global epidemiology and burden of hepatitis C. Microbes Infect. 4, 1219–1225.

[2] Lavanchy, D. (2009) The global burden of hepatitis. Liver Int. 29, 74–81.

[3] Manns, M.P., McHutchison, J.G., Gordon, S.C., Rustgi, V.K., Shiffman, M., Reindollar, R., Goodman, Z.D., Koury, K., Ling, M. and Albrecht, J.K. (2001) Peginterferon alfa-2b plus ribavirin compared with interferon alfa-2b plus ribavirin for initial treatment of chronic hepatitis C: a randomised trial. Lancet 358, 958–965.

[4] Fried, M.W., Shiffman, M.L., Reddy, K.R., Smith, C., Marinos, G., Gonçales Jr., F.L., Häussinger, D., Diago, M., Carosi, G., Dhumeaux, D., Craxi, A., Lin, A., Hoffman, J. and Yu, J. (2002) Peginterferon alfa-2a plus ribavirin for chronic hepatitis C virus infection. N. Engl. J. Med. 347, 975–982.

[5] Hadziyannis, S.J., Sette Jr., H., Morgan, T.R., Balan, V., Diago, M., Marcellin, P., Ramadori, G., Bodenheimer Jr., H., Bernstein, D., Rizzetto, M., Zeuzem, S., Pockros, P.J., Lin, A., Ackrill, A.M. and PEGASYS International Study Group (2004) Peginterferon-alpha2a and ribavirin combination therapy in chronic hepatitis C: a randomized study of treatment duration and ribavirin dose. Ann. Intern. Med. 140, 346–355.

[6] Fried, M.W. (2002) Side effects of therapy of hepatitis C and their management. Hepatology 36, S237–S244.

[7] Tanaka, Y., Nishida, N., Sugiyama, M., Kurosaki, M., Matsuura, K., Sakamoto, N., Nakagawa, M., Korenaga, M., Hino, K., Hige, S., Ito, Y., Mita, E., Tanaka, E., Mochida, S., Murawaki, Y., Honda, M., Sakai, A., Hiasa, Y., Nishiguchi, S., Koike, A., Sakaida, I., Imamura, M., Ito, K., Yano, K., Masaki, N., Sugauchi, F., Izumi, N., Tokunaga, K. and Mizokami, M. (2009) Genome-wide association of IL28B with response pegylated interferon-α and ribavirin therapy for chronic hepatitis C. Nat. Genet. 41, 1105–1109.

[8] Ge, D., Fellay, J., Thompson, A.J., Simon, J.S., Shianna, K.V., Urban, T.J., Heinzen, E.L., Qiu, P., Bertelsen, A.H., Muir, A.J., Sulkowski, M., McHutchison, J.G. and Goldstein, D.B. (2009) Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance. Nature 461, 399–401.

[9] Thomas, D.L., Thio, C.L., Martin, M.P., Qi, Y., Ge, D., O'Huigin, C., Kidd, J., Kidd, K., Khakoo, S.I., Alexander, G., Goedert, J.J., Kirk, G.D., Donfield, S.M., Rosen, H.R., Tobler, L.H., Busch, M.P., McHutchison, J.G., Goldstein, D.B. and Carrington, M. (2009) Genetic variation in IL28L and spontaneous clearance of hepatitis C virus. Nature 461, 798–801.

[10] Fellay, J., Thompson, A.J., Ge, D., Gumbs, C.E., Urban, T.J., Shianna, K.V., Little, L.D., Qiu, P., Bertelsen, A.H., Watson, M., Warner, A., Muir, A.J., Brass, C., Albrecht, J., Sulkowski, M., McHutchison, J.G. and Goldstein, D.B. (2010) ITPA gene variants protect against anaemia in patients treated for chronic hepatitis C. Nature 464, 405–408.

[11] Suppiah, V., Moldovan, M., Ahlenstiel, G., Berg, T., Weltman, M., Abate, M.L., Bassendine, M., Spengler, U., Dore, G.J., Powell, E., Riordan, S., Sheridan, D., Smedile, A., Fragomeli, V., Müller, T., Bahlo, M., Stewart, G.J., Booth, D.R. and George, J. (2009) IL28B is associated with response to chronic hepatitis C interferon-alpha and ribavirin therapy. Nat. Genet. 41, 1100–1104.

[12] Quinlan, J.R. (1986) Induction of decision trees. Machine Learning 1, 81–106.

[13] Shafer, J., Agrawal, M. and Mehta, M. (1996) A scalable parallel classifier for data mining in: Proc. of 22nd VLDB Conference, pp. 544–555.

[14] Lin, C.Y., Chen, J.Y., Lin, T.N., Jeng, W.J., Huang, C.H., Huang, C.W., Chang, S.W. and Sheen, I.S. (2011) IL2B SNP rs12979860 is a critical predictor for on-treatment and sustained virologic response in patients with hepatitis C virus genotype-1 infection. PLoS ONE 6, e18322.