Bioorganic & Medicinal Chemistry 19 (2011) 6892–6905

Contents lists available at SciVerse ScienceDirect

Bioorganic & Medicinal Chemistry

ELSEVIER

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

A new method for induced fit docking (GENIUS) and its application to virtual screening of novel HCV NS3-4A protease inhibitors

Daisuke Takaya[a], Atsuya Yamashita[b], Kazue Kamijo[a], Junko Gomi[a], Masahiko Ito[b], Shinya Maekawa[c], Nobuyuki Enomoto[c], Naoya Sakamoto[d,e], Yoshiaki Watanabe[f], Ryoichi Arai[f], Hideaki Umeyama[f], Teruki Honma[a], Takehisa Matsumoto[a], Shigeyuki Yokoyama[a,g,*]

[a] *RIKEN Systems and Structural Biology Center, 1-7-22 Suehiro-cho, Tsurumi, Yokohama 230-0045, Japan*

[b] *Department of Microbiology, Division of Medicine, Interdisciplinary Graduate School of Medicine and Engineering, University of Yamanashi, 1110 Shimokato, Chuo, Yamanashi 409-3898, Japan*

[c] *First Department of Internal Medicine, Faculty of Medicine, University of Yamanashi, 1110 Shimokato, Chuo, Yamanashi 409-3898, Japan*

[d] *Department of Gastroenterology and Hepatology, Tokyo Medical and Dental University, 1-5-45 Yushima, Bunkyo-ku, Tokyo 113-8519, Japan*

[e] *Department for Hepatitis Control, Tokyo Medical and Dental University, 1-5-45 Yushima, Bunkyo-ku, Tokyo 113-8519, Japan*

[f] *School of Pharmacy, Kitasato University 5-9-1 Shirokane, Minato-ku, Tokyo 108-8641, Japan*

[g] *Department of Biophysics and Biochemistry, Graduate School of Science, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan*

ARTICLE INFO

Article history:
Received 12 July 2011
Revised 12 September 2011
Accepted 12 September 2011
Available online 16 September 2011

Keywords:
HCV
NS 3-4A protease
Structure based drug design
Ligand docking
Virtual screening

ABSTRACT

Hepatitis C virus (HCV) is an etiologic agent of chronic liver disease, and approximately 170 million people worldwide are infected with the virus. HCV NS3-4A serine protease is essential for the replication of this virus, and thus has been investigated as an attractive target for anti-HCV drugs. In this study, we developed our new induced-fit docking program (GENIUS), and applied it to the discovery of a new class of NS3-4A protease inhibitors (IC_{50} = 1–10 μM including high selectivity index). The new inhibitors thus identified were modified, based on the docking models, and revealed preliminary structure–activity relationships. Moreover, the GENIUS in silico screening performance was validated by using an enrichment factor. We believe our designed scaffold could contribute to the improvement of HCV chemotherapy.

1. Introduction

Hepatitis C virus (HCV) is an etiologic agent of chronic liver disease,[1,2] and approximately 170 million people worldwide are infected with the virus.[3] Chronic hepatitis C can lead to severe liver diseases, including fibrosis, cirrhosis, and hepatocellular carcinoma.[4] The current standard therapy for chronic hepatitis C consists of pegylated interferon in combination with ribavirin.[5] Unfortunately, this therapy results in sustained antiviral activity in only about 50–60% of the patients, and is associated with serious side effects. Thus, the development of alternative and more effective anti-HCV agents has been eagerly anticipated.

HCV NS3-4A serine protease is essential for the replication of this virus, and has been investigated as an attractive target for anti-HCV drugs. Several three-dimensional structures of HCV NS3-4A protease have been deposited in the Protein Data Bank (PDB).[6] Therefore, Structure Based Drug Design (SBDD) is a promising approach for the discovery of new NS3-4A protease inhibitors. The NS3-4A protease has the catalytic triad with the anion hole, commonly found among serine protease family members. The NS3-4A protease consists of two domains: a protease domain of 180 residues and a helicase domain of 420 residues.[7] The protease domain contains the protease activity, and thus it is appropriate to use only this domain as the receptor coordinates for SBDD.[8] On the other hand, docking calculations to a complex with a helicase domain have also been performed.[9] Different receptor structures were used in the docking calculations, because no experimentally determined full-length NS3-4A protease structures complexed with small molecule inhibitors were available, as of 2011.

In recent years, many peptide or peptide-mimic inhibitors that inhibit HCV NS3-4A protease have been developed, including SCH-503034,[10] VX-950,[11] BILN-2061,[12] TMC-435,[13] ITMN-191[14] and MK-7009,[15] as specifically targeted anti-viral agents for HCV (STAT-C).[16] These compounds, which competitively inhibit the protease activity, were roughly classified into two types: the mimic type inhibitors (SCH-503034, VX-950), which have a peptide bond, and the macrocyclic compounds (BILN-2061, TMC-435350, ITMN-191, MK-7009), which have a macrocyclic ring. Recently,

* Corresponding author.
E-mail address: yokoyama@biochem.s.u-tokyo.ac.jp (S. Yokoyama).

0968-0896/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2011.09.023

ACH-806[17] was reported as an HCV NS3-4A non-peptide inhibitor, and it works in harmony with the NS3-4A protease inhibitor or the NS5B polymerase. Clinical trials (Phase III) of SCH503034 and VX-950 have been performed.[18] However, cardiotoxicity in monkeys was reported for BILN-2061, one of the macrocyclic compounds, and thus its clinical development has been interrupted.[19] Moreover, macrocyclic compounds also have a problematic ADME profile, mainly due to their large molecular weights, and the synthetic optimization of the inhibitors is difficult. In addition, various mutations, especially A156T in the active site,[20] confer resistance to these protease inhibitors, such as SCH503034, BILN-2061 and VX-950[21] Since drug-resistant viruses have readily appeared in monotherapy, a multiple drug regimen has been widely applied for anti-HCV therapy. Therefore, good ADME properties are important for the next generation of HCV NS3-4A inhibitors.[18] Generally, since peptide inhibitors lack chemical stability in relation to racemization, peptide compounds are not being pursued in the development of more effective anti-HCV drugs. Thus, a new class of non-peptide inhibitors is still expected, and an inhibitor of this protease, designed by SBDD, would be valuable for anti-HCV chemotherapy. For example, in recent years, Ismail and Hattori designed a new inhibitor with an indole skelton by a molecular modeling approach,[22] based on the structure complexed with an inhibitor bearing an indole skeleton (PDB code: 1W3C) reported by Ontoria et al.[23]

From a three-dimensional point of view, many HCV NS3-4A protease structures have been reported. In the PDB, 53 BLAST hits (*E*-value <10.0) on a query sequence obtained from the NS3-4A protease (PDB code 1DXW.A[24]) were found, as of January 2011. Almost all of the structures were determined by X-ray analyses. For example, Cummings et al. determined the complex structure of TMC-435 with the protease, and reported that the protease inhibitor interacts with the protease domain by forming non-covalent bonds (PDB code 3KEE).[25] Moreover, Hangel et al. reported the structure complexed with an inhibitor that interacts with the non-catalytic cysteine of the protease.[26] However, the structures of some HCV NS3-4A proteases have also been determined by NMR analyses. Among the BLAST hits, 3 structures determined by NMR were found. Barbato et al. reported two structures (PDB codes 1BT7,[27] 1DXW), and recently, Gallo et al. reported that of the NS3 protease, in the absence of the NS-4A co-factor, complexed with a non-covalent inhibitor (PDB code 2K1Q[28]).

Many programs are available to predict the binding modes of small molecules. Docking programs, such as AUTODOCK,[29] DOCK[30] and GOLD,[31] dock a ligand by changing their conformations to a fixed coordinate receptor and evaluating the fit by various experiential energy functions (i.e., Flexible Ligand Docking). These docking programs are useful for relatively non-flexible proteins; however, the conformations of many proteins are changed by different ligand molecules (induced fit). In such cases, conventional flexible ligand docking is not suitable for the prediction of the binding mode. To solve this induced fit problem, there are many docking programs and protocols in which the dock changes not only the conformations of the ligand but also the coordinates of the receptor, to consider the flexibility of the receptor (Flexible *Receptor* Docking or Induced-Fit Docking).

The induced-fit ligand docking methods are mainly classified into two groups,[32] soft-docking and ensemble docking. In soft-docking, the flexibility of a receptor is considered by changing the repulsion term of the protein ligand interaction in scoring functions, such as the Lennard-Jones potential term. In ensemble docking, one ligand is docked to multiple receptor conformation groups. For example, the soft-docking program Glide[33] enables scaling of the VDW radii, to relax the repulsion of the protein–ligand atoms. As an ensemble docking method, RosettaLigand considers the induced fit of the side chain, using a Backbone-dependent Rotamer Library.[34,35] Moreover, to release the volume occupied by the side chain, Glide performs an alanine substitution of the side chain in contact with the docking ligand. The open space is used for the binding pocket in the first docking, for predicting tentative binding modes. After the ligand is docked, the removed side chain is reconstructed by homology modeling using Prime, and the ligands are re-docked into the constructed protein models. These induced fit docking programs make it possible to predict interactions in difficult predictions, by only using the coordinates of one fixed receptor structure.

In this study, we developed our new induced-fit docking program (GENIUS), and applied it to the discovery of a new class of HCV NS3-4A protease inhibitors. In our program, the induced fit of protein side chains was considered by incorporating the dynamic information in solution. Among the available experimental coordinates of the NS3-4A protease, the NMR structure (PDB code 1DXW) was chosen as the receptor ensemble for docking. The collision tolerance was set for each atom of the receptor, based on the degree of preservation of the side chain torsion angle in the ensemble. Moreover, Essential Interaction Pairs (EIPs) were newly defined to interact with not only the active site but also the hydrophobic atoms on the planar beta sheet of the protease, as a constraint for ligands. The GENIUS docking system enables induced fit docking (Fig. 1), and combines ensemble docking to use the conformation cluster of the receptors, and soft-docking to set the coefficient for every atom of the receptor and to relax the collisions between protein and ligand atoms. The GENIUS docking system using EIPs was employed for the in silico screening of the NS3-4A protease inhibitors, and the selected compounds were evaluated by HCV NS3-4A enzymatic and cell-based assays. The new inhibitors thus discovered were modified based on the docking models, and revealed their preliminary structure–activity relationships.

2. Result and discussion

This study was performed by combining in silico and in vitro screening techniques. For the in silico screening part, we developed

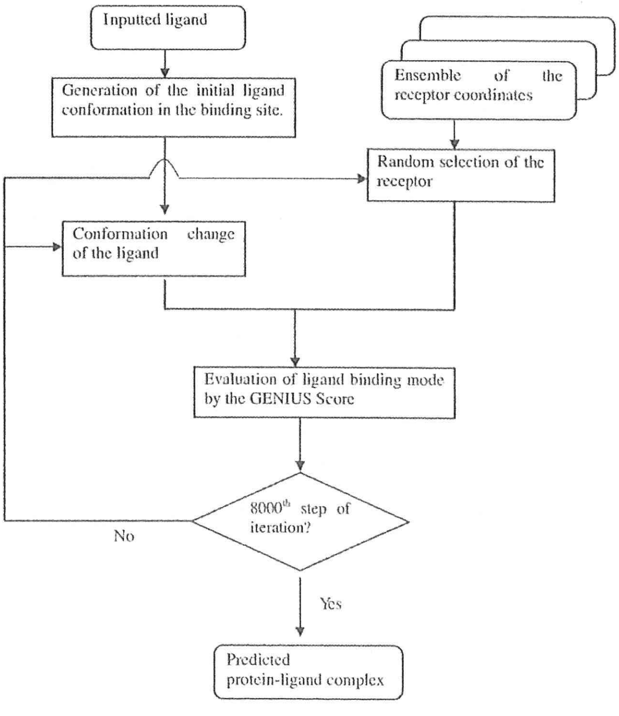
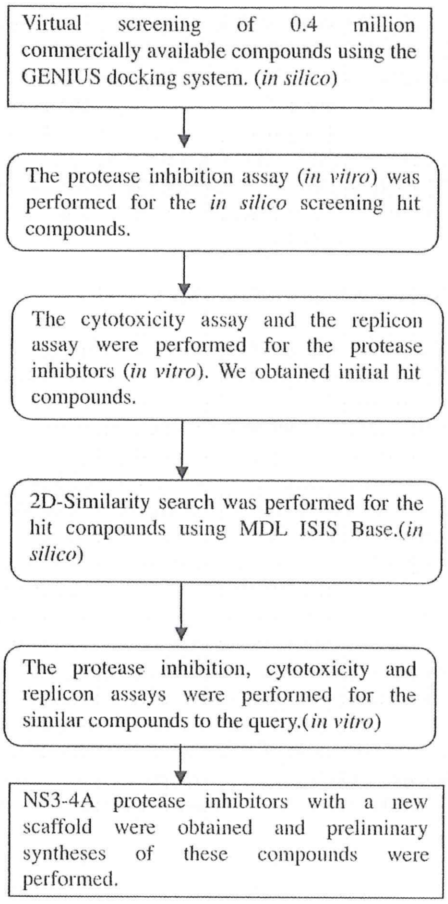

Inputted ligand
Generation of the initial ligand conformation in the binding site.
Ensemble of the receptor coordinates
Random selection of the receptor
Conformation change of the ligand
Evaluation of ligand binding mode by the GENIUS Score
8000th step of iteration?
No
Yes
Predicted protein-ligand complex

Figure 1. Flowchart of the GENIUS docking system.



Virtual screening of 0.4 million commercially available compounds using the GENIUS docking system. (*in silico*)

The protease inhibition assay (*in vitro*) was performed for the *in silico* screening hit compounds.

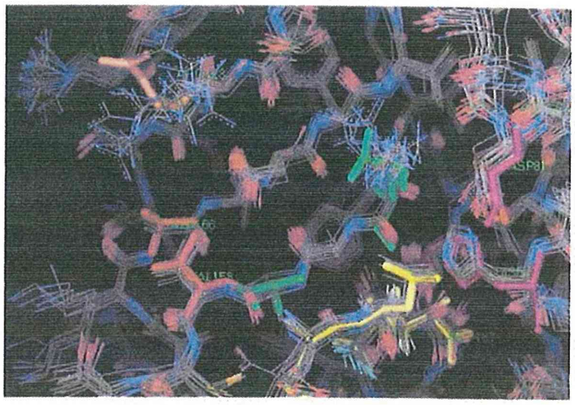
The cytotoxicity assay and the replicon assay were performed for the protease inhibitors (*in vitro*). We obtained initial hit compounds.

2D-Similarity search was performed for the hit compounds using MDL ISIS Base.(*in silico*)

The protease inhibition, cytotoxicity and replicon assays were performed for the similar compounds to the query.(*in vitro*)

NS3-4A protease inhibitors with a new scaffold were obtained and preliminary syntheses of these compounds were performed.

Figure 2. Flowchart of HCV NS3-4A in silico and in vitro.



(a): 20 NMR structures (PDB code: 1DXW)

```
# anion hole, anchor zone
KEYATM O.3   100 2.58 NE2 HISA_57
KEYATM O.co2 100 2.60 N   GLYA_137

# hydrogen-bond interactions
KEYATM DONOR 100 3.40  O  ARGA_155
KEYATM ACPTR 100 2.60  N  ALAA_157
KEYATM DONOR 100 2.60  O  ALAA_157

# hydrophobic interactions on the beta-sheet
KEYATM C.3   100 2.60  CB  ALAA_166
KEYATM C.3   100 3.80  CB  VALA_158
```

(b) The indicated interaction points on the NS3-4A protease.

Figure 3. (a) The line representations are the ensemble of 20 NMR structures of the NS3-4A protease domain. All hydrogen atoms were removed. The fraction residue on the beta sheet, Arg 123, is shown as an orange stick representation. The catalytic triad residues, His57, Asp81, and Ser139, are shown as a pink stick representation. The hydrophobic residues on the beta sheet, Val158 and Ala166, are shown as a red stick representation. The residue involved in the anion hole formation, Gly 137, is shown as a blue stick representation. The residues involved in the hydrogen-bond interaction, Arg155 and Ala157, are colored green on the beta sheet. Val158 and Ala166 are shown as red stick representations. The inhibitor, 3-amino-5,5-di-fluoro- 2-keto-pentan-1-oic acid, which forms a covalent bond with Ser139, is shown as a yellow stick representation. (b) The line that begins with"KEYATM" means one of the EIPs. The second column string, such as 0.3, O.co2, means the designated atom type that the docking ligand must have in the docking calculation. The third column means the constraint value for the EIP term in the GENIUS scoring function. The fourth column means the equivalent distance of pairwise atoms between receptor and ligand. The 5-th and 6-th columns mean the atom type and the amino acid involved in the protein–ligand interaction on the receptor, respectively.

a new method for induced fit docking, called the GENIUS docking system (Fig. 1 and see details in the Section 4), and the system was utilized for HCV NS3-4A protease in silico screening, based on NMR structural ensembles. Subsequently, the EIP for HCV NS3-4A protease inhibitors was set. For the in vitro assays, the protease inhibition activity and efficacy in HCV infected cells (i.e., the replicon) were assessed for the compounds selected in the in silico screen. Finally, preliminary syntheses to analyze the structure–activity relationships for the effective compounds in the in-vitro assays were performed (Fig. 2, see details in the Section 4).

2.1. Setting of ligand binding site and EIP for in silico screening

Since this research commenced before the structure complexed with a non-covalent inhibitor was reported (PDB code 2K1Q), 1DXW.A was used for the docking receptor. In the GENIUS docking system, the definition of a binding site was required, as in other docking algorithms. The binding site for docking was defined at 16.0 Å around every atom of the ligand (3-amino-5 and 5-di-fluoro-2-keto-pentan-1-oic acid) contained in the coordinates. The ensemble of receptor coordinates was clustered, in order to analyze the induced fit of the receptor. The atoms conserved in the average torsion angle range from −18 to 18 degrees in 99% of the population were collected from the binding site, and were ignored in the calculation of the collision term (Table S1 Supplementary data).

Next, the EIP used in this study was set up for the docking conditions of GENIUS. The receptor conformation group revealed that the active site residues displayed minimal fluctuations between the NMR structures. On the other hand, for Arg123 on the β sheet, the fluctuation between each coordinate was large (Fig. 3a). The hydrophobic residues (Val158, Ala166) on the β sheet are exposed, as a result of the motion of Arg123. The EIP was then prepared, by reference to the interactions generated as a result of the dynamics (Fig. 3b).

The final EIP is described below. Since the HCV NS3-4A protease is a serine protease, it contains the catalytic triad and the oxyanion hole that cleave the peptide bond of the substrate, as in other serine proteases. In order to obtain the interaction with the oxyanion hole, Gly137 was assigned to the EIP setting. Furthermore, we set an EIP with a hydrophobic interaction between the atoms on the β sheet (Val158, Ala166) and the atoms of the ligand (Fig. 3b). This EIP was used for the docking conditions.

2.2. Docking by GENIUS docking system

In silico screening by GENIUS, using the obtained EIP, was performed for 166,206 compounds. Based on their GENIUS docking scores, 42,504 compounds were ranked, because compounds lacking the atoms specified in the EIP could not be docked. The ranked compounds were also verified by visual inspection from top to bottom, because the EIP term is not always valid for all docking compounds. Finally, 97 compounds were selected, based on their high scores in the docking calculation, as meeting the criteria specified in the obtained EIP and the visual inspection.

2.3. In vitro evaluation of the selected compounds

Among the 97 compounds, 27 compounds showed more than 50% protease inhibition activity at 100 μM. In addition, compounds CP3-0032 (**1**) and CP3-0084 (**2**) (Fig. 4) exhibited HCV growth inhibitory activity at 13 and 23 μM in the replicon assay, respectively, and lacked toxicity (CC_{50}(MTS) >125 μM). (Table 1) Compounds **1** and **2** have a common skeleton, featuring an acyl diazene (–N=N–) and a biarylester (Fig. 4). To clarify the structure–activity relationship of this chemical series, similar compounds were selected from commercially available compounds. In total, 140 compounds were selected as derivatives with the common substructure and a similar skeleton by a 2D-similarity search, and the protease inhibition assay was performed. Among the similar compounds, eight compounds (**3**–**10**) exhibited protease inhibition activities ranging from 1.01 to 64.3 μM of the IC_{50} values. The IC_{50}, EC_{50}, CC_{50} and selectivity index values for these compounds are summarized in Table 1. Among these compounds, CP3-3284-125 (**3**) and CP3-3284-126 (**4**) exhibited strong inhibition of the protease activity at IC_{50} = 1.06 and 1.01 μM, respectively. Moreover, in the replicon assay, their EC_{50} values were 19.5 and 12.5 μM, respectively (Table 1). However, these compounds showed relatively strong toxicity in the ATP assays. In contrast, CP3-3284-53 (**10**) exhibited moderate protease inhibition activity (IC_{50} = 8.59 μM), as compared with compounds **3** and **4**; however, in the cell-based assays, the EC_{50} was 12.0 μM with a high selectivity index (>9.3).

2.4. Synthesis of compounds 10, 11, 12 and 13

Since the purity of compound **10** was unknown (we assumed 100% purity in the in vitro assay), compound **10** was synthesized (Scheme 1 and see details in the Section 4). In addition, compounds **11**, **12** and **13** were synthesized, and a preliminary synthetic modification was performed for compound **10**, based on the predicted binding mode. First, to enhance the hydrophobic interaction between these compounds and the receptor, a methyl (compound **11**) or ethyl (compound **12**) group was introduced to the central benzene ring. Moreover, this compound contained multiple nitro groups (Fig. 4). Next, the effect of introducing a nitro group to compound **10** was examined (compound **13**). However, the inhibition activity was not significantly different (Table 1). Generally, since a nitro group is disadvantageous from the viewpoint of solubility, this functional group is removed or converted to an amino group, which can form a hydrogen bond to the receptor atoms.

In summary, compound **1**, compound **2** and the CP3-3284 series (**3**–**10**) obtained in this research represent a new, unique class of non-peptide HCV NS3-4A inhibitors, because no similar HCV

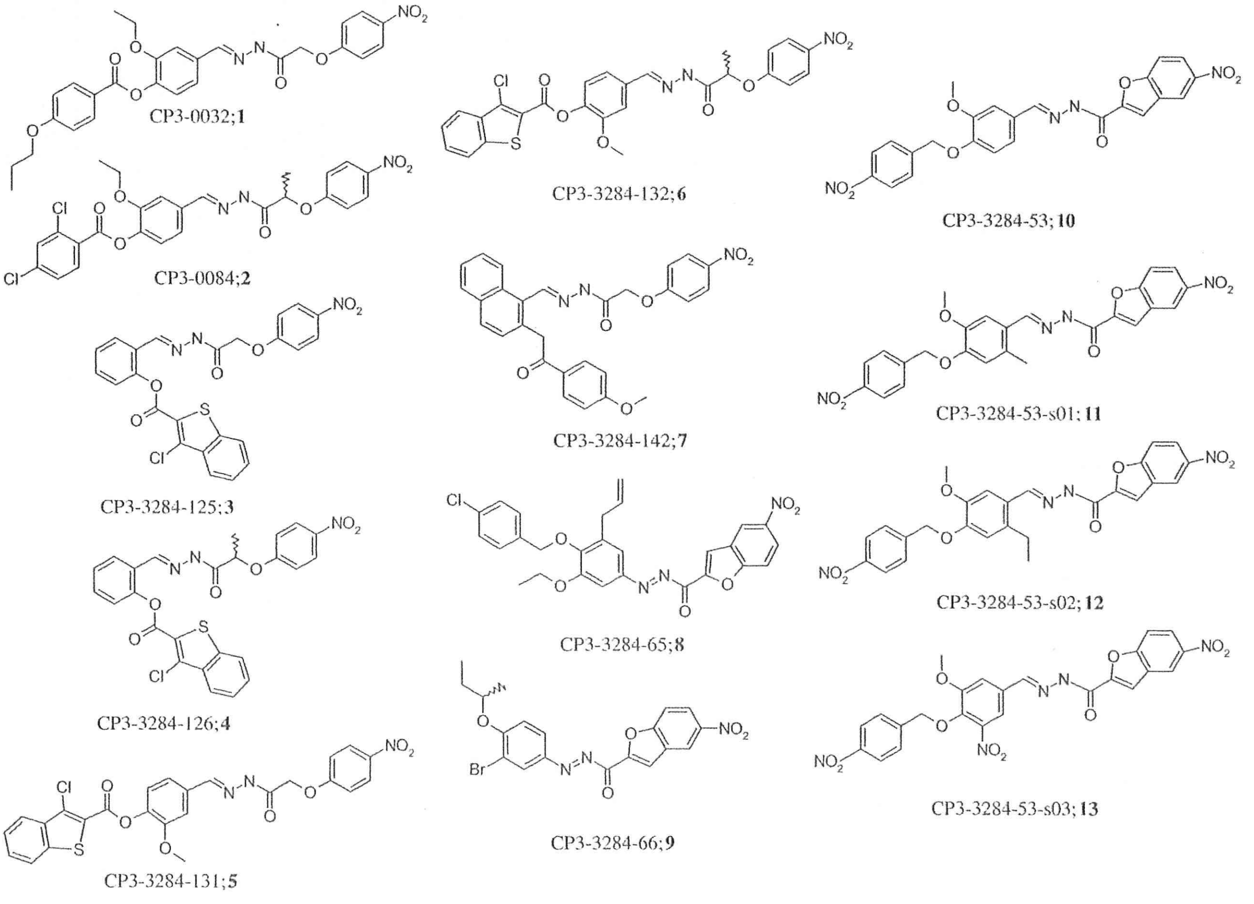
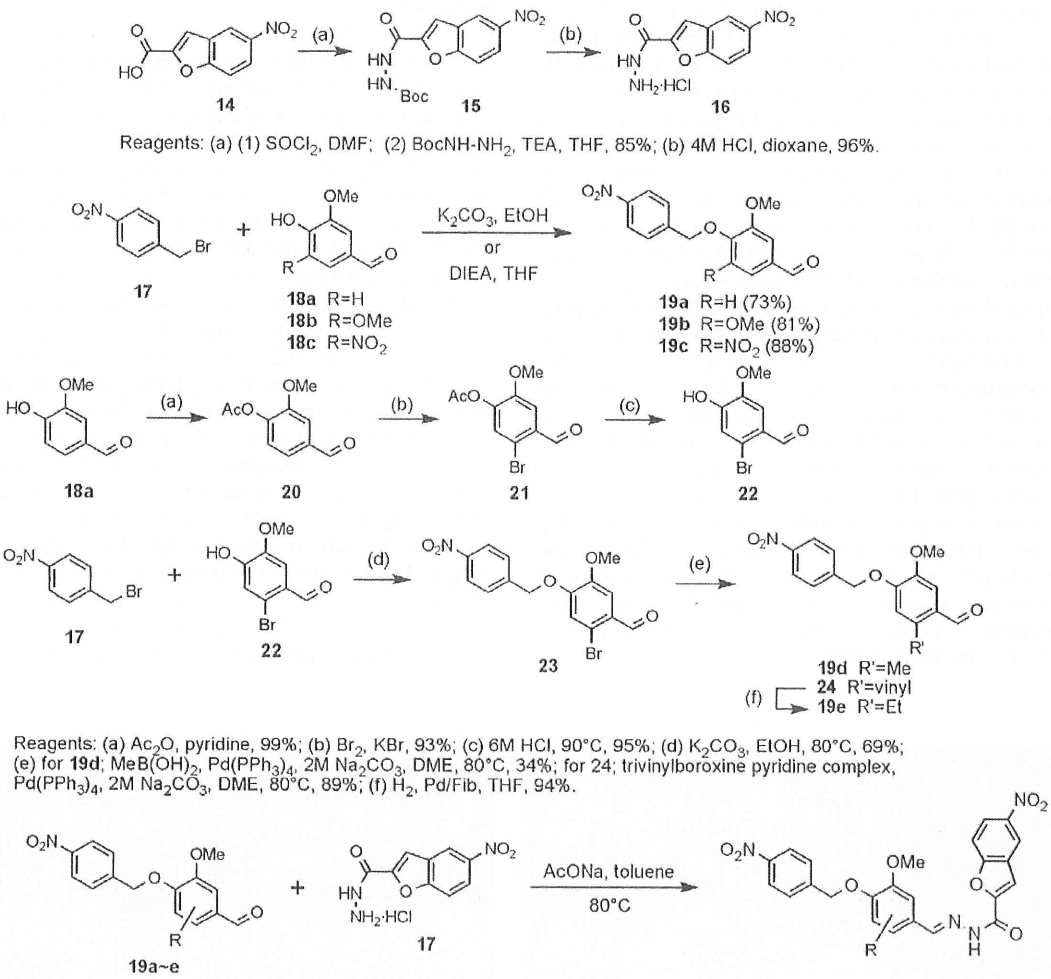


Figure 4. 2D Structures of the discovered protease inhibitors.

Table 1
In-vitro assay data of the discovered protease inhibitors

ID; serial number	Inhibition at 100 μM(%)	IC_{50}	EC_{50}	CC_{50} (MTS)	CC_{50} (ATP)	SI[a]	*A*Log*P*[b]
CP3-0032;**1**	38		13	>125		>9.6	5.63
CP3-0084;**2**	42.9		23	>125		>5.4	6.58
CP3-3284-125;**3**		1.06	19.5	>125	40	2.1	6.25
CP3-3284-126;**4**		1.01	12.5	>125	19	1.5	6.74
CP3-3284-131;**5**		12.3	93	>125			6.24
CP3-3284-132;**6**		4.08	121	>125			6.72
CP3-3284-142;**7**		64.3	8.5	>125	9	1.1	5.34
CP3-3284-65;**8**		8.07	>125	>125			8.72
CP3-3284-66;**9**		22.7	13.5	57	36	2.6	7.13
CP3-3284-53;**10**		8.59	12	>125	>80	>9.3	5.80
CP3-3284-53-s01;**11**		17.1					6.29
CP3-3284-53-s02;**12**		11.9					6.74
CP3-3284-53-s03;**13**		8.34					6.29

[a] The selectivity index (SI) is the ratio of the smaller CC50 value (either CC50(MTS) or CC50(ATP)) to the EC50 value.
[b] ALogP was calculated by PipeLinePilot 8.0.1(Accelrys Software Inc.).



Reagents: (a) (1) $SOCl_2$, DMF; (2) BocNH-NH_2, TEA, THF, 85%; (b) 4M HCl, dioxane, 96%.

Reagents: (a) Ac_2O, pyridine, 99%; (b) Br_2, KBr, 93%; (c) 6M HCl, 90°C, 95%; (d) K_2CO_3, EtOH, 80°C, 69%; (e) for **19d**; $MeB(OH)_2$, $Pd(PPh_3)_4$, 2M Na_2CO_3, DME, 80°C, 34%; for 24; trivinylboroxine pyridine complex, $Pd(PPh_3)_4$, 2M Na_2CO_3, DME, 80°C, 89%; (f) H_2, Pd/Fib, THF, 94%.

Scheme 1. Synthetic routes of Compounds **10–15**.

NS3-4A inhibitors (Tanimoto coefficient >= 0.7) are currently registered in SciFinder.[36]

2.5. Features of the hit compounds

The CP3-3284 series compounds have a skeleton with a diazene in common. In addition to the diazene, compounds **3**, **4**, **5** and **6** have a benzothiophene ring, and their predicted binding modes with the NS3-4A protease were almost the same, involving a hydrophobic interaction between the skeleton and various residues, such as Val158 or Ala166. (Fig. 5a,b, those of compounds **3** & **6** are in Figs. S1 and S2).

The predicted binding mode of compound **10** involved interactions with Val158 and Ala166, which are close to the side chain of Arg123 (Fig. 5a). One of the reasons why the predicted binding mode was not stable is that the side chain of Arg123 is also not stabilized, since it is influenced by the multiple side chain conformations in the receptor ensembles (Fig. 5b). Moreover, most of the side chain atoms of Arg123 were ignored in the collision term of the GENIUS docking system (Table S1). Therefore, an undesirable angle in the hydrogen bond between the N atom of Arg123 and the O atom of NO_2 (Fig. 5a) would be observed in the flexible region of the receptor. The diazene moiety of the identified inhibitors formed a hydrogen bond with the oxygen atom of the main chain of Ala157, and the carbonyl group of the inhibitors also formed a hydrogen bond with the nitrogen atom of the main chain of Ala157 (Fig. 5a).

The IC_{50} values of compounds **3** and **4** were 4- to 12-fold lower than those of compounds **5** and **6**. In compounds **3** and **4**, the diazene and benzothiophene are ortho-substituted on the central benzene ring, while they are para-substituted in compounds **5** and **6**. The ortho-substituted benzothiophene moiety is predicted to interact more tightly with a hydrophobic surface.

In the replicon assay, the EC_{50} values of the four compounds (**3**–**6**) were approximately 10-fold larger than their IC_{50} values. In terms of hydrophobicity, very high calculated log*P* values (6.24–6.74) were observed for these compounds. Generally, hydrophobic compounds demonstrate good cell permeability. However, strong hydrophobicity also causes non-specific binding to the cell membrane. Therefore, these compounds would be less potent in the cell-based assay, as compared to the enzyme assay. In terms of cell toxicity in the ATP assay, compounds **3**, **4**, **7**, and **9** were more toxic than compound **10**. To clarify the preliminary structure–activity relationship, the R1 or R2 part (Fig. 6) of compound **10** was modified, by introducing methyl, ethyl, and nitro groups (Fig. 4). The inhibition activity of the derivatives was not significantly changed. In compound **7**, which has a naphthalene ring instead of the central benzene ring, the inhibition activity was decreased to 64.3 μM, because the atomic collision increased due to the larger volume of the sub-structure, extended by changing the substituent from benzene to naphthalene. The nitrobenzene group was commonly found at the T1 position of the active compounds (Fig. 6). The nitrobenzene group is an electron-poor aromatic ring, and is suitable to tightly bind to the electron-rich aromatic ring of His57. The nitro group also formed a weak hydrogen bond with Arg123 (Fig. 5a). In a future study, we will generate new compounds by introducing other electron-deficient substituents to interact with His57 and more powerful H-bonding acceptors to interact with Arg123, based on these structure–activity relationships.

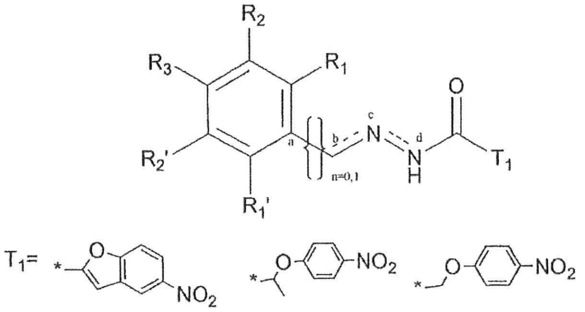
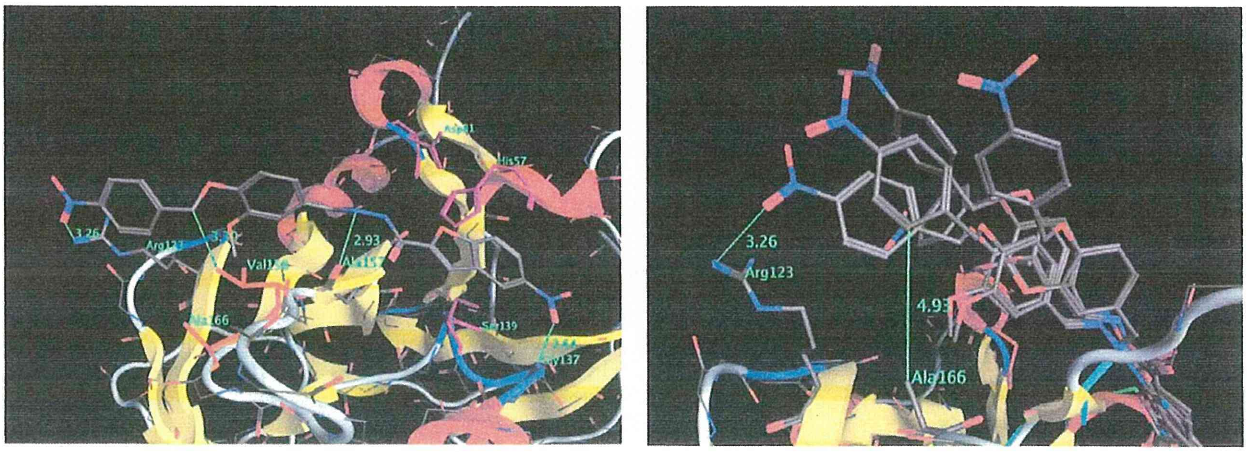


Figure 6. The common scaffold among CP3-3284 series. T1 means substructures in the CP3-3284 series.

2.6. Consideration of the predicted binding modes of the hit compounds

Since the CP3-3284 series compounds inhibited the protease activity and the cell viability, these compounds were considered to be promising as competitive inhibitors of the HCV NS3-4A protease. In a recent study, the interactions around the catalytic triad have been regarded as being important in NS3-4A protease inhibitor design.[10,11] Since the NS3-4A protease involves four connections of the HCV protein precursors, such as NS3-NS4A, NS4A-NS4B, NS4B-NS5A and NS5A-NS5B,[37] it is likely to identify peptide-type inhibitors. Generally, docking software emphasizes hydrophilic interactions, such as H-bonds, as compared with hydrophobic interactions, such as the interaction on the planar β sheet. To evaluate that kind of interaction and to identify the compounds that interact with the planar β sheet more accurately, it is necessary to determine the residues that interact with the ligand.[38] To overcome the problems with the conventional docking software, we set the hydrophobic interactions with the planar β sheet (Val158 and Ala166). Since the potent compounds **3** and **6** (IC_{50} values 1.06 and 4.08 μM, respectively) were discovered to form hydrophobic interactions between the 3-chlorobenzothiophene ring and the β sheet (the predicted binding modes are included in the Supplementary data), our pharmacophore constraints (that is, the EIPs) were effective to detect a new class of non-peptide inhibitors that interact with the planar β sheet.



(a) One of the predicted binding modes of CP3-3284-53

(b) All of the predicted binding modes of CP3-3284-53

Figure 5. Predicted binding modes of CP3-3284-53(10); Ribbon representation: one of the conformations of the NS3-4A protease. Thick stickrepresentation: predicted binding mode(s) of CP3-3284-53. Purple: the catalytic triad, red: hydrophobic residue on the β sheet.

2.7. Validation of specificity for CP3-3284 series compounds and known-inhibitors by the GENIUS docking system with the EIP

In this research, to identify the HCV NS3-4A protease inhibitors that differ from the conventional macrocyclic or peptide type inhibitors, the EIP was set to interact with not only the active site but also the β sheet. We validated the effectiveness of the GENIUS docking system to detect the CP3-3284 series compounds, using the obtained EIP for the NS3-4a protease. The docking and the subsequent GENIUS score ranking of well-known protease inhibitors and 166,206 compounds (described in the Section 4) used as decoy compounds were performed. If the active compounds were ranked higher than the decoy compounds, then the in silico screening procedure can detect the inhibitors efficiently. The enrichment factor (EF) is one of the popular metrics for screening efficiency.[39] In this case, the EF(x) values were EF(1%) = 14.3, EF(5%) = 11.4, and EF(10%) = 7.1 (x means the top x% of the total number of all calculated compounds). The number of active compounds is 21, including the CP3-3284 series (Fig. 4) and the macro-cyclic and peptide inhibitors. Moreover, the rank orders between the CP3-3284 series and the other active compounds were compared, to validate the specificity for the CP3-3284 series of the obtained EIP for the NS3-4A protease. All of the CP3-3284 series compounds were ranked higher than the other active compounds. In addition, the Wilcoxon rank sum test indicated a significant difference between the distributions of the ranking between the CP3-3284 series and the other active compounds (p-value <5.76e−11). This result shows that the EIP obtained for the NS3-4A protease had specificity for the CP3-3284 series compounds. Moreover, the ranking of the CP3-3284 series was higher than that of the macrocyclic compounds by the GENIUS score (Table 2). The peptide inhibitors could not be docked by the EIP because they lacked the ligand atoms specified in the obtained EIP. It was demonstrated that the GENIUS docking system, using the combination of the induced fit and the obtained EIP, had the capability to selectively detect a new class of inhibitors (CP3-3284 series compounds) that are neither peptide-type nor macrocyclic inhibitors.

2.8. Validation of the detection capability for the CP3-3284 series compounds in terms of induced-fit and no-induced fit in the GENIUS scoring function

In order to clarify the effects of induced fit docking by GENIUS on the discovery of the CP3-3284 series compounds, a docking without consideration of induced fit was performed. To cancel the consideration of induced fit, the X-ray structure complexed with TMC-435 (PDB code: 3KEE) was used, instead of the receptor conformation ensemble. In addition, no collisions between ligand atoms and receptor atoms were allowed. Except for the receptor coordinates and the collision term, the docking calculation conditions were the same as those of the previous GENIUS induced fit docking calculation. Five docking calculations were performed with the receptor, and as a result, the average value of the GENIUS score with induced fit was about three times better than that with the fixed receptor (Fig. 7). The obtained EIP contributed to the discovery of compounds that formed hydrophobic interactions with Val158 and Ala166 on the β sheet, arising from induced fit. The reason for the worse score of CP3-3284-53 (**10**) in the case of the fixed receptor is mainly due to the collision with Arg123, which was permitted in the case of an induced fit setting. We have shown that our defined EIP functions are effective with a receptor that functions by induced fitting, including side chain fluctuations. Moreover, the average score of the decoy distribution was better (6227.2) than that of the induced fit receptor mode. Additionally, in the case of the induced fit mode, the standard deviation of the score was larger than that of the fixed mode. Since the induced fit mode used multiple receptor conformations for the docking calculation, more diverse binding modes were generated, thus enlarging the standard deviation of the induced fit mode. Interestingly, the top scores of the decoy compounds for both the fixed and induced fit modes were almost the same (although the compounds with the top scores were not the same). The average score of compound **10** in the induced fit mode (965.3) and its rank (1st place) were quite improved, as compared with those in the fixed mode. Although the ranking is largely influenced by the selection of decoy compounds, the induced fit mode played very important roles in the discovery of the non-peptide inhibitor **10**.

Table 2
Ranking of the discovered compounds and the macrocyclic inhibitors and peptide-mimic inhibitors by the GENIUS docking system

ID	Rank	SD.
CP3-0032	2163	636
CP3-0084	4410	1336
CP3-3284-53	1089	445
CP3-3284-65	1542	604
CP3-3284-66	2291	170
CP3-3284-125	12260	178
CP3-3284-126	10511	437
CP3-3284-131	1810	368
CP3-3284-132	4245	1254
CP3-3284-142	11047	441
CP3-3284-164	3056	478
CP3-3284-53-s01	3424	686
CP3-3284-53-s02	1802	856
CP3-3284-53-s03	3200	346
CP3-3284-53-s04	1415	369
BILN-2061	23876	984
ITMN-191	18140	962
MK-7009	22402	3966
TMC-435	19032	1788
VX-950	N/A	N/A
SCH-503034	N/A	N/A

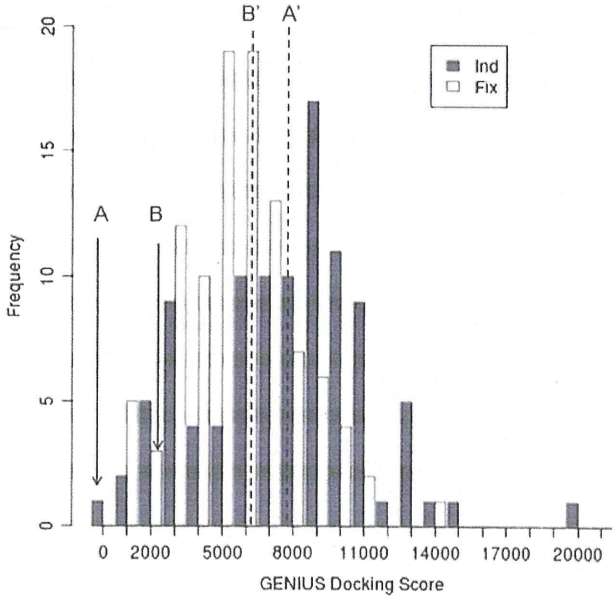


Figure 7. Distributions (histogram) of the docking scores of the decoy compounds and CP3-3284-53, using the induced-fit and the fixed receptor modes; 'Ind' means the trial using the induced-fit receptor. 'Fix' means the trial using the fixed receptor. A: The average score of CP3-3284-53 in the induced-fit receptor mode. B: The average score of CP3-3284-53 in the fixed receptor mode. A': The average score of the decoy compounds in the induced-fit receptor mode. B': The average score of the decoy compounds in the fixed receptor mode. Raw data are available in Supplementary data.

2.9. Validation of the effectiveness of the obtained EIP for the CP3-3284 series compounds

To verify the effectiveness of the obtained EIP for the HCV NS3-4A protease docking, four different EIPs derived from the obtained EIP were used for in silico screenings. The EIPs are listed in Table 3a. As active compounds, 15 of the CP3-3284 series compounds were used, and as the decoy compounds, 3,000 compounds randomly selected from PubChem were used. For each decoy compound, the GENIUS score was calculated once. For each CP3-3284 series compound, the average score of five calculations was used. The EFs are listed in Table 3b. In the case of the EIP(1) condition, the EFs were quite poor. This result shows that it is difficult to obtain active compounds when only the active site atoms in the EIP are specified. In the case of EIP(4), the EFs were better than those of the other conditions. This result shows that EIP(4) was optimized for the CP3-3284 series. Therefore, if EIP(4) had not been used, then the CP3-3284 series compounds probably would not have been detected. However, when EIP(3) was used, the EF(5%) and EF(10%) values gave good results, and the EF(10%) value using EIP(2) was reasonable. In a future study, in order to obtain different compounds from the CP3-3284 series, we plan to perform a docking calculation with a new EIP, with KEYATM added, on the basis of EIP(2) or EIP(3). This GENIUS docking system, using these EIPs, is expected to identify a new class of HCV NS3-4A protease inhibitors that interact with the flexible region, in addition to the inhibitors detected by the conventional docking.

2.10. Consideration of the collision term in the GENIUS docking system

In general, it is hard to determine only one structure coordinate by NMR, because only a few restrictions, such as NOEs and the torsion angles, are available. However, NMR structures include information related to the flexibility of the protein molecule in solution. Therefore, it is likely that the flexible atoms in the NMR structure ensemble are ignored in the calculation of the collisions between the protein and the ligand. Table S3 summarizes the atoms that were judged as being flexible, based on a cluster analysis of the torsion angles, and thus were ignored in the collision term calculation. Interestingly, while all of the atoms of His57 in the active site were flexible, different atoms were flexible in Arg119, Arg123, and Arg155, because the flexible regions of the side chains were different. In Arg119, the NE, CZ, NH1, and NH2 atoms were permitted to collide. However, in Arg123 and Arg155, CG and CD were also added. For example, in Glide, to express the induced fit of the receptor, an intermolecular collision can be relaxed by scaling each VDW radius. According to Table S3, the flexibilities of receptor atoms are dramatically different, even in the same residue. Therefore, if very small scaling coefficients were uniformly set for all of the atoms in a binding site, then most of the real inhibitors could be docked into the active site without any collision. However, many inactive compounds would also fit, and the screening efficiency would be very low. Therefore, the individual assignment of each atom, which permits a collision using the degree of torsion angle preservation derived from experimental structures (NMR or multiple X-ray structures), was effective to address the local softness of the receptor.

Table 3
Partially-divided EIPs and Enrichment Factors for each partially divided EIP

KEYATM	EIP (1)	EIP (2)	EIP (3)	EIP (4)
O.3 100 2.58 NE2 HISA_36	1	1	1	1
O.co2 100 2.60 N GLYA_116	1	1	1	1
DONOR 100 3.40 O ARGA_134		1	1	1
ACPTR 100 2.60 N ALAA_136		1	1	1
DONOR 100 2.60 O ALAA_136		1	1	1
C.3 100 2.60 CB ALAA_145			1	1
C.3 100 3.80 CB VALA_137				1

	EF (1%)	EF (5%)	EF (10%)
EIP(1)	0.0	0.0	1.3
EIP(2)	0.0	4.0	8.0
EIP(3)	6.7	20.0	10.0
EIP(4)	46.7	20.0	10.0

(a) The upper table; Enable KEYATMS in each partially-divided EIP. If the KEYATM was valid in the EIP, then the corresponding column bit was on. EIP (4) is the same as the EIP set up for the HCV NS3-4A in silico screening in this study. EIP (1): used KEYATMS only near the active-site 3 residue. EIP (2): used the hydrogen bond interactions and EIP (1); EIP (3): used part of the hydrophobic interaction and EIP (3). (b) The lower table; EF(x%) for each partially-divided EIP.

3. Conclusion

A new induced fit docking system, GENIUS, was developed, using collision term modification based on an experimentally determined protein structure ensemble and the essential interaction pair (EIP). The GENIUS system was applied to virtually screen HCV NS3-4A protease inhibitors, and a new class of non-peptide inhibitors was successfully identified. The EIPs for the induced fit of Arg123 on the β sheet and the hydrophobic interaction with the ligand in the open space were extracted by analyses of the binding site. Based on the ranking of the compounds by the GENIUS score, 97 compounds were selected and purchased. Among them, 27 compounds exhibited >50% inhibition at 100 μM in the protease inhibition assay. In the cell-based infection inhibition assay (replicon assay), two compounds showed 10 μM level potency (EC_{50}: 13 and 23 μM).

From a 2D similarity search of the chemical series, 140 compounds were obtained, and five compounds with IC_{50} values lower than 10 μM were identified. In particular, compound **3** was the most potent, with an IC_{50} of 1.06 μM. Unfortunately, since it exhibited cytotoxicity, this compound is not suitable as a seed molecule for drug development. Instead, compound **10**, which has 10 μM level potency (IC_{50}: 8.59 μM and EC_{50}: 12 μM) and no toxicity at >80 μM, was selected, and the preliminary structure–activity relationship was analyzed. We believe that compound **10** is promising as a seed for future synthetic development. The discovered compounds represent a new class of non-peptide HCV NS3-4A protease inhibitors. Furthermore, the new chemical series lacks an asymmetric carbon, unlike the existing inhibitor, and does not have a macrocyclic structure. Therefore, in terms of the synthetic feasibility and the ADME profile, the discovered chemical series has chemical tractability, as compared with the conventional peptide-type or macrocyclic NS3-4A inhibitors. The obtained EIP was capable of selectively identifying the CP3-3284 series, based on the validation results using both the induced fit and fixed receptor modes of GENIUS. In the validation, the score of compound **10** was greatly improved when induced fit was enabled. The rank of compound **10** over the decoy compounds and the EF of the CP3-3284 series were also superior in the induced fit mode. The effectiveness of the EIP was validated using the EF values under different EIP conditions. To improve this docking system, the collision coefficient was not set as a binary bit (0 or 1) for every atom of the receptor, but instead to a value between 0 and 1, by the clustering of the receptor conformation ensemble. It is hoped that a compound with the new skeleton identified by this research will be useful for future HCV therapies.

4. Materials and methods

4.1. In silico experiment schema

4.1.1. Receptor coordinates for docking calculations

The NMR structure of the HCV NS3-4A protease complexed with an inhibitor (PDB code 1DXW[24]) was used as the receptor for this in silico screening. The structure was complexed with the peptide mimic inhibitor (3-amino-5,5-di-fluoro-2-keto-pentan-1-oic acid), which forms a covalent bond with Ser139 in the active site. The 20 registered structures were used for the receptor conformation ensemble. We considered the atomic coordinates in which the torsion angle is not maintained among the NMR conformations to have a low possibility for interaction with ligands in the stable conformation of the NS3-4A receptor. Thus, the collisions between the receptor-ligand atoms in the flexible regions were tolerated. The criteria of flexibility were determined based on the preservation of the corresponding torsion angle of the receptor ensemble by clustering, as mentioned later.

4.1.2. Clustering of the ensemble of receptor conformations

The ensembles of the receptor conformation were clustered, in order to consider induced fit by the receptor. All of the side chain torsion angles maintained in the parent population, in the range of variation around the average angle of α % and plus or minus β degrees, were collected. The collected residues were referred to as the rigid residues. However, when the χ angle of the origin of the side chain was not maintained, it was assumed that the following atoms in the side chain were also not maintained, and these residues were referred to as the flexible residues. The side chain atoms of the flexible residues were ignored in the collision term of the GENIUS scoring function, which evaluates interactions between the receptor and the docking ligand. In the case of the NS3-4A protease, collisions between the docking ligand and the main chain atoms were not permitted. The details of the defined scoring functions are mentioned below. GENIUS (GENerating InduCed Systems),[40,41] which we encoded, implemented flexible ligand docking and induced-fit ligand docking algorithms, using the above scoring function.

4.1.3. Introducing of EIP

GENIUS requires three-dimensional receptor coordinate(s), ligand structures and essential interaction pairs (EIP). One EIP entry consists of an interaction pair that specifies the atom types of both the receptor and ligand atoms, the equilibrium distance, and the strength of the constraint. For example, if the CB atom of Val137 in the receptor interacts with the SP3 carbon atom in the ligand with an equilibrium distance of 3.8 Å, and using the constraint value of 100, its EIP is described as follows:

```
KEYATM C.3 100 3.80 CB VALA_137
```

In the PDB format, the character string of amino acid residues is normally presented with capital letters. Therefore, it was similarly treated by the EIP. The designation of the hydrogen donor and acceptor is also possible, in addition to the character string full match of the atomic species. One or more combination(s) of the designation are available for the EIP. When at least one of the EIP criteria cannot be fulfilled, because the indicated atom type does not exist in the docking ligand, the docking calculation can be skipped.

4.1.4. Generation of the initial interaction structure in the binding site

First, the initial binding mode of each docking ligand was prepared. Dummy atoms were generated around the atoms of the receptor specified by the EIP(s), and the atoms between the ligand and the dummy were structurally aligned using the DALI[42]-like algorithm, while maintaining the initial ligand conformation. The formula is provided below:

$$S = \sum_{i=1}^{N} \sum_{j=1}^{N} \phi(i,j) \tag{1}$$

$$\omega = \exp(-|d^A_{i,j} - d^B_{i,j}|)^2 \tag{2}$$

$$\phi(i,j) = \begin{cases} \theta(1.0 - \lambda) & (i = j) \\ \frac{\theta}{\mu}(\omega - \lambda) & (i \neq j) \end{cases} \tag{3}$$

where N is the number of joints, $d^A{}_{i,j}$ is the distance between the i-th and j-th dummy atoms, $d^B{}_{i,j}$ is the distance between the i-th and j-th ligand atoms, and μ is the average distance of $d^A{}_{i,j}$ and $d^B{}_{i,\cdot}$. θ and λ are constants (1.535 and 0.81, respectively). To obtain the maximized S, the correspondence atom relationship between the dummy and the ligand was randomly generated 10,000 times.

4.1.5. GENIUS scoring function

A binding mode with a smaller score has an advantage in a protein–ligand interaction. To optimize the interaction of the initial ligand pose, the conformational changes of the ligand, translation and rotation, are repeated 8,000 times. In the case of using more than two receptor structures, one coordinate included in the ensemble of receptor conformations was randomly selected for every step. In addition, slight conformational changes (between plus or minus 1 degree) of the ligand were performed 5,000 times. The definition of the GENIUS scoring function U_{optimum} is described below.

$$U_{optimum} = U_{sar} + U_{hydrogenbond} + U_{hydrophobic} + U_{stacking} + U_{collision} + U_{ligand-internal} \tag{4}$$

The atomic radius and the distance of the interatomic interaction were determined by reference to the AMBER99[43] and MM3[44] parameters.

4.1.5.1. EIP term. One of the features of the GENIUS docking system is U_{sar}, which considers the EIP in the score function. This term is effective to make a specific ligand atom interact with a restricted binding site in the receptor. The formula is defined below:

$$U_{sar} = \sum_{i=1}^{N} \varphi_{sar}(i,j) \tag{5}$$

$$\varphi_{sar}(i,j) = K_{sar}(R_{sar} - R)^2 - \delta \tag{6}$$

where R_{sar} is the i-th equilibrium distance, R is the distance between the i-th specified atoms of the ligand and the receptor, K_{sar} is the i-th strength of the restraint, and δ is a constant equal to −20.0. When the interaction distance of the binding mode is close to the specified equilibrium distance, this term judges that the interaction is favorable.

4.1.5.2. Hydrogen bond (hb) term. The hydrogen bonding score is calculated for the acceptor (or donor) atom of the receptor closest the donor (or acceptor) atom of the ligand. These atom types were previously defined. The formula is shown below:

$$U_{hydrogenbond} = \sum_{i=1}^{N} \varphi_{hb}(i) \tag{7}$$

$$\varphi_{hb}(i) = \begin{cases} -\frac{K_{hb}(i)}{|R - R_{hb}(i)| + 1.0} & (\theta \leqslant 30.0) \\ -\frac{K_{hb}(i)}{(|R - R_{hb}(i)| + 1.0)\theta} & (\theta > 30.0) \end{cases} \tag{8}$$

where N is the number of hydrogen bonds, and R is the distance between the two atoms that formed each hydrogen bond. $R_{hb}(i)$ and $K_{hb}(i)$ are the equilibrium distance and a constant of the strength of the atom pair forming the hydrogen bond, respectively. θ is the angle of the hydrogen bond, in degrees (Fig. 8). If the hydrogen bonding angle exceeds 30 degrees, then the score rapidly worsens.

4.1.5.3. Hydrophobic bond (hyd) term.

The hydrophobic score is calculated between the atoms of Ala, Cys, Phe, Ile, Leu, Met, Pro, Val, Trp, and Tyr (–OH is ignored) and the atoms of the ligand, defined as the hydrophobic atom within a fixed distance, by the following formula:

$$U_{hydrophobic} = \sum_{i=1}^{M} \sum_{j=1}^{N} \varphi_{hyd}(i,j) \tag{9}$$

$$\varphi_{hyd}(i,j) = \begin{cases} -\frac{K_{hyd}(i,j)}{R-R_{hyd}(i,j)+1.0} & (R \geqslant R_{hyd}(i,j)) \\ -K_{hyd}(i,j) & (R < R_{hyd}(i,j)) \end{cases} \tag{10}$$

where N and M are the numbers of atoms that could form hydrophobic interactions in the ligand and the receptor, respectively (cutoff: 8.0 Å). $R_{hyd}(i, j)$ and $K_{hyd}(i,j)$ are the equilibrium distance and a constant defined for every interaction pair, respectively. R is the distance between the i-th ligand atom and the j-th receptor atom.

4.1.5.4. Stacking term.

The stacking score was calculated if the distance between the i-th receptor aromatic atom and the j-th ligand aromatic atom is less than 5.0 Å. The aromatic ring center where the i-th atom belongs to the receptor side, is defined as i′, the nearest aromatic atom is defined as j′, from the j-th atom of the ligand, and the score was calculated by the following formula (Fig. 9):

$$U_{stacking} = \sum_{i=1}^{M} \sum_{j=1}^{N} \varphi_{stacking}(i,j) \tag{11}$$

$$\varphi_{stacking} = \begin{cases} -K_{stacking}(i,j)R_{boundary} & (R_{boundary} < 0.0) \\ -K_{stacking(i,j)}\theta_{boundary} & (R_{boundary} \geqslant 0.0) \end{cases} \tag{12}$$

$$R_{boundary} = 1.0 - (R_{stacking}(i,j) - R)^2 \tag{13}$$

$$\theta_{boundary} = |1.0 - \Theta| \tag{14}$$

$$\Theta = \min\{\frac{\pi}{180.0}(\theta - 90.0)^2\}(\theta : \theta_{i'j'} or \theta_{ij'}) \tag{15}$$

where N and M are the numbers of atoms that could form stacking interactions in the ligand and the receptor, respectively. $R_{stacking}(i,j)$ and $K_{stacking}(i,j)$ are the equilibrium distance and a constant defined for every interaction pair, respectively.

4.1.5.5. Intermolecular collision term.

The intermolecular collision score was calculated for the atoms of the main chains and the rigid side chains, if the receptor ensemble was used. If the interatomic distance R between the i-th atom of the receptor and the j-th atom of the ligand is within the defined collision distance, then the following formula was applied:

$$U_{collision} = \sum_{i=1}^{M} \sum_{j=1}^{N} \varphi_{collision}(i,j) \tag{16}$$

$$\varphi_{collision}(i,j) = K_{collision}\varepsilon(i)(R_{collision}(i,j) - R)^2 \tag{17}$$

$$\varepsilon(i) = \begin{cases} 0 \\ 1 \end{cases} \tag{18}$$

where M is the number of receptor atoms, N is the number of ligand atoms, and $K_{collision}$ is a constant equal to 1,000.0. $R_{collision}(i,j)$ is the summation of van der Waals radii of the i-th ligand atom and the j-th receptor atom. ε (i) is the collision coefficient, set to 1 or 0 for the receptor atoms through clustering of the receptor ensembles. If the i-th atom is the ignored atom, then it is set to 0. Otherwise, it is set to 1.

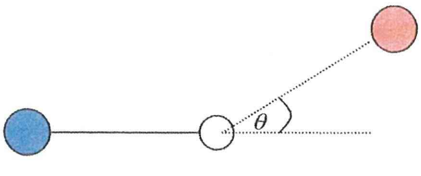


Figure 8. Definition of the hydrogen bond interaction. The red circle is the acceptor atom, the blue circle is the acceptor atom, and the white circle is a hydrogen atom. θ is the hydrogen bond angle.

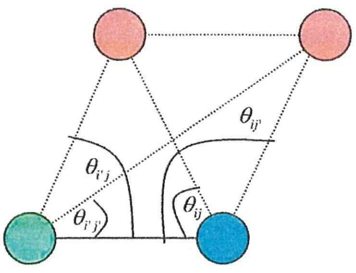


Figure 9. Definition of the stacking interaction. The cyan circle is the i-th atom in the aromatic ring of the receptor. The green circle means centroid of the aromatic ring including the i-th atom. The red circle is the j-th atom in the aromatic ring of the ligand.

4.1.5.6. Internal ligand term.

In order to avoid ligand docking poses with collapsed internal structures, such as when the bond length is broken by repeating the rotation, and by intramolecular collisions, a very strong restraint was added by the following formula:

$$U_{ligand-\mathrm{internal}} = \sum_{i=1}^{L} \varphi_{bond-length}(i) + \sum_{i=1}^{N} \sum_{j=1}^{S} \varphi_{\mathrm{internal-collision}}(i,j) \tag{19}$$

$$\varphi_{bond-length}(i) = K_{bond-length}\{(R_{bond-length}(i) - R_1)^2\} \tag{20}$$

$$\varphi_{\mathrm{internal-collision}}(i,j) = K_{\mathrm{internal-collision}}(R_{\mathrm{internal-collision}} - R_2)^2 \tag{21}$$

where L is the number of rotated bonds. N is the number of atoms of the ligand. S is the number of the i-th atom and the atoms that do not form a covalent bond. $K_{bond\text{-}length}$ is a constant equal to 100,000.0. $R_{bond\text{-}length}(i)$ is the bond length of the ligand in the initial structure. $K_{internal\text{-}collision}$ is a constant equal to 150.0. $R_{internal\text{-}collision}$ is a constant equal to 2.2(Å). R_1 is the distance for two atoms that form a covalent bond. R_2 is the distance between the i-th atom and an atom that does not form a covalent bond to the i-th atom.

4.1.6. Setup of the EIP used in the NS3-4A protease in-silico screening

The EIP can be automatically set up when a previously reported interaction is available from an X-ray structure or the associated literature. Nevertheless, in order to accurately dock a ligand to the important position of the receptor, the EIP should be determined manually. The EIP setting and the docking calculation were repeated until it was judged that the drug-like skeletons of compounds and appropriate binding modes were included in the ranking.

4.1.7. Compound database used for in silico screening

The MDL Available Chemical Directory 2005(ACD)[45] was used as the compound database for in silico screening (total 371,040 compounds). The database included the 2D structures of compounds that are commercially available. Ranking by the GENIUS score was performed for 166,206 compounds with molecular weights between 300 and 800. Generally, to reduce the number of docking compounds, drug like filter(s), such as Lipinski's Rule of five,[46] were applied to the compound database. In the GENIUS docking system, the compounds without the atomic type specified in the EIP were removed from the docking calculation. For example, a ligand without a donor atom could not be docked, if the donor is specified in the EIP. This is equivalent to performing a pre-docking filtering of compounds by simple atomic species. After the in silico screening, the selection of the compounds that satisfy the EIP and a visual inspection of the predicted interaction status of higher ranked compounds were performed. After the NS3-4A inhibition assay, structurally similar compounds of the hit compounds were selected by a 2D-similarity search of the MDL ISIS Base[45] and were purchased.

4.2. In vitro experiment schema

The protease inhibition activities of the compounds selected by in silico screening were measured, as the primary screen. For the hit compounds by the enzyme assay, one or both of the two cell viability test(s) described below were applied. Moreover, the concentration required for RNA generation inhibition in an HCV-infected cell was measured.

4.2.1. Enzyme assays

The recombinant NS3 protease protein was prepared for compound screening, as an engineered single-chain NS3-protease (scNS3).[47] The DNA sequence of the recombinant protein encoding the NS4A peptide (residues 21–33; GSVVIVGRIILSG) was genetically fused via a short linker (SGS), capable of making a beta-turn, to the N-terminus of the NS3 protease domain (residues 2–180, corresponding to 1,208-1,386 in the polypeptide). The gene encoding scNS3, with an N-terminal histidine-tag, was cloned into the pET32a(+) vector, and the protein was overexpressed in *Escherichia coli* (KRX). The scNS3 protein is reportedly soluble and fully active, with kinetic parameters virtually identical to those of the NS3/NS4A non-covalent complex. The protein was purified by chromatography on a HisTrap HP column (GE Healthcare), a HiPrep26/60 desalting column (GE Healthcare) and then a HiTrap Q column (GE Healthcare). Finally, the purified protein was concentrated on a HiLoad Superdex75p.g. 16/60 column in 20 mM Tris-HCl buffer (pH 8.0), containing 300 mM NaCl and 2 mM dithiothreitol.

The NS3 serine protease activity was measured by the fluorogenic assay based on intramolecular fluorescence resonance energy transfer, reported previously.[48] A quenched-fluorogenic substrate, Mca-Asp-Asp-Ile-Val-Pro-Cys-Ser-Met-Lys(Dnp)-Arg-Arg (QF-2), derived from the NS5A/5B junction of the HCV polyprotein, was synthesized by Toray Research Center (Kamakura, Japan). The purity of the synthetic peptide was more than 95%, based on an HPLC analysis.

The enzyme was pre-incubated with or without chemical compounds dissolved in dimethylsulfoxide (DMSO), in a reaction mixture containing 50 mM Tris HCl (pH 7.8), 30 mM NaCl, 5 mM $CaCl_2$ and 10 mM dithiothreitol, at 37 °C for 30 min, and then the reaction was started by the addition of QF-2 at a final concentration of 26 μM. The enzyme reaction was incubated at 37 °C. The progress of the enzyme reaction was detected in a 96-well black plate with a Twinkle LB970 multiwell plate reader (Berthold Technologies GmbhH & Co, Bad Wildbad/Germany), using F340 and F440 filters for excitation and emission, respectively.

4.2.2. Replicon assay

4.2.2.1. Cell culture. An HCV replicon harboring cell line, Huh7/Rep-Feo,[49,50] which expressed a chimeric gene encoding firefly luciferase and neomycin phosphotransferase, was used for the in vitro replication assay. Huh7/Rep-Feo cells were maintained in Dulbecco's modified Eagle's medium (DMEM), supplemented with 10% fetal calf serum and 250 μg/mL of G418.

4.2.2.2. Anti-HCV assay in Huh7/Rep-Feo cells. Huh7/Rep-Feo cells were seeded in a 48-well plate at a density of 2 × 10^4 cells per well. Two-fold serial dilutions of the test compounds in culture medium were added. After 72 h of culture, the expression levels of the HCV replicon were measured, using the luciferase assay system (Promega, Madison, WI, USA), and a JNR AB-2100 detector (Atto, Tokyo, Japan). The 50% effective concentration (EC_{50}) was defined as the concentration of compound that reduced the luciferase signal by 50%.

4.2.2.3. Cell viability assays. Huh7/Rep-Feo cells were seeded in a 96 well plate, at a density of 1 × 10^4 cells per well, and were incubated in the presence of various compounds. The 50% cytotoxicity concentration (CC_{50}) was determined 72 h after compound addition, using the cell titer 96 aqueous one solution cell proliferation assay (Promega, USA) (represented as "CC50 MTS" in this paper) or the cell titer-glo luminescent cell viability assay (Promega, USA) (represented as 'CC50 ATP' in this paper), according to the manufacturer's protocol.

4.3. Compounds

4.3.1. Commercially available compounds 1–10

Compound **1** was SALOR-INT L29,866-2, compound **2** was SALOR-INT L39,343-6, compound **3** was R941689, and compound **4** was R942251. All were purchased from SALOR-INT. Compound **5** was R985147, compound **6** was R988529, and compound **7** was L268399, all purchased from SALOR-INT. Compound **8** was NATR212614, compound **9** was NATR206554, and compound **10** was NATR206692, all purchased from Vitas-M. The purity of these compounds was unknown, and thus 100% purity was assumed in the enzyme assay and the replicon assay.

4.3.2. Synthesis of compounds 11–24

4.3.2.1. Purity analysis of the synthesized compounds. A Waters 996 PDA (254 nm) was used for detection. The column was a GL Science Inertsil ODS-3 (4.6 × 75 mm). The mobile phase gradient was a mixture of H_2O and CH_3CN (80:20, 0 min), (0:100, 5 min) with formic acid (0.1%). ^{1}H NMR spectra were obtained on a JEOL JNM ECP300 FT NMR system. Liquid chromatograph mass spectra (LC–MS) were detected in the ES positive mode.

4.3.2.2. Compound 15 (tert-butyl N-[(5-nitro-1-benzofuran-2-carbonyl)amino]carbamate). 5-Nitroglycerine benzofuran 2-carboxylic acid (**14**) (1.0 g, 4.8 mmol) was dissolved in methylene chloride (10 mL). Thionyl chloride (1.15 mL, 1.58 mmol) and *N,N*-dimethylformamide (30 μL) were added, and the resultant solution was stirred at 40 °C for 3.5 h. After cooling in air, the mixture was concentrated under reduced pressure. The residue was dissolved in THF (10 mL). Triethylamine (1.0 mL, 7.2 mmol) and tert-butyl-carbazate (0.77 g, 5.8 mmol) were then added, and the resultant solution was stirred at 0 °C for 1 h. Water was added to the reaction mixture, followed by ethyl acetate. The organic layer was dried with anhydrous sodium sulfate.

The combined organic solutions were filtered, and concentrated under reduced pressure. The resultant solid was washed with ethyl acetate/hexane to give compound **15** (1.3 g, 4.1 mmol, 85% yield).

^{1}H NMR (300 MHz, DMSO-d6) δ 10.69 (s, 1H), 9.12 (s, 1H), 8.83 (d, *J* = 2.4 Hz, 1H), 8.36 (dd, *J* = 2.4 Hz, *J* = 9.3 Hz, 1H), 7.95 (d, *J* = 9.3 Hz, 1H), 7.86 (s, 1H), 1.45 (s, 9H).

4.3.2.3. Compound 16 (5-nitro-1-benzofuran-2-carbohydrazide). Compound **2** (300 mg, 0.93 mmol) was mixed with hydrochloric acid (3 mL, 4 mol/L in dioxane) at 0 °C, and then stirred for 2 h at the same temperature. The solution was concentrated under reduced pressure, and the resultant solid was washed with isopropyl ether to give compound **16** (230 mg, 0.89 mmol, 96% yield).

^{1}H NMR (300 MHz, DMSO-d6) δ 8.86 (d, *J* = 2.4 Hz, 1H), 8.39 (dd, *J* = 2.4 Hz, *J* = 9.3 Hz, 1H), 8.06 (s, 1H), 7.99 (d, *J* = 9.3 Hz, 1H).

4.3.2.4. Compound 19a (3-methoxy-4-[(4-nitrophenyl)methoxy]benzaldehyde). Vanillin (**18a**) (0.50 g, 3.3 mmol) and 4-nitrobenzyl bromide (**17**) (0.71 g, 3.3 mmol) were suspended in ethanol (2.5 mL), and potassium carbonate (0.23 g, 1.7 mmol) was added. The resultant solution was stirred at 80 °C for 15 h. After cooling in air, the resultant solid was filtered and washed with ethanol, water, and ethanol to give compound **19a** (0.68 g, 2.4 mmol, 73% yield).

^{1}H NMR (300 MHz, $CDCl_3$) δ 9.87 (s, 1H), 8.26 (d, J = 8.7 Hz, 2H), 7.64 (d, J = 8.7 Hz, 2H), 7.46 (d, J = 1.5 Hz, 1H), 7.42 (dd, J = 1.5 Hz, J = 8.1 Hz, 1H), 6.97 (d, J = 8.1 Hz, 1H), 5.34 (s, 2H), 3.98 (s, 3H).

4.3.2.5. Compound 19b (3,5-dimethoxy-4-[(4-nitrophenyl)methoxy]benzaldehyde). Syringaldehyde (**18b**) (0.50 g, 2.7 mmol) and 4-nitrobenzyl bromide (**17**) (0.59 g, 2.7 mmol) were suspended in ethanol (3 mL), and potassium carbonate (0.38 g, 2.7 mmol) was added. The resultant solution was stirred at 80 °C for 3 h. After cooling in air, the resultant solid was filtered and washed with ethanol, water, and ethanol to give compound (**19b**) (0.70 g, 2.2 mmol, 81% yield) ^{1}H NMR (300 MHz, $CDCl_3$) δ 9.89 (s, 1H), 8.23 (d, J = 8.4 Hz, 2H), 7.68 (d, J = 8.4 Hz, 2H), 7.14 (s, 2H), 5.22 (s, 2H), 3.93 (s, 6H).

4.3.2.6. Compound 19c (3-methoxy-5-nitro-4-[(4-nitrophenyl)methoxy]benzaldehyde). 5-Nitrovanillin (**18c**) (0.50 g, 2.5 mmol) and 4-nitrobenzyl bromide (4) (0.55 g, 2.5 mmol) were suspended in THF (3 mL), and N, N-diisopropylethylamine was added (0.47 mL, 2.8 mmol). The resultant solution was stirred at room temperature for 12 h and at 65 °C for 2 h. After cooling in air, the insoluble matter was filtered, and the filtrate was concentrated under reduced pressure. The resultant solid was washed with diethyl ether to give compound **19c** (0.73 g, 2.2 mmol, 88% yield).

^{1}H NMR (300 MHz, $CDCl_3$) δ 9.95 (s, 1H), 8.26 (d, J = 8.4 Hz, 2H), 7.88 (d, J = 1.8 Hz, 1H), 7.68 (d, J = 1.8 Hz, 1H), 8.65 (d, J = 8.4 Hz, 2H), 5.38 (s, 2H), 4.03 (s, 3H).

4.3.2.7. Compound 20 ((4-formyl-2-methoxyphenyl) acetate). Vanillin (**18a**) (9.99 g, 65.6 mmol) was suspended in methylene chloride (50 mL), and then acetic anhydride (7.5 mL, 79 mmol) and pyridine (6.4 mL, 79 mmol) were added. The resultant solution was stirred at room temperature for 18 h. Water was added to the reaction mixture, and ethyl acetate was then added. The organic layer was washed with 1 N HCl, a saturated sodium hydrogen carbonate aqueous solution, and saturated brine, and was dried with anhydrous sodium sulfate. The organic layer was filtered, and concentrated under reduced pressure to give pure compound **20** (12.6 g, 64.9 mmol, 99% yield).

^{1}H NMR (300 MHz, $CDCl_3$) δ 9.95 (s, 1H), 7.51 (d, J = 2.1 Hz, 1H), 7.48 (dd, J = 2.1 Hz, J = 7.8 Hz, 1H), 7.23 (d, J = 7.8 Hz, 1H), 3.91 (s, 3H), 2.35 (s, 3H).

4.3.2.8. Compound 21 ((5-bromo-4-formyl-2-methoxyphenyl) acetate). Compound **20** (1.0 g, 5.1 mmol) and potassium bromide (2.0 g, 17 mmol) were suspended in water (10 mL), and then bromine (0.29 mL, 5.7 mmol) was added at 0 °C. The solution was stirred at room temperature for 15 h. The resultant solid was filtered, washed with water and dried to give pure compound **21** (1.29 g, 4.72 mmol, 93% yield).

^{1}H NMR (300 MHz, DMSO-d6) δ 10.15 (s, 1H), 7.68 (s, 1H), 7.52 (s, 1H), 3.86 (s, 3H), 2.30 (s, 3H).

4.3.2.9. Compound 22 (2-bromo-4-hydroxy-5-methoxybenzaldehyde). HCL (25 mL, 6 mol/L) was added to compound **21** (1.0 g, 3.7 mmol), and stirred at 90 °C for 4 h. After cooling in air, the resultant solid was filtered, washed with water and dried to give compound **22** (0.80 g, 3.5 mmol, 95% yield).

^{1}H NMR (300 MHz, DMSO-d6) δ 10.01 (s, 1H), 7.34 (s, 1H), 7.11 (s, 1H), 3.83 (s, 3H).

4.3.2.10. Compound 23 (2-bromo-5-methoxy-4-[(4-nitrophenyl)methoxy]benzaldehyde). Compound **22** (0.30 g, 1.3 mmol) and 4-nitrobenzyl bromide (**17**) (0.28 g, 1.3 mmol) were suspended in ethanol (3 mL), and then potassium carbonate was added. The resultant solution was stirred at 80 °C for 5 h. After cooling in air, water was added to the solution. The resultant solid was filtered, washed with water, followed by ethanol, and dried to give compound **23** (0.33 g, 0.90 mmol, 69% yield).

^{1}H NMR (300 MHz, $CDCl_3$) δ 10.19 (s, 1H), 8.28 (d, J = 8.4 Hz, 2H), 7.63 (d, J = 8.4 Hz, 2H), 7.47 (s, 1H), 7.07 (s, 1H), 5.29 (s, 2H), 3.94 (s, 3H).

4.3.2.11. Compound 19d (5-methoxy-2-methyl-4-[(4-nitrophenyl)methoxy]benzaldehyde). Compound **23** (1.0 g, 2.7 mmol) and methylboronic acid (245 mg, 4.09 mmol) were dissolved in dimethoxyethane (10 mL), and tetrakistriphenyl phosphine palladium (0.16 g, 0.14 mmol) and 2 mol/L sodium carbonate aqueous solution (4.1 mL, 8.2 mmol) were added to the solution. The resultant solution was stirred at 80 °C for 18 h. Methylboronic acid (245 mg, 4.09 mmol) and tetrakistriphenyl phosphine palladium (0.16 g, 0.14 mmol) were added to the solution, which was stirred at the same temperature for 24 h. After cooling in air, ethyl acetate was added, and the organic layer was washed with water and saturated brine, and dried with anhydrous sodium sulfate. The resultant mixture was filtered and concentrated under reduced pressure. The residue was fractionated by silica gel column chromatography to give compound **19d** (0.28 g, 0.93 mmol, 34% yield).

^{1}H NMR (300 MHz, $CDCl_3$) δ 10.22 (s, 1H), 8.26 (d, J = 9.0 Hz, 2H), 7.63 (d, J = 9.0 Hz, 2H), 7.40 (s, 1H), 6.68 (s, 1H), 5.31 (s, 2H), 3.94 (s, 3H), 2.59 (s, 3H).

4.3.2.12. Compound 24 (2-ethenyl-5-methoxy-4-[(4-nitrophenyl)methoxy]benzaldehyde). Compound **23** (0.30 g, 0.82 mmol) and tri-vinyl boroxine pyridine complex (99 mg, 0.41 mmol) were dissolved in dimethoxyethane (3 mL), and tetrakistriphenyl phosphine palladium (47 mg, 0.041 mmol) and 2 mol/L sodium carbonate aqueous solution were added. The resultant solution was stirred at 80 °C for 7 h. After cooling in air, ethyl acetate was added to the mixture. The organic layer was washed water and saturated brine, and was dried with anhydrous sodium sulfate. The resultant mixture was filtered and concentrated under reduced pressure. The residue was fractionated by silica gel column chromatography to give compound **24** (228 mg, 0.728 mmol, 89% yield).

^{1}H NMR (300 MHz, $CDCl_3$) δ 10.26 (s, 1H), 8.27 (d, J = 8.7 Hz, 2H), 7.65 (d, J = 8.7 Hz, 2H), 7.42 (s, 1H), 7.40 (dd, J = 17.4 Hz, J = 10.8 Hz, 1H), 6.97 (s, 1H), 5.53 (dd, J = 17.4 Hz, J = 0.9 Hz, 1H), 5.47 (dd, J = 10.8 Hz, J = 0.9 Hz, 1H), 5.34 (s, 2H), 3.97 (s, 3H).

4.3.2.13. Compound 19e (2-ethyl-5-methoxy-4-[(4-nitrophenyl)methoxy]benzaldehyde). Compound **24** (228 mg, 0.728 mmol) was dissolved in THF (2.3 mL), and palladium-fibroin (26 mg) was added. The resultant solution was stirred in a hydrogen atmosphere at room temperature for 20 h. Palladium-fibroin (22 mg) was added to the mixture, which was stirred in a hydrogen atmosphere at room temperature for 20 h. The resultant mixture was filtered and concentrated under reduced pressure. The residue was fractionated by silica gel column chromatography to give compound **19e** (215 mg, 0.682 mmol, 94% yield).

^{1}H NMR (300 MHz, CDCl$_3$) δ 10.22 (s, 1H), 8.26 (d, J = 8.7 Hz, 2H), 7.63 (d, J = 8.7 Hz, 2H), 7.42 (s, 1H), 6.71 (s, 1H), 5.31 (s, 2H), 3.94 (s, 3H), 2.96 (q, J = 7.5 Hz, 2H), 1.23 (t, J = 7.5 Hz, 3H).

4.3.2.14. Compound 10 (N-[(E)-[3-methoxy-4-[(4-nitrophenyl)methoxy]phenyl]methylideneamino]-5-nitro-1-benzofuran-2-carboxamide). Compound **16** (100 mg, 0.388 mmol) was suspended in toluene, and compound **19a** (122 mg, 0.425 mmol) and sodium acetate (36 mg, 0.44 mmol) were added. The resultant solution was stirred at 80 °C for 20 h. After cooling in air, the solution was filtered, and washed with water and toluene to give compound **10** (114 mg, 0.232 mmol, 60% yield, 100% purity).

^{1}H NMR (300 MHz, DMSO-d6) δ 8.83 (d, J = 2.4 Hz, 1H), 8.44 (s, 1H), 8.35 (dd, J = 9.0 Hz, J = 2.4 Hz, 1H), 7.27 (d, J = 8.7 Hz, 2H), 7.95 (d, J = 9.0 Hz, 1H), 7.91 (s, 1H), 7.73 (d, J = 8.7 Hz, 2H), 7.40 (s, 1H), 7.26–7.10 (m, 2H), 5.34 (s, 2H), 3.87 (s, 3H).

MS calcd. for $C_{24}H_{18}N_4O_8$ (M+H)$^+$ 491.11, found 490.8.

4.3.2.15. Compound 11 (N-[(E)-[5-methoxy-2-methyl-4-[(4-nitrophenyl)methoxy]phenyl]methylideneamino]-5-nitro-1-benzofuran-2-carboxamide). Compound **16** (116 mg, 0.450 mmol) was suspended in toluene (3 mL), and compound **19d** (149 mg, 0.495 mmol) and sodium acetate (41 mg, 0.50 mmol) were added. The resultant solution was stirred at room temperature for 1 h, at 80 °C for 5 h and at 100 °C for 4 h. After cooling in air, the solution was filtered and suspended in dimethoxyethane at 80 °C for 1 h. The mixture was then cooled in air, filtered and washed with dimethoxyethane to give compound **11** (122 mg, 0.242 mmol, 54% yield, 95.5% purity).

^{1}H NMR (300 MHz, DMSO-d6) δ 8.83 (s, 1H), 8.76 (s, 1H), 8.36 (d, J = 9.0 Hz, 1H), 8.27 (d, J = 7.8 Hz, 2H), 7.95 (d, J = 9.0 Hz, 1H), 7.90 (s, 1H), 7.72 (d, J = 7.8 Hz, 2H), 7.43 (s, 1H), 6.98 (s, 1H), 5.31 (s, 2H), 3.83 (s, 3H), 2.56–2.37 (s, 3H).

MS calcd. for $C_{25}H_{20}N_4O_8$ (M+H)$^+$ 505.13, found 504.8.

4.3.2.16. Compound 12 (N-[(E)-[2-ethyl-5-methoxy-4-[(4-nitrophenyl)methoxy]phenyl]methylideneamino]-5-nitro-1-benzofuran-2-carboxamide). Compound **16** (100 mg, 0.388 mmol) was suspended in toluene, and compound **19e** (134 mg, 0.425 mmol) and sodium acetate (35 mg, 0.43 mmol) were added. The resultant solution was stirred at 80 °C for 16 h. After cooling in air, the obtained solid was filtered and suspended in dimethoxyethane (4 mL). The resultant solution was stirred at 80 °C for 16 h. After cooling in air, the mixture was filtered and washed with dimethoxyethane to give compound **12** (128 mg, 0.247 mmol, 64% yield, 100% purity).

^{1}H NMR (300 MHz, DMSO-d6) δ 8.82 (s, 1H), 8.80 (s, 1H), 8.36 (d, J = 9.3 Hz, 1H), 8.27 (d, J = 7.8 Hz, 2H), 7.95 (d, J = 9.0 Hz, 1H), 7.90 (s, 1H), 7.73 (d, J = 7.8 Hz, 2H), 7.45 (s, 1H), 6.97 (s, 1H), 5.32 (s, 2H), 3.83 (s, 3H), 2.71 (q, J = 7.5 Hz, 2H), 1.61 (t, J = 7.5 Hz, 3H). MS calcd. for $C_{26}H_{22}N_4O_8$ (M+H)$^+$ 519.14, found 518.7.

4.3.2.17. Compound 13 (N-[(E)-[3-methoxy-5-nitro-4-[(4-nitrophenyl)methoxy]phenyl]methylideneamino]-5-nitro-1-benzofuran-2-carboxamide). Compound **16** (100 mg, 0.388 mmol) was suspended in toluene (2 mL), and compound **19c** (142 mg, 0.427 mmol) and sodium acetate (35 mg, 0.43 mmol) were added. The resultant solution was stirred at room temperature for 15 h and at 80 °C for 5 h. After cooling in air, the obtained solution was filtered and washed with water and toluene to give compound **13** (148 mg, 0.276 mmol, 71% yield, 95.9% purity).

^{1}H NMR (300 MHz, DMSO-d6) δ 8.80 (d, J = 2.1 Hz, 1H), 8.53 (s, 1H), 8.33 (dd, J = 9.0 Hz, J = 2.1 Hz, 1H), 8.26 (d, J = 9.0 Hz, 2H), 7.94 (d, J = 9.0 Hz, 1H), 7.92 (s, 1H), 7.78 (s, 1H), 7.73 (s, 1H), 7.70 (d, J = 9.0 Hz, 2H), 5.31 (s, 2H), 4.00 (s, 3H).

MS calcd. for $C_{24}H_{17}N_5O_{10}$ (M+H)$^+$ 536.1, found 535.8.

Acknowledgments

This work was supported by Health and Labor Sciences Research Grants for Research on Hepatitis, as well as by Grants-in-Aid for research on hepatitis and BSE, from the Ministry of Health, Labor and Welfare and by a contract research fund for the Program of Funding Research Centers for Emerging and Reemerging Infectious Diseases, from the Ministries of Education, Culture, Sports, Science, and Technology, Japan. This work was supported by the RIKEN Structural Genomics/Proteomics Initiative (RSGI), the National Project on Protein Structural and Functional Analyses, and the Targeted Proteins Research Program (TPRP), the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan. We thank Dr. Shunta Sasaki, from Structure Based Drug Design Research, for the MS data analysis.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bmc.2011.09.023.

References and notes

1. Alter, H. J.; Purcell, R. H.; Shih, J. W.; Melpolder, J. C.; Houghton, M.; Choo, Q.-L.; Kuo, G. N. *Engl. J. Med.* **1989**, *321*, 1494.
2. Choo, Q. L.; Kuo, G.; Weiner, A. J.; Overby, L. R.; Bradley, D. W.; Houghton, M. *Science* **1989**, *244*, 359.
3. Qureshi, S.; Qureshi, H.; Hameed, A. *Eur. J. Clin. Microbiol. Infect. Dis.* **2009**, *28*, 1409.
4. Alter, M. J. *Hepatology* **1997**, *26*, 62S.
5. Fried, M. W.; Shiffman, M. L.; Reddy, K. R.; Smith, C.; Marinos, G.; Gonçales, F. L., Jr.; Häussinger, D.; Diago, M.; Carosi, G.; Dhumeaux, D.; Craxi, A.; Lin, A.; Hoffman, J.; Yu, J. N. *Engl. J. Med.* **2002**, *347*(13), 975.
6. Berman, H. M. *Acta Crystallogr., Sect. A* **2008**, *64*, 88.
7. Yao, N.; Reichert, P.; Taremi, S. S.; Prosise, W. W.; Weber, P. C. *Structure* **1999**, *15*, 1353.
8. Thibeault, D.; Massariol, M.-J.; Zhao, S.; Welchner, E.; Goudreau, N.; Gingras, R.; Llinàs-Brunet, M.; White, P. W. *Biochemistry* **2009**, *48*, 744.
9. Liverton, N. J.; Holloway, M. K.; McCauley, J. A.; Rudd, M. T.; Butcher, J. W.; Carroll, S. S.; DiMuzio, J.; Fandozzi, C.; Gilbert, K. F.; Mao, S.-S.; McIntyre, C. J.; Nguyen, K. T.; Romano, J. J.; Stahlhut, M.; Wan, B.-L.; Olsen, D. B.; Vacca, J. P. *J. Am. Chem. Soc.* **2008**, *130*, 4607.
10. Berman, K.; Kwo, P. Y. *Clin. Liver. Dis.* **2009**, *13*(3), 429.
11. Perni, R. B.; Almquist, S. J.; Byrn, R. A.; Chandorkar, G.; Chaturvedi, P. R.; Courtney, L. F.; Decker, C. J.; Dinehart, K.; Gates, C. A.; Harbeson, S. L.; Heiser, A.; Kalkeri, G.; Kolaczkowski, E.; Lin, K.; Luong, Y.-P.; Rao, B. G.; Taylor, W. P.; Thomson, J. A.; Tung, R. D.; Wei, Y.; Kwong, A. D.; Lin, C. *Antimicrob. Agents Chemother.* **2006**, *50*(3), 899.
12. Lamarre, D.; Anderson, P. C.; Bailey, M.; Beaulieu, P.; Bolger, G.; Bonneau, P.; Bös, M.; Cameron, D. R.; Cartier, M.; Cordingley, M. G.; Faucher, A.-M.; Goudreau, N.; Kawai, S. H.; Kukolj, G.; Lagacé, L.; LaPlante, S. R.; Narjes, H.; Poupart, M.-A.; Rancourt, J.; Sentjens, R. E.; StGeorge, R.; Simoneau, B.; Steinmann, G.; Thibeault, D.; Tsantrizos, Y. S.; Weldon, S. M.; Yong, C. L.; Llinas-Brunet, M. *Nature* **2003**, *13*, 186.
13. Lin, T.-I.; Lenz, O.; Fanning, G.; Verbinnen, T.; Delouvroy, F.; Scholliers, A.; Vermeiren, K.; Rosenquist, A.; Edlund, M.; Samuelsson, B.; Vrang, L.; de Kock, H.; Wigerinck, P.; Raboisson, P.; Simmen, K. *Antimicrob. Agents Chemother.* **2009**, *53*(4), 1377.
14. Rajagopalan, R.; Misialek, S.; Stevens, S. K.; Myszka, D. G.; Brandhuber, B. J.; Ballard, J. A.; Andrews, S. W.; Seiwert, S. D.; Kossen, K. *Biochemistry* **2009**, *48*, 2559–2568.
15. McCauley, J. A.; McIntyre, C. J.; Rudd, M. T.; Nguyen, K. T.; Romano, J. J.; Butcher, J. W.; Gilbert, K. F.; Bush, K. J.; Holloway, M. K.; Swestock, J.; Wan, B. L.; Carroll, S. S.; DiMuzio, J. M.; Graham, D. J.; Ludmerer, S. W.; Mao, S. S.; Stahlhut, M. W.; Fandozzi, C. M.; Trainor, N.; Olsen, D. B.; Vacca, J. P.; Liverton, N. J. *J. Med. Chem.* **2010**, *25*, 2443.
16. Thompson, A. J. V.; McHutchison, J. G. *J. Viral Hepat.* **2009**, *16*, 377–387.
17. Wyles, D. L.; Kaihara, K. A.; Schooley, R. T. *Antimicrob. Agents Chemother.* **2008**, 1862.
18. Thompson, C. A. J. V.; McHutchison, J. G. *Aliment. Pharmacol. Ther.* **2009**, *29*, 689.
19. http://www.natap.org/2004/HCV/113004_01.htm.
20. Cubero, M.; Esteban, J. I.; Otero, T.; Sauleda, S.; Bes, M.; Esteban, R.; Guardia, J.; Quer, J. *Virology* **2008**, *370*, 237.
21. Yi, M.; Tong, X.; Skelton, A.; Chase, R.; Chen, T.; Prongay, A.; Bogen, S. L.; Saksena, A. K.; Njoroge, F. G.; Veselenak, R. L.; Pyles, R. B.; Bourne, N.; Malcolm, B. A.; Lemon, S. M. *J. Biol Chem.* **2006**, *281*, 8205.
22. Ismail, N. S. M.; Hattori, M. *Bioorg. Med. Chem.* **2011**, *19*, 374.

23. Ontoria, J. M.; Di Marco, S.; Conte, I.; Di Francesco, M. E.; Gardelli, C.; Koch, U.; Matassa, V. G.; Poma, M.; Steinkühler, C.; Volpari, C.; Harper, S. *J. Med. Chem.* **2004**, *47*, 6443.

24. Barbato, G.; Cicero, D. O.; Cordier, F.; Narjes, F.; Gerlach, B.; Sambucini, S.; Grzesiek, S.; Matassa, V. G.; De Francesco, R.; Bazzo, R. *EMBO J.* **2000**, *19*, 1195.

25. Cummings, M. D.; Lindberg, J.; Lin, T.-I.; de Kock, H.; Lenz, O.; Lilja, E.; Felländer, S.; Baraznenok, V.; Nyström, S.; Nilsson, M.; Vrang, L.; Edlund, M.; Rosenquist, A.; Samuelsson, B.; Raboisson, P.; Simmen, K. *Angew. Chem., Int. Ed.* **2010**, *22*, 1652.

26. Hagel, M.; Niu, D.; Martin, T. S.; Sheets, M. P.; Qiao, L.; Bernard, H.; Karp, R. M.; Zhu, Z.; Labenski, M. T.; Chaturvedi, P.; Nacht, M.; Westlin, W. F.; Petter, R. C.; Singh, J. *Nat. Chem. Biol.* **2011**, *7*, 22.

27. Barbato, G.; Cicero, D. O.; Nardi, M. C.; Steinkühler, C.; Cortese, R.; De Francesco, R.; Bazzo, R. *J. Mol. Biol.* **1999**, *289*, 371.

28. Gallo, M.; Pennestri, M.; Bottomley, M. J.; Barbato, G.; Eliseo, T.; Paci, M.; Narjes, F.; De Francesco, R.; Summa, V.; Koch, U.; Bazzo, R.; Cicero, D. O. *J. Mol. Biol.* **2009**, *385*, 1142.

29. Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. *J. Comput. Chem.* **1998**, *19*, 1639–1662.

30. Ewing, T. J. A.; Makino, S.; Skillman, A. G.; Kuntz, I. D. *J. Comput.Aided Mol. Des.* **2001**, *15*, 411.

31. Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. *J. Mol. Biol.* **1997**, *267*, 727.

32. Koska, J.; Spassov, V. Z.; Maynard, A. J.; Yan, L.; Austin, N.; Flook, P. K.; Venkatachalam, C. M. *J. Chem. Inf. Model.* **2008**, *48*, 1965.

33. Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. *J. Med. Chem.* **2004**, *47*, 1739.

34. Dunbrack, R. L., Jr.; Karplus, M. *J. Mol. Biol.* **1993**, *230*, 543.

35. Meiler, J.; Baker, D. *Proteins: Struct., Funct., Bioinf.* **2006**, *65*, 538.

36. https://scifinder.cas.org.

37. Chevaliez, S.; Jean-Michel Pawlotsky, J.-M.; Tan, S.-L., Ed.; Norfolk (UK): Horizon Bioscience; 2006. Chapter 1. HCV Genome and Life Cycle; http://www.ncbi.nlm.nih.gov/books/NBK1613/.

38. Schneider, G. *Nat. Rev. Drug Disc.* **2010**, *9*, 273.

39. Bender, A.; Glen, R. C. *J. Chem. Inf. Model.* **2005**, *45*, 1369.

40. Umeyama, H.; Watanabe, Y.; Arai, R., Japan Patent 4314128, 2009.

41. Umeyama, H.; Watanabe, Y.; Arai, R., Japan Patent 4314206, 2009;.

42. Holm, L.; Sander, C. *J. Mol. Biol.* **1993**, *5*, 123.

43. Wang, J.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21*, 1049.

44. Ma, B.; Lii, J.-H.; Allinger, N. L. *J. Comput. Chem.* **2000**, *21*, 813.

45. Symyx Technologies, Inc. Corporate Address: 3100 Central Expressway, Santa Clara, CA 95051.

46. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. *Adv. Drug Delivery Rev.* **2001**, *46*(1–3), 3.

47. Dimasi, N.; Pasquo, A.; Martin, F.; Di Marco, S.; Steinkühler, C.; Cortese, R.; Sollazzo, M. *Protein Eng. Des. Sel.* **1998**, *11*(12), 1257.

48. Kakiuchi, N.; Nishikawa, S.; Hattori, M.; Shimotohno, K. *J. Virol. Methods.* **1999**, *80*, 77.

49. Tanabe, Y.; Sakamoto, N.; Enomoto, N.; Kurosaki, M.; Ueda, E.; Maekawa, S.; Yamashiro, T.; Nakagawa, M.; Chen, C.-H.; Kanazawa, N.; Kakinuma, S.; Watanabe, M. *J. Infect. Dis.* **2004**, *189*, 1129.

50. Yokota, T.; Sakamoto, N.; Enomoto, N.; Tanabe, Y.; Miyagishi, M.; Maekawa, S.; Yi, L.; Kurosaki, M.; Taira, K.; Watanabe, M.; Mizusawa, H. *EMBO Rep.* **2003**, *4*, 602.

Research Article

EASL EUROPEAN ASSOCIATION FOR THE STUDY OF THE LIVER | JOURNAL OF HEPATOLOGY

Pre-treatment prediction of response to pegylated-interferon plus ribavirin for chronic hepatitis C using genetic polymorphism in *IL28B* and viral factors

Masayuki Kurosaki[1], Yasuhito Tanaka[2], Nao Nishida[3], Naoya Sakamoto[4], Nobuyuki Enomoto[5], Masao Honda[6], Masaya Sugiyama[2], Kentaro Matsuura[2], Fuminaka Sugauchi[2], Yasuhiro Asahina[1], Mina Nakagawa[4], Mamoru Watanabe[4], Minoru Sakamoto[5], Shinya Maekawa[5], Akito Sakai[6], Shuichi Kaneko[6], Kiyoaki Ito[7], Naohiko Masaki[7], Katsushi Tokunaga[3], Namiki Izumi[1,*], Masashi Mizokami[2,7]

[1]*Division of Gastroenterology and Hepatology, Musashino Red Cross Hospital, Tokyo, Japan;* [2]*Department of Virology, Liver Unit, Nagoya City University, Graduate School of Medical Sciences, Nagoya, Japan;* [3]*Department of Human Genetics, Graduate School of Medicine, University of Tokyo, Tokyo, Japan;* [4]*Department of Gastroenterology and Hepatology, Tokyo Medical and Dental University, Tokyo, Japan;* [5]*First Department of Internal Medicine, University of Yamanashi, Yamanashi, Japan;* [6]*Department of Gastroenterology, Kanazawa University, Graduate School of Medicine, Kanazawa, Japan;* [7]*Research Center for Hepatitis and Immunology, International Medical Center of Japan, Konodai Hospital, Ichikawa, Japan*

Background & Aims: Pegylated interferon and ribavirin (PEG-IFN/RBV) therapy for chronic hepatitis C virus (HCV) genotype 1 infection is effective in 50% of patients. Recent studies revealed an association between the *IL28B* genotype and treatment response. We aimed to develop a model for the pre-treatment prediction of response using host and viral factors.

Methods: Data were collected from 496 patients with HCV genotype 1 treated with PEG-IFN/RBV at five hospitals and universities in Japan. *IL28B* genotype and mutations in the core and IFN sensitivity determining region (ISDR) of HCV were analyzed to predict response to therapy. The decision model was generated by data mining analysis.

Results: The *IL28B* polymorphism correlated with early virological response and predicted null virological response (NVR) (odds ratio = 20.83, p <0.0001) and sustained virological response (SVR) (odds ratio = 7.41, p <0.0001) independent of other covariates. Mutations in the ISDR predicted relapse and SVR independent of *IL28B*. The decision model revealed that patients with the minor *IL28B* allele and low platelet counts had the highest NVR (84%) and lowest SVR (7%), whereas those with the major *IL28B* allele and mutations in the ISDR or high platelet counts had the lowest NVR (0–17%) and highest SVR (61–90%). The model had high reproducibility and predicted SVR with 78% specificity and 70% sensitivity.

Conclusions: The *IL28B* polymorphism and mutations in the ISDR of HCV were significant pre-treatment predictors of response to PEG-IFN/RBV. The decision model, including these host and viral factors may support selection of optimum treatment strategy for individual patients.

Introduction

Hepatitis C virus (HCV) infection is the leading cause of cirrhosis and hepatocellular carcinoma worldwide [1]. The successful eradication of HCV, defined as a sustained virological response (SVR), is associated with a reduced risk of developing hepatocellular carcinoma. Currently, pegylated interferon (PEG-IFN) plus ribavirin (RBV) is the most effective standard of care for chronic hepatitis C but the rate of SVR is around 50% in patients with HCV genotype 1 [2,3], the most common genotype in Japan, Europe, the United States, and many other countries. Moreover, 20–30% of patients with HCV genotype 1 have a null virological response (NVR) to PEG-IFN/RBV therapy [4]. The most reliable method for predicting the response is to monitor the early decline of serum HCV-RNA levels during treatment [5] but there is no established method for prediction before treatment. Because PEG-IFN/RBV therapy is costly and often accompanied by adverse effects such as flu-like symptoms, depression and hematological abnormalities, pre-treatment predictions of those patients who are unlikely to benefit from this regimen enables ineffective treatment to be avoided.

Recently, it has been reported through a genome-wide association study (GWAS) of patients with genotype 1 HCV that single nucleotide polymorphisms (SNPs) located near the *IL28B* gene are strongly associated with a response to PEG-IFN/RBV therapy in

Keywords: *IL28B*; ISDR; Peg-interferon; Ribavirin; Data mining; Decision tree.
Received 14 March 2010; received in revised form 22 June 2010; accepted 7 July 2010

* Corresponding author. Address: Division of Gastroenterology and Hepatology, Musashino Red Cross Hospital, 1-26-1 Kyonan-cho, Musashino-shi, Tokyo 180-8610, Japan. Tel.: +81 422 32 3111; fax: +81 422 32 9551.
E-mail address: nizumi@musashino.jrc.or.jp (N. Izumi).

Please cite this article in press as: Kurosaki M et al. Pre-treatment prediction of response to pegylated-interferon plus ribavirin for chronic hepatitis C using genetic polymorphism in *IL28B* and viral factors. J Hepatol (2010), doi:10.1016/j.jhep.2010.07.037

Table 1. Baseline characteristics of all patients, and patients assigned to the model building or validation groups.

	All patients n = 496	Model group n = 331	Validation group n = 165
Gender: male	250 (50%)	170 (51%)	80 (48%)
Age (years)	57.1 ± 9.9	56.8 ± 9.7	57.5 ± 10.2
ALT (IU/L)	78.6 ± 60.8	78.1 ± 61.4	79.7 ± 59.6
GGT (IU/L)	59.3 ± 63.6	58.9 ± 62.0	60.2 ± 66.9
Platelets (10^9/L)	154 ± 53	153 ± 52	154 ± 56
Fibrosis: F3-4	121 (24%)	80 (24%)	41 (25%)
HCV-RNA: >600,000 IU/ml	409 (82%)	273 (82%)	136 (82%)
ISDR mutation: ≤1	220 (88%)	290 (88%)	145 (88%)
Core 70 (Arg/Gln or His)	293 (59%)/203 (41%)	197 (60%)/134 (40%)	96 (58%)/69 (42%)
Core 91 (Leu/Met)	299 (60%)/197 (40%)	200 (60%)/131 (40%)	99 (60%)/66 (40%)
IL28B: Minor allele	151 (30%)	101 (31%)	50 (30%)
SVR	194 (39%)	129 (39%)	65 (39%)
Relapse	152 (31%)	103 (31%)	49 (30%)
NVR	150 (30%)	99 (30%)	51 (31%)

ALT, alanine aminotransferase; GGT, gamma-glutamyltransferase; ISDR, interferon sensitivity determining region; Arg, arginine; Gln, glutamine; His, histidine; Leu, leucine; Met, methionine; Minor, heterozygote or homozygote of minor allele; SVR, sustained virological response; NVR, null virological response.

Japanese [6], European [7], and a multi-ethnic population [8,9]. The last three studies focused on the association of SNPs in the *IL28B* region with SVR [7–9] but we found a stronger association with NVR [6]. In addition to these host genetic factors, we have reported that mutations within a stretch of 40 amino acids in the NS5A region of HCV, designated as the IFN sensitivity determining region (ISDR), are closely associated with the virological response to IFN therapy: a lower number of mutations is associated with treatment failure [10–13]. Amino acid substitutions at positions 70 and 91 of the HCV core region (Core70, Core91) also have been reported to be associated with response to PEG-IFN/RBV therapy: glutamine (Gln) or histidine (His) at Core70 and methionine (Met) at Core91 are associated with treatment resistance [4,14]. The importance of substitutions in the HCV core and ISDR was confirmed recently by a Japanese multicenter study [15]. How these viral factors contribute to response to therapy is yet to be determined. For general application in clinical practice, host genetic factors and viral factors should be considered together.

Data mining analysis is a family of non-parametric regression methods for predictive modeling. Software is used to automatically explore the data to search for optimal split variables and to build a decision tree structure [16]. The major advantage of decision tree analysis over logistic regression analysis is that the results of the analysis are presented in the form of flow chart, which can be interpreted intuitively and readily made available for use in clinical practice [17]. The decision tree analysis has been utilized to define prognostic factors in various diseases [18–25]. We have reported recently its usefulness for the prediction of an early virological response (undetectable HCV-RNA within 12 weeks of therapy) to PEG-IFN/RBV therapy in chronic hepatitis C [26].

This study aimed to define the pre-treatment prediction of response to PEG-IFN/RBV therapy through the integrated analysis of host factors, such as the *IL28B* genetic polymorphism and various clinical covariates, as well as viral factors, such as mutations in the HCV core and ISDR and serum HCV-RNA load. In addition, for the general application of these results in clinical practice, decision models for the pre-treatment prediction of response were determined by data mining analysis.

Materials and methods

Patients

This was a multicentre retrospective study supported by the Japanese Ministry of Health, Labor and Welfare. Data were collected from a total of 496 chronic hepatitis C patients who were treated with PEG-IFN alpha and RBV at five hospitals and universities throughout Japan. Of these, 98 patients also were included in the original GWAS analysis [6]. The inclusion criteria in this study were as follows (1) infection by genotype 1b, (2) lack of co-infection with hepatitis B virus or human immunodeficiency virus, (3) lack of other causes of liver disease, such as autoimmune hepatitis, and primary biliary cirrhosis, (4) completion of at least 24 weeks of therapy, (5) adherence of more than 80% to the planned dose of PEG-IFN and RBV for the NVR patients, (6) availability of DNA for the analysis of the genetic polymorphism of *IL28B*, and (7) availability of serum for the determination of mutations in the ISDR and substitutions of Core70 and Core91 of HCV. Patients received PEG-IFN alpha-2a (180 μg) or 2b (1.5 μg/kg) subcutaneously every week and were administered a weight adjusted dose of RBV (600 mg for <60 kg, 800 mg for 60–80 kg, and 1000 mg for >80 kg daily) which is the recommended dosage in Japan. Written informed consent was obtained from each patient and the study protocol conformed to the ethical guidelines of the Declaration of Helsinki and was approved by the institutional ethics review committee. The baseline characteristics are listed in Table 1. For the data mining analysis, 67% of the patients (331 patients) were assigned randomly to the model building group and 33% (165 patients) to the validation group. There were no significant differences in the clinical backgrounds between these two groups.

Laboratory and histological tests

Blood samples were obtained before therapy and were analyzed for hematologic tests and for blood chemistry and HCV-RNA. Sequences of ISDR and the core region of HCV were determined by direct sequencing after amplification by reverse-transcription and polymerase chain reaction as reported previously [4,11]. Genetic polymorphism in one tagging SNP located near the *IL28B* gene (rs8099917) was determined by the GWAS or DigiTag2 assay [27]. Homozygosity (GG) or heterozygosity (TG) of the minor sequence was defined as having the *IL28B* minor allele, whereas homozygosity for the major sequence (TT) was

Please cite this article in press as: Kurosaki M et al. Pre-treatment prediction of response to pegylated-interferon plus ribavirin for chronic hepatitis C using genetic polymorphism in *IL28B* and viral factors. J Hepatol (2010), doi:10.1016/j.jhep.2010.07.037

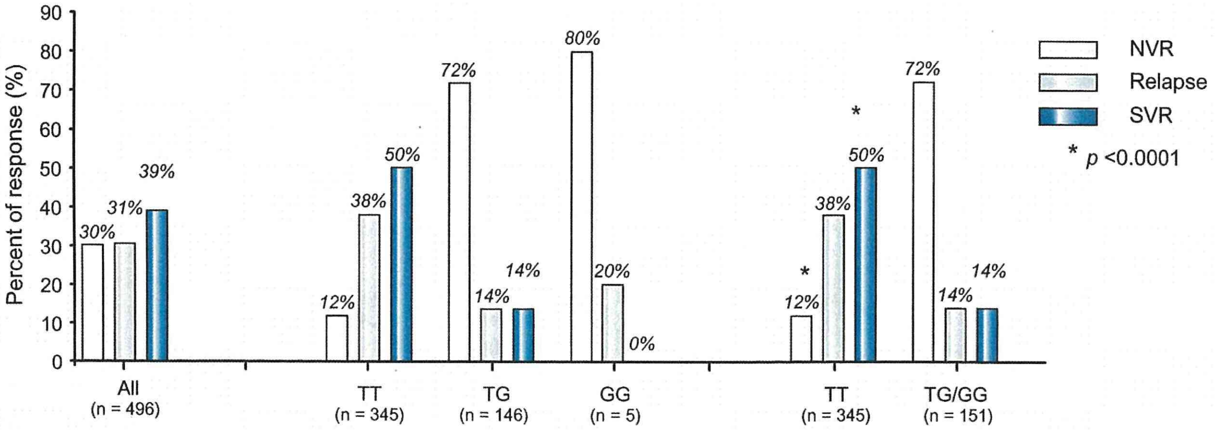


Fig. 1. Association between the IL28B genotype (rs8099917) and treatment response. The rates of response to treatment are shown for each rs8099917 genotype. The rate of null virological response (NVR), relapse, and sustained virological response (SVR) is shown. The *p* values are from Fisher's exact test. The rate of NVR was significantly higher (*p* <0.0001) and the rate of SVR was significantly lower (*p* <0.0001) in patients with the *IL28B* minor allele compared to those with the major allele. [This figure appears in colour on the web.]

defined as having the *IL28B* major allele. In this study, NVR was defined as a less than 2 log reduction of HCV-RNA at week 12 and detectable HCV-RNA by qualitative PCR with a lower detection limit of 50 IU/ml (Amplicor, Roche Diagnostic systems, CA) at week 24 during therapy. RVR (rapid virological response) and complete early virological response (cEVR) were defined as undetectable HCV-RNA at 4 weeks and 12 weeks during therapy and SVR was defined as undetectable HCV-RNA 24 weeks after the completion of therapy. Relapse was defined as reappearance of HCV-RNA after the completion of therapy. The stage of liver fibrosis was scored according to the METAVIR scoring system: F0 (no fibrosis), F1 (mild fibrosis: portal fibrosis without septa), F2 (moderate fibrosis: few septa), F3 (severe fibrosis: numerous septa without cirrhosis) and F4 (cirrhosis). Percentage of steatosis was quantified in 111 patients by determining the average proportion of hepatocytes affected by steatosis.

Statistical analysis

Associations between pre-treatment variables and treatment response were analyzed by univariate and multivariate logistic regression analysis. Associations between the *IL28B* polymorphism and sequences of HCV were analyzed by Fisher's exact test. SPSS software v.15.0 (SPSS Inc., Chicago, IL) was used for these analyses. For the data mining analysis, IBM-SPSS Modeler version 13.0 (IBM-SPSS Inc., Chicago, IL) software was utilized as reported previously [26]. The patients used for model building were divided into two groups at each step of the analysis based on split variables. Each value of each variable was considered as a potential split. The optimum variables and cut-off values were determined by a statistical search algorithm to generate the most significant division into two prognostic subgroups that were as homogeneous as possible for the probability of SVR. Thereafter, each subgroup was evaluated again and divided further into subgroups. This procedure was repeated until no additional significant variable was detected or the sample size was below 15. To avoid over-fitting, 10-fold cross validation was used in the tree building process. The reproducibility of the resulting model was tested with the data from the validation patients.

Results

Association between the IL28B (rs8099917) genotype and the PEG-IFN/RBV response

The rs8099917 allele frequency was 70% for TT (*n* = 345), 29% for TG (*n* = 146), and 1% for GG (*n* = 5). We defined the *IL28B* major allele as homozygous for the major sequence (TT) and the *IL28B* minor allele as homozygous (GG) or heterozygous (TG) for the minor sequence. The rate of NVR was significantly higher (72% vs. 12%, *p* <0.0001) and the rate of SVR was significantly lower (14% vs. 50%, *p* <0.0001) in patients with the *IL28B* minor allele compared to those with the major allele (Fig. 1).

Effect of the IL28B polymorphism, substitutions in the ISDR, Core70, and Core91 of HCV on time-dependent clearance of HCV

Patients were stratified according to their IL28B allele type, the number of mutations in the ISDR, the amino acid substitutions in Core70 and Core91, and the rate of undetectable HCV-RNA at 4, 8, 12, 24, and 48 weeks after the start of therapy was analyzed (Fig. 2A–D). The rate of undetectable HCV-RNA was significantly higher in patients with the IL28B major allele than the minor allele, in patients with two or more mutations in the ISDR compared to none or only one mutation, in patients with arginine (Arg) at Core70 rather than Gln/His, and in patients with leucine (Leu) at Core91 rather than Met. The difference was most significant when stratified by the *IL28B* allele type. The rate of RVR and cEVR was significantly more frequent in patients with the *IL28B* major allele compared with those with the *IL28B* minor allele: 9% vs. 3% for RVR (*p* <0.005) and 57% vs. 11% for cEVR (*p* <0.0001). These findings suggest that IL28B has the greatest impact on early virological response to therapy.

Association between substitutions in the ISDR and relapse after the completion of therapy

Patients were stratified according to the *IL28B* allele, number of mutations in the ISDR, and amino acid substitutions of Core70 and Core91, and the rate of relapse was analyzed (Fig. 3A and B). Among patients who achieved cEVR, the rate of relapse was significantly lower in patients with two or more mutations in the ISDR compared to those with only one or no mutations (15% vs. 31%, *p* <0.005) (Fig. 3 B). On the other hand, the relapse rate was not different between the *IL28B* major and minor alleles within patients who achieved RVR (3% vs. 0%) or cEVR (28% vs. 29%) (Fig. 3A). Amino acid substitutions of Core70 and Core91 were not associated with the rate of relapse (data not shown).

Factors associated with response by multivariate logistic regression analysis

By univariate analysis, the minor allele of *IL28B* (*p* <0.0001), one or no mutations in the ISDR (*p* = 0.03), high serum level of

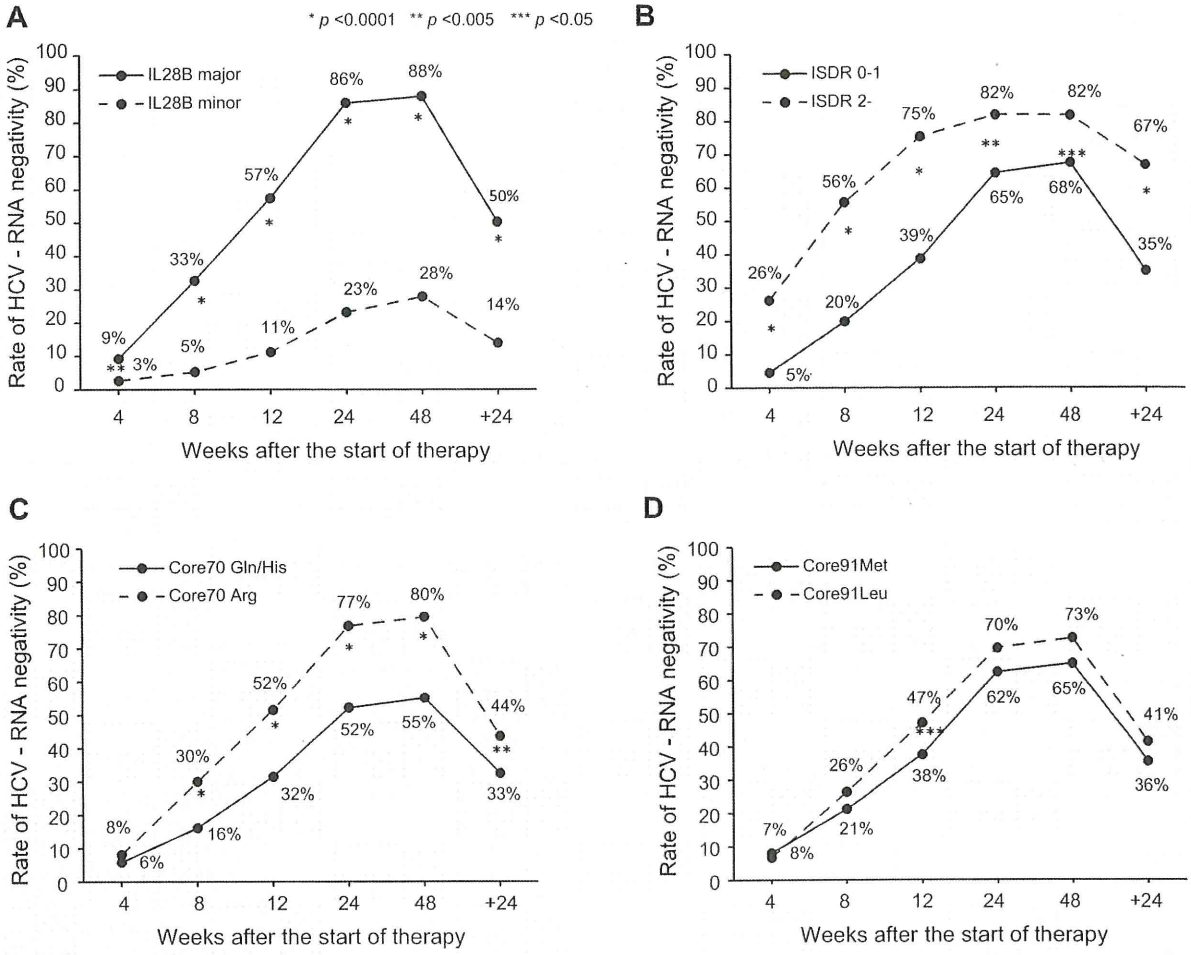


Fig. 2. Effect of *IL28B* mutations in the ISDR, Core70 and Core91 of HCV on time-dependent clearance of HCV. The rate of undetectable HCV-RNA was plotted for serial time points after the start of therapy (4, 8, 12, 24, and 48 weeks) and for 24 weeks after the completion of therapy. Patients were stratified according to (A) the *IL28B* allele (minor allele vs. major allele), (B) the number of mutations in the ISDR (0–1 mutation vs. 2 or more mutations), amino acid substitutions of (C) Core70 (Gln/His vs. Arg), and (D) Core91 (Met vs. Leu). The *p* values are from Fisher's exact test.

HCV-RNA ($p = 0.035$), Gln or His at Core70 ($p < 0.0001$), low platelet counts ($p = 0.009$), and advanced fibrosis ($p = 0.0002$) were associated with NVR. By multivariate analysis, the minor allele of *IL28B* (OR = 20.83, 95%CI = 11.63–37.04, $p < 0.0001$) was associated with NVR independent of other covariates (Table 2). Notably, mutations in the ISDR ($p = 0.707$) and at amino acid Core70 ($p = 0.207$) were not significant in multivariate analysis due to the positive correlation with the *IL28B* polymorphism ($p = 0.004$ for ISDR and $p < 0.0001$ for Core70, Fig. 4).

Genetic polymorphism of *IL28B* also was associated with SVR (OR = 7.41, 95% CI = 4.05–13.57, $p < 0.0001$) independent of other covariates, such as platelet counts, fibrosis, and serum levels of HCV-RNA. Mutation in the ISDR was an independent predictor of SVR (OR = 2.11, 95% CI = 1.06–4.18, $p = 0.033$) but the amino acid at Core70 was not (Table 3).

Factors associated with the IL28B polymorphism

Patients with the *IL28B* minor allele had significantly higher serum level of gamma-glutamyltransferase (GGT) and a higher frequency of hepatic steatosis (Table 4). When the association between the *IL28B* polymorphism and HCV sequences was analyzed, Gln or His at Core70, that is linked to resistance to PEG-IFN and RBV therapy [4,14,15], was significantly more frequent in patients with the minor *IL28B* allele than in those with the major allele (67% vs. 30%, $p < 0.0001$) (Fig. 4). Other HCV sequences with an IFN resistant phenotype also were more prevalent in patients with the minor *IL28B* allele than those with the major allele: Met at Core91 (46% vs. 37%, $p = 0.047$) and one or no mutations in the ISDR (94% vs. 85%, $p = 0.004$) (Fig. 4).

Data mining analysis

Data mining analysis was performed to build a model for the prediction of SVR and the result is shown in Fig. 5. The analysis selected four predictive variables, resulting in six subgroups of patients. Genetic polymorphism of *IL28B* was selected as the best predictor of SVR. Patients with the minor *IL28B* allele had a lower probability of SVR and a higher probability of NVR than those with the major *IL28B* allele (SVR: 14% vs. 50%, NVR: 72% vs.

Please cite this article in press as: Kurosaki M et al. Pre-treatment prediction of response to pegylated-interferon plus ribavirin for chronic hepatitis C using genetic polymorphism in *IL28B* and viral factors. J Hepatol (2010), doi:10.1016/j.jhep.2010.07.037

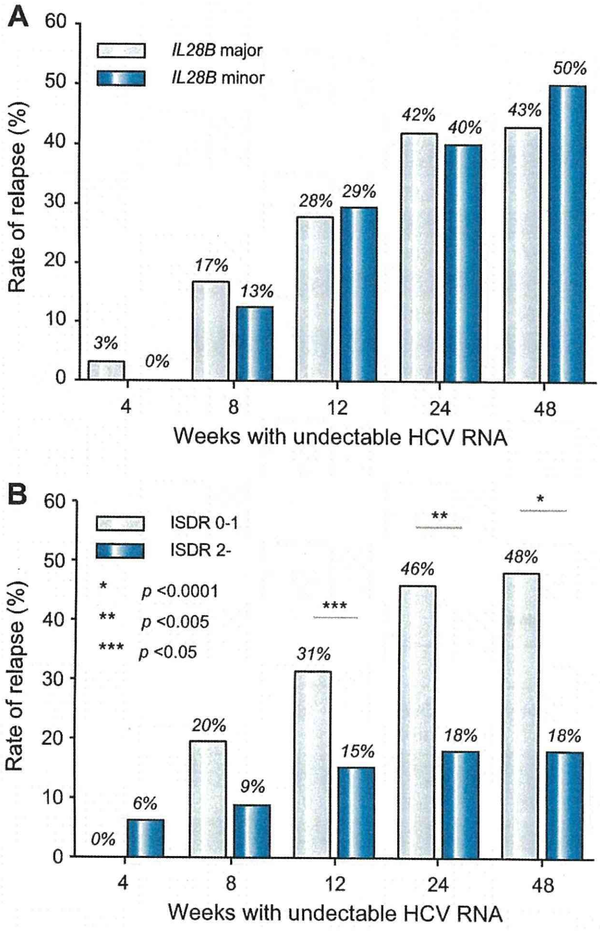


Fig. 3. Association between relapse and the *IL28B* allele or mutations in the ISDR. The rate of relapse was calculated for patients who had undetectable HCV-RNA at serial time points after the start of therapy (4, 8, 12, 24, and 48 weeks). Patients were stratified according to (A) the *IL28B* allele (minor allele vs. major allele) and (B) the number of mutations in the ISDR (0–1 mutation vs. 2 or more mutations). The *p* values are from Fisher's exact test. [This figure appears in colour on the web.]

12%). After stratification by the *IL28B* allele, patients with low platelet counts (<140 × 10^9/L) had a lower probability of SVR and higher probability of NVR than those with high platelet counts (⩾140 × 10^9/L): for the minor *IL28B* allele, SVR was 7% vs. 19%, and NVR was 84% vs. 62%, and for the major *IL28B* allele, SVR was 32% vs. 66% and NVR was 16% vs. 8%. Among patients with the major *IL28B* allele and low platelet counts, those with two or more mutations in the ISDR had a higher probability of SVR and lower probability of relapse than those with one or no mutations in the ISDR (SVR: 75% vs. 27%, and relapse: 8% vs. 57%). Among patients with the major *IL28B* allele and high platelet counts, those with a low HCV-RNA titer (<600,000 IU/ml) had a higher probability of SVR and lower probability of NVR and relapse than those with a high HCV-RNA titer (SVR: 90% vs. 61%, NVR: 0% vs. 10%, and relapse: 10% vs. 29%). The sensitivity and specificity of the decision tree were 78% and 70%, respectively. The area under the receiver operating characteristic (ROC) curve of the model was 0.782 (data not shown). The proportion of patients with advanced fibrosis (F3-4) was 39% (84/217) in patients with low platelet counts (<140 × 10^9/L) compared to 13% (37/279) in those with high platelet counts (⩾140 × 10^9/L).

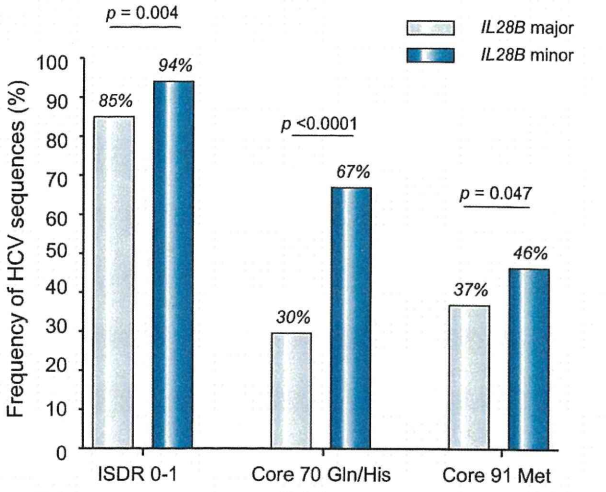


Fig. 4. Associations between the *IL28B* allele and HCV sequences. The prevalence of HCV sequences predicting a resistant phenotype to IFN was higher in patients with the minor *IL28B* allele than those with major allele. (A) 0 or 1 mutation in the ISDR of NS5A, (B) Gln or His at Core70, and (C) Met at Core91. *p* values are from Fisher's exact test. [This figure appears in colour on the web.]

Validation of the data mining analysis

The results of the data mining analysis were validated with 165 patients who differed from those used for model building. Each patient was allocated to one of the six subgroups for the validation using the flow-chart form of the decision tree. The rate of SVR and NVR in each subgroup was calculated. The rates of SVR and NVR for each subgroup of patients were closely correlated between the model building and the validation patients (r^2 = 0.99 and 0.98) (Fig. 6).

Discussion

The rate of NVR after 48 weeks of PEG-IFN/RBV therapy among patients infected with HCV of genotype 1 is around 20–30%. Previously, there have been no reliable baseline predictors of NVR or SVR. Because more potent therapies, such as protease and polymerase inhibitor of HCV [28,29] and nitazoxianide [30], are in clinical trials and may become available in the near future, a pre-treatment prediction of the likelihood of response may be helpful for patients and physicians, to support clinical decisions about whether to begin the current standard of care or whether to wait for emerging therapies. This study revealed that the *IL28B* polymorphism was the overwhelming predictor of NVR and is independent of host factors and viral sequences reported previously. The *IL28B* encodes a protein also known as IFN-lambda 3, which is thought to suppress the replication of various viruses including HCV [31,32]. The results of the current study and the findings of the GWAS studies [6–9] may provide the rationale for developing diagnostic testing or an IFN-lambda based therapy for chronic hepatitis C in the future.

Please cite this article in press as: Kurosaki M et al. Pre-treatment prediction of response to pegylated-interferon plus ribavirin for chronic hepatitis C using genetic polymorphism in *IL28B* and viral factors. J Hepatol (2010), doi:10.1016/j.jhep.2010.07.037

Table 2. Factors associated with NVR analyzed by univariate and multivariate logistic regression analysis.

	Univariate			Multivariate		
	Odds ratio	95%CI	*p* value	Odds ratio	95%CI	*p* value
Gender: female	0.98	0.67-1.45	0.938	1.29	0.75-2.23	0.363
Age	1.01	0.97-1.01	0.223	0.99	0.97-1.02	0.679
ALT	1.00	1.00-1.00	0.867	1.00	0.99-1.00	0.580
GGT	1.004	1.00-1.01	0.029	1.00	1.00-1.00	0.715
Platelets	0.95	0.91-0.99	0.009	0.92	0.87-0.98	0.006
Fibrosis: F3-4	2.23	1.46-3.42	0.0002	1.97	1.09-3.57	0.025
HCV-RNA: ≥600,000 IU/ml	1.83	1.05-3.19	0.035	2.49	1.17-5.29	0.018
ISDR mutation: ≤1	2.14	1.08-4.22	0.030	0.96	0.78-1.18	0.707
Core 70 (Gln/His)	3.23	2.16-4.78	<0.0001	1.41	0.83-2.42	0.207
Core 91 (Met)	1.39	0.95-2.06	0.093	1.21	0.72-2.04	0.462
IL28B: Minor allele	19.24	11.87-31.18	<0.0001	20.83	11.63-37.04	<0.0001

ALT, alanine aminotransferase; GGT, gamma-glutamyltransferase; ISDR, interferon sensitivity determining region; Gln, glutamine; His, histidine; Met, methionine; Minor allele, heterozygote or homozygote of minor allele.

Table 3. Factors associated with SVR analyzed by univariate and multivariate logistic regression analysis.

	Univariate			Multivariate		
	Odds ratio	95%CI	*p* value	Odds ratio	95%CI	*p* value
Gender: female	0.81	0.56-1.16	0.253	0.86	0.55-1.35	0.508
Age	0.97	0.95-0.99	0.0003	0.99	0.96-1.01	0.199
ALT	1.00	1.00-1.00	0.337	1.00	1.00-1.01	0.108
GGT	1.00	1.00-1.00	0.273	1.00	1.00-1.00	0.797
Platelets	1.12	1.01-116	<0.0001	1.13	1.08-1.19	<0.0001
Fibrosis: F0-2	2.64	1.65-4.22	<0.0001	1.87	1.07-3.28	0.029
HCV-RNA: <600,000 IU/ml	2.49	1.55-3.98	0.0001	2.75	1.55-4.90	0.001
ISDR mutation: 2≤	3.78	2.14-6.68	<0.0001	2.11	1.06-4.18	0.033
Core 70 (Arg)	1.61	1.11-2.28	0.012	0.84	0.52-1.35	0.470
Core 91 (Leu)	1.28	0.88-1.85	0.185	1.26	0.81-1.96	0.300
IL28B: Major allele	6.21	3.75-10.31	<0.0001	7.41	4.05-13.57	<0.0001

ALT, alanine aminotransferase; GGT, Gamma-glutamyltransferase; ISDR, interferon sensitivity determining region; Arg, arginine; Leu, leucine; Major allele, homozygote of major allele.

Among baseline factors, IL28B was the most significant predictor of NVR and SVR. Moreover, the IL28B allele type was also correlated with early virological response: the rate of RVR and cEVR was significantly high for the IL28B major allele compared to the IL28B minor allele: 9% vs. 3% for RVR and 57% vs. 11% for cEVR (Fig. 2). On the other hand, the relapse rate was not different between the IL28B genotypes within patients who achieved RVR or cEVR (Fig. 3). We believe that optimal therapy should be based on baseline features and a response-guided approach. Our findings suggest that the IL28B genotype is a useful baseline predictor of virological response which should be used for selecting the treatment regimen: whether to treat patients with PEG-IFN and RBV or to wait for more effective future therapy including direct acting antiviral drugs. On the other hand, baseline IL28B genotype might not be suitable for determining the treatment duration in patients who started PEG-IFN/RBV therapy and whose virological response is determined because the IL28B genotype is not useful for the prediction of relapse. The duration of therapy should be personalized based on the virological response. Future studies need to explore whether the combination of baseline IL28B genotype and response-guided approach further improves the optimization of treatment duration.

The SVR rate in patients having the IL28B minor allele was 14% in the present study while it was 23% in Caucasians and 9% in African Americans in a study by McCarthy et al. [33]. On the other hand, the SVR rate in patients having the IL28B minor allele was 28% in genotypes 1/4 compared to 80% in genotypes 2/3 in a study by Rauch et al. [9]. These data imply that the impact of the IL28B polymorphism on response to therapy may be different in terms of race, geographical areas, or HCV genotypes, and that our data need to be validated in future studies including different populations and geographical areas before generalization.

Please cite this article in press as: Kurosaki M et al. Pre-treatment prediction of response to pegylated-interferon plus ribavirin for chronic hepatitis C using genetic polymorphism in *IL28B* and viral factors. J Hepatol (2010), doi:10.1016/j.jhep.2010.07.037