Coevolution of the Hepatitis C Virus Polyprotein Sites in Patients on Combined Pegylated Interferon and Ribavirin Therapy

James Lara,* Guoliang Xia, Mike Purdy, and Yury Khudyakov*

Molecular Epidemiology & Bioinformatics Laboratory, Laboratory Branch, Division of Viral Hepatitis, Centers for Disease Control and Prevention, 1600 Clifton Road, Atlanta, Georgia 30333

Received 19 October 2010/Accepted 7 January 2011

Hepatitis C virus (HCV) is the major etiologic agent of blood-borne non-A, non-B hepatitis (25). Chronic HCV infection is an established risk factor for the development of liver diseases, such as fibrosis, cirrhosis, and hepatocellular carcinoma (33, 124, 125). Approximately 70% to 80% of HCV-infected patients fail to clear the virus and progress to chronicity (89a). At present, there are no preventive vaccines against HCV. The current, accepted therapeutic approach to treating chronic hepatitis C infection involves a 24- or 48-week course of pegylated alpha interferon (IFN-α) combined with ribavirin (RBV) (i.e., IFN-RBV therapy) (48, 52). Because only 50% to 70% of chronically infected patients develop a sustained virologic response (SVR) to this treatment (48, 55, 80) and because patient intolerance to such therapy is common (61, 68, 120), the development and application of other therapeutic approaches using antiviral compounds that act against HCV more efficaciously and yet generate lower rates of adverse effects are major clinical management and public health objectives. Therapeutic failure presents in two forms: (i) complete resistance to treatment (no response [NR]) and (ii) unsustainable response (UR), which is characterized by an increase in HCV load observed during therapy after an initial period of decline in viral load (breakthrough) or observed after cessation of therapy (relapse) (52).

Several factors are known to affect therapy outcome in HCV-infected patients, most notably the infecting HCV genotype. There are six major HCV genotypes, 1 to 6 (108, 109). Patients infected with genotype 2 are the most responsive, with SVR being achievable in 70% to 80% of cases (52, 80). In contrast, only 50% to 60% of genotype 1-infected patients achieve SVR (48, 55, 80, 90). Genotype 1 is the most prevalent genotype worldwide (78). The dependence of IFN-RBV response rates on HCV genotype (48, 52, 55, 80) implies that the composition of the HCV genome plays a role in influencing therapy outcome.

The mechanism of IFN action against HCV is not fully known. It was shown that treatment with IFN activates the host's innate antiviral immune responses by inducing IFN-stimulated genes (47, 59, 64, 84). Several HCV genomic regions have been found to be associated with resistance to IFN treatment (74). Since responses to IFN differ among HCV strains, associations between IFN therapy outcome and HCV genomic variability in regions such as hypervariable region 1 (HVR1) of E2 (87, 118) and the V3 domain of NS5A (34, 79) have been frequently investigated. A correlation was reported between NR and the high complexity of HVR1 variants before treatment (87, 118), but it was not confirmed in a subsequent study (75). A high level of V3 heterogeneity was associated with IFN sensitivity (34, 99, 119). Specific mutations in the core protein have also been suggested to determine the early response to IFN-RBV therapy (36).

Both E2 and NS5A proteins have been implicated in binding to the IFN-inducible, double-stranded, RNA-activated protein kinase R (PKR), which is involved in the IFN-induced antiviral
response (49). A 12-amino-acid (aa) region located between positions 659 and 670 in E2 known as the PKR-eIF2α phosphorylation homology domain (PePHD) was shown to bind PKR in vitro (113). The PePHD sequence has similarity to the autophosphorylation sites of PKR and the phosphorylation site in eIF2α. This similarity is greater for HCV genotype 1 than genotype 2 or 3. However, the association between PePHD sequence and therapy outcomes has not been consistently shown (1, 98). A PKR-binding domain is located in the C-terminal region of NS5A (49). A variable 40-aa region of this domain, termed the interferon sensitivity determining region (ISDR), was reported to play a key role in the IFN therapy response (37, 38). Analysis of HCV 1b sequences showed an association between the number of ISDR mutations and the response to the IFN therapy (92). However, studies of HCV genotype 2b and 3a did not find such a relation between SVR and NS5A variability (8, 89). Additionally, no binding between PKR and the genotype 3a NS5A from the IFN-resistant HCV strains was observed in vitro (20).

RBV, a guanosine nucleotide analog, is ineffective against HCV when used alone but when combined with IFN therapy dramatically improves viral clearance and decreases relapse rates (42). The mechanism by which RBV improves treatment responses is not well understood. Several mechanisms of its therapeutic action have been proposed, including inosine monophosphate dehydrogenase inhibition (133), viral inhibition (77), facilitation of Th1 immunoresponses (111), mutagenesis (27, 76), inhibition of 5’ cap formation on mRNAs (53), and upregulation of genes involved in IFN signaling (44, 132). However, none of these mechanisms has been convincingly shown to be responsible for its efficacy when combined with IFN (42). Nonetheless, RBV was recently shown to improve early responses to IFN (43), thus supporting its role in enhancing IFN signaling (44, 132) and emphasizing the leading role of IFN in combination therapy.

Host factors have been also found to affect both the natural course of HCV infection and the outcome of treatment (116). For example, common-source HCV infections frequently lead to differential outcomes among incident cases, with some patients resolving the infection and some developing chronic hepatitis C (122, 123), or patients chronically infected with the same genotype respond differently to IFN-RBV treatment despite carrying similar HCV viral loads (55, 103). In addition to genotype, demographic factors such as ethnicity and gender have been associated with therapy outcomes (48, 80, 103). Several studies reported the role of the host genetic polymorphism, e.g., in the IL28B locus, in defining the rate of spontaneous clearance (114) and IFN-RBV SVR (50, 110).

Many host selection pressures, including innate and adaptive immune responses, shape HCV evolution, and their effects should be reflected in HCV genetic composition and epistatic connectivity among genomic sites. Indeed, polymorphic sites within the HCV genome have been shown to be organized as a network of coordinated substitutions (17), with the topology of the network being different for HCV strains that are resistant or sensitive to treatment (7). Although indicating a strong association of many HCV sites with outcomes of therapy, these networks, however, do not provide quantitative measures for viral genomic parameters related to IFN treatment.

In this paper, we report modeling of quantitative associations between a global epistatic connectivity among the HCV polymorphic amino acid sites and UR/NR outcomes of the IFN-RBV therapy. While NR represents complete resistance to IFN-RBV, UR reflects incomplete suppression of HCV or the intrahost HCV evolution toward IFN-RBV resistance (93). Both UR and NR are associated with HCV persistence despite treatment (52). With HCV available for analysis at the start and end of therapy, these outcomes provide an important setting for analyzing genetic changes in the HCV genome associated with resistance.

**MATERIALS AND METHODS**

Sequence data. Analyses were conducted using the HCV 1a full-length polyprotein consensus sequences from 20 patients (10 UR and 10 NR cases) identified through the Viracep-C study (18, 26). Sequences in the Viracep-C study were sampled from patients before (n = 20) and at the end of treatment (n = 20) with pegylated IFN-α2a and RBV. Analyses included all sites from the entire HCV polyprotein except for the most C-terminal 56 aa from the NS5B protein. This sequence data set served as a training set for developing models for prediction of therapy outcomes. For some analyses, a total of 298 HCV 1a full-length consensus polyprotein sequences from GenBank were used. In addition, full-length NS5A protein consensus sequences from 18 treatment-naive patients (6 UR and 12 NR) identified through the HALT-C trial (131) were used as a test data set to validate the predictive models constructed from the Viracep-C data. A full listing of the GenBank accession numbers of all sequences used in this study can be found in the supplemental material.

An alignment of the HCV viral sequences from all three data sets was generated using the Clustal W program (115) implemented in BioEdit v7.0.5.3 (58). HCV H77 (GenBank accession no. AF009060) was used as the reference sequence. In alignment, alignments of consensus sequences for individual gene products were generated using the Viracep-C data. Each amino acid site was numbered according to its position in the HCV polyprotein. For modeling, each sequence was associated with the IFN-RBV therapy outcome, UR or NR. Together, the sequences and assigned therapy outcome attributes constituted the entire set of viral features representing each HCV variant. These viral features of the Viracep-C data were used for modeling dependencies among sites in relation to treatment response.

Conditional independence analysis. Pairwise conditional independencies (CI) among HCV viral features (amino acid sites and therapy outcome) were examined using full-length polyprotein consensus sequences from the Viracep-C study (18, 26). Testing for CI was performed in the form of undirected independence graphs (71), which present the CI among a collection of variables. Nodes in the graph represent the HCV polyprotein sites and therapy outcome, while links between nodes represent dependencies among the features.

The CI testing was used to validate dependency among the polyprotein sites in relation to the therapy outcome. Only polymorphic sites were considered for finding CI from the data. The identified dependency between two features was shown in the graph as a link. This type of statistical analysis assumes the null hypothesis of independence between any two given features. Relative strengths assigned to links in the graph were based on the marginal dependencies between observed associations. Marginal dependence for each link connecting variables A and B was quantified through P value. For each set, C, of conditioning variables, a P value for \{A, B\} was computed, which expresses the probability that A and B are conditionally independent given C. The marginal P value is the value corresponding to \( P(A, B) = 1.0 \) is defined as minus the marginal P value associated with \{A, B\}, where a marginal dependence of 0 means that A and B are completely independent and 1 means that they are completely dependent. The CI among the features was measured at several different levels of significance (thresholds between 0.05 and 5 \( \times 10^{-4} \)). Undirected independence graphs and statistical computations of CI were conducted as implemented in the commercially available software package Hugin Researcher (v6.8).

**Bayesian network (BN).** Relationships among amino acid sites of the HCV polyprotein and therapy outcome were examined using probabilistic graphical models in the form of a Bayesian network (BN) (63), where nodes in the graph represent variables (here, amino acid sites and therapy outcome) and links between the nodes represent relationship. Unlike the undirected independence graphs, BNs provide a more complex notion of the relationships. This includes the notion of the conditional probability and directionality of the relationship.
Links connecting two variables (nodes in the graph) are represented as arcs, which may project toward the node (incoming links) or from the node (outgoing links), thus specifying the direction of influences among variables. Relationships between variables in a BN may be interpreted as causal (22). The conditional probability distributions are represented in the conditional probability tables (CPTs) of the variables (features) in the network. CPTs of the BNs in this study represent amino acid probability distributions at each site and probabilities associated with therapy outcome. Inference of the network structure and parameter estimation (i.e., CPTs) of all BNs constructed for this study was performed through Bayesian artificial intelligence learning algorithms.

BNs were inferred from the Virahep-C data using the HCV polyprotein consensus sequences and associated therapy outcomes. The objectives of analysis with BN were to examine the complexity of the probabilistic interrelationship and measure the importance (or strength) of links among the HCV amino acid sites and therapy outcome (variables). Measurements of the importance of links were used to identify the most influential amino acid sites in the polyprotein BN. The importance of a variable can be estimated using the number and strength of links associated with the corresponding node in the BN. The amino acid sites that most strongly influence the probabilities of the treatment outcome were of a particular interest.

The greedy thick thinning (GTT) method (31) was used to infer the BN structure for the purpose of examining complexity of interrelationships among the variables. The number of incoming links to any given node was constrained to 3 and 10. Parameter estimation of the CPTs was performed using the K2 priors (28) of each variable in the network. Complexity of the probabilistic interrelationships among amino acid sites and therapy outcomes was also examined by individual protein regions. BNs were constructed for each individual protein region using methods as described learning and network parameterization (GTT and the K2 priors, respectively). BNs were constructed using the GeNie software (http://genie.sis.pitt.edu/).

It is important to note that with the increase in the number of variables, the number of possible networks grows superexponentially and computation of the probabilities of all links becomes NP-hard (24). Therefore, a search heuristics method was adopted to compute the strengths of the links in order to derive measures of the importance of relationships among amino acid sites and therapy outcome. The maximum spanning tree (MST) algorithm was used to infer the BN structure from the data.

The strength of the probabilistic relationships (or force of the influences) among variables (amino acid sites and therapy outcome) was inferred by computing the Kullback-Leibler (KL) divergence (69) between the joint probability distribution with and without the link. The greater the KL divergence between these two distributions, the greater the strength of the link, hence, the importance of the relationship it represents. The global importance of an amino acid site was calculated as the sum of strength of incoming and outgoing links associated with the node representing this site in the network. The overall strength of links for individual protein regions and relevance to the therapy outcome was calculated by summing the strength of incoming (incoming strength) and outgoing (outgoing strength) associated with each region.

The relative importance of amino acid sites was measured by the contribution of each amino acid site independently provided to the knowledge of therapy outcome was determined using a naive BN (28) approach. The BN structure was inferred from the Virahep-C data using the MST algorithm. This approach identifies associations between the therapy outcome and amino acid sites, with sites considered to be independent from each other. Mutual information was used to measure contribution of each site to the knowledge of therapy outcome (29). All algorithms based on heuristic methods used here to infer the BN structures as well as computation of link strength and relevance of variables were carried out as implemented in the Professional Edition of BayesiaLab software (Bayesia SAS, Laval, France). The Pearson correlation coefficient was calculated using SAS (version 9.2; SAS Institute Inc., Cary, NC).

Bayesian network classifier (BNC). BNC was constructed for E2 and NS5A. Both BNCs can infer the probabilities of the UR/NR responses to IFN-RBV treatment directly from amino acid sequence. The E2 BNC and NS5A BNC were inferred from the Virahep-C data as follows: (i) the network was initialized as a naive BN (28), where the therapy outcome was directly linked to all amino acid sites; (ii) conditional probabilities for amino acid sites were computed. The K2 learning algorithm (28) was used to infer BN structure. The maximum number of incoming links associated with each node (feature) in BN was constrained to 4. Parameter estimation of CPTs of each feature in the BN was empirically derived from the data.

The NS5A BNC based on the selected amino acid sites was constructed using the hybrid decision table-naive Bayes method (DTNB) (56). The DTNB is a BN where CPTs are represented by a decision table. This method has been shown to perform well when applied with feature selection (56). The DTBN model splits the features into two groups: one group assigns class probabilities based on naive Bayes, and the other assigns probability class based on a decision table. The resulting probability estimates are then combined to estimate the probability of the outcome class association.

Physicochemical properties of HCV variants. Each amino acid can be represented as a set of physicochemical properties. Using these properties, the HCV polyprotein consensus sequences from the Virahep-C data set were converted into the respective physicochemical vectors, which were subsequently used to identify their association with therapy outcome. Analyses were conducted for the HCV polyprotein and individual gene products. Conserved positions were not considered for the physicochemical representation of HCV variants. Position numbering of polymorphic sites was maintained according to the HCV polyprotein. Sequence alignments comprised of polymorphic sites were transformed into $N \times 5$ dimensional numerical vectors, where $N$ is the sequence length and 5 represents the number of physicochemical values assigned to each amino acid site in the sequence. The five physicochemical factors used in this study have been previously described (6). Each vector was then associated with the known therapy outcome (18, 26).

Physicochemical mapping of the data was conducted using a projection pur- suit-based technique in the form of a two-dimensional linear projection (LP) (57). A projection pursuit-based method was used for a combination of the physicochemical vectors (projections) that most accurately separates HCV variants into two classes: UR and NR therapy outcomes. The LP mapping can be tested on new data without having to reconstruct the original mapping (32).

Feature selection was used to identify amino acid sites and their properties most relevant to the therapy outcome-based clustering of the HCV variants. A minimal subset of features (features for structure or properties of a protein) was derived, using a heuristic method (73), to search for “interesting” projections that were most associated with the therapy outcome. Projections were evaluated during the global and local searching that was performed using the $k$-nearest neighbor method ($k=10$) and tested by 10-fold cross validation (10-fold CV) for classification correctness. Correctness estimation was based on the average probability of a projection to be assigned to the correct therapy outcome class. During the global and local searches, 5 x $10^3$ and 3 x $10^5$ projections, respectively, were evaluated.

Feature selection (FS). FS was applied to alignments of the full-length consensus polyprotein sequences and individual gene products of the Virahep-C data to determine which amino acid sites were most associated with therapy outcome. FS reduces dimensionality of the data and improves the prediction performance of BNCs. The usefulness of each amino acid site for the prediction of the therapy outcome was evaluated using FS techniques for ranking or selecting an optimal subset of features. Feature ranking was conducted using divide-and-conquer approaches (decision trees) and information-based metrics. Correlation was used as the filtering metric to search for optimal subsets of features. Given that FS techniques have biases known to affect the variable selection optimization method (30, 54), several FS methods were applied.

Three FS techniques based on information theory were used: information gain (101), Gini gain (16), and gain ratio (101). These methods rank the elevation of information gain obtained on a score that is relative to the contribution of each amino acid site independently provided to the knowledge of therapy outcome. For the UR/NR therapy outcome was evaluated using FS techniques for ranking or selecting an optimal subset of features. Each feature receives a score based on its degree of correlation to the class variable (therapy outcome) and low intercorrelation between features. This method was used to search for a minimal subset of complementary amino acid sites to improve the BNC accuracy.

Evaluation and validation of the therapy outcome predictors. The E2 and NS5A BNCs were evaluated by 10-fold CV. Briefly, the HCV variants represented by all polymorphic sites or selected amino acid sites from E2 or NS5A were randomly divided into 10 parts of equal size. Each part was held out strictly as a testing data set to evaluate the prediction accuracy of the BNC trained with the remaining nine parts of the data. This process was executed until the BNC was evaluated with all 10 parts. The 10 accuracy estimates were then averaged to estimate the overall accuracy of the BNC.

Also, BNCs trained with data sets—where the E2 and NS5A protein sequences were randomly assigned with UR/NR outcome—were evaluated for prediction accuracy. The results were then compared to the accuracy obtained from the BNCs trained with the correct outcome assignment in order to account for any random statistical correlations present in the Virahep-C data.
Two measures of accuracy were used for classification performance: overall percent classification correctness and precision. The overall percent correctness was measured as \[
\frac{\text{no. correctly classified instances}}{\text{total no. of instances}} \times 100\%.
\]
Precision was determined in the following manner (where \(TP\) is the number of true positives, \(TN\) is the number of true negatives, \(FP\) is the number of false positives, and \(FN\) is the number of false negatives): precision = \[
\frac{TP}{TP + FP} \times 100\%.
\]

The validation of the NS5A predictive models was conducted using the consensus sequences of the NS5A protein from the HALT-C study (131), which were not part of any of the analyses described herein. Estimation of the NS5A BNC and NS5A-LP accuracy of prediction of treatment outcome for HCV NS5A variants from the HALT-C study was based on the overall percent classification correctness.

RESULTS

Complex interdependence between polymorphic sites and therapy outcome. CI analysis of interdependencies between those polymorphic amino acid sites and their potential linkage to the UR/NR outcome of IFN-RBV therapy was conducted using 40 HCV full-genome sequences obtained before and at the end of therapy from 10 UR and 10 NR patients in the Virahep-C data. Feature pairs whose dependencies exceed the threshold are linked. HCV polypeptide sites are grouped by region; from left to right: core, E2, NS2, NS4A, and NS5A (upper row), and E1, P7, NS3, NS4B, NS5B (lower row). Therapy outcome is shown as a single node at the top of the graphs. (a) Initial explorative search for significant dependencies \((P < 0.05)\) followed by gradual decrease in thresholds: \(0.0032\) (b), \(3 \times 10^{-4}\) (c), \(6 \times 10^{-5}\) (d), and \(5 \times 10^{-6}\) (e). (f) Reduced diagram of summarized dependencies between HCV polypeptide sites and treatment outcome \((P \leq 0.003)\).

FIG. 1. Undirected independence graphs showing relative strengths of the dependencies (links in the graph) found among HCV polypeptide sites (nodes in the graph) and UR/NR outcomes following IFN-RBV therapy from 40 sequences obtained from 10 UR and 10 NR patients in the Virahep-C data. Feature pairs whose dependencies exceed the threshold are linked. HCV polypeptide sites are grouped by region; from left to right: core, E2, NS2, NS4A, and NS5A (upper row), and E1, P7, NS3, NS4B, NS5B (lower row). Therapy outcome is shown as a single node at the top of the graphs. (a) Initial explorative search for significant dependencies \((P < 0.05)\) followed by gradual decrease in thresholds: \(0.0032\) (b), \(3 \times 10^{-4}\) (c), \(6 \times 10^{-5}\) (d), and \(5 \times 10^{-6}\) (e). (f) Reduced diagram of summarized dependencies between HCV polypeptide sites and treatment outcome \((P \leq 0.003)\).
resistant as undirected links or edges. The undirected graph displayed numerous links representing a dense and complex network of dependencies $(P < 4 \times 10^{-8})$ between amino acid polymorphic sites across the entire HCV polyprotein and therapy outcome. A large number of links among sites within and between individual proteins remained present up to a threshold value of $2 \times 10^{-8}$. E2 protein sites formed the strongest dependencies. For example, site 612 of E2 is strongly connected to site 233 in E1 ($P = 3 \times 10^{-8}$), and site 642 of E2 to site 1756 in NS4B ($P = 2 \times 10^{-8}$). Also, sites 482 and 612 are strongly connected to site 642 ($P = 2 \times 10^{-8}$). It is important to note that links connecting amino acid sites to therapy outcome were among the strongest ($P \leq 7 \times 10^{-5}$). As shown in Fig. 1, therapy outcome was strongly linked ($P \leq 0.003$) to amino acid sites from the E1 (site 242), E2 (sites 397, 434, 524), P7 (site 790), NS3 (site 1090), NS5A (sites 2280, 2283, 2320, 2366, 2411, 2413, and 2414), and NS5B (sites 2530, 2633, 2730, and 2747) regions. The strongest dependencies were found with sites from P7 (site 790; $P = 2 \times 10^{-8}$), NS5A (site 2280 and 2283; $P = 3 \times 10^{-8}$ and $P = 7 \times 10^{-5}$, respectively), and NS5B (site 2633; $P = 9 \times 10^{-4}$). These data suggest strong coordination of substitutions at sites along the entire HCV polyprotein and association between polymorphic sites and therapy outcome.

**Contribution of different proteins to therapy outcome.** To infer a more insightful representation of the relationships among polymorphic sites and therapy outcome, a Bayesian network (BN) approach (63) was used. The complexity of relationships among HCV polymorphic sites and UR/NR outcome was evaluated by inferring BNs from the full-length HCV polyprotein consensus sequences. The properties of the network are listed in Table 1. In concordance with the undirected interdependence graph findings, interrelationships among all polymorphic sites were found to be highly complex. Figure 2 shows the structure of the polyprotein BN containing 551 polymorphic amino acid sites and their association to therapeutic outcome. Although all sites are interdependent, the number of links broadly varies from 1 to 30 among sites. Sites contributing into this polyprotein BN are listed in Table S1 in the supplemental material.

As shown in Table 1, HCV proteins do not contribute equally to the network topology. The E2 and NS5A regions of the HCV polyprotein are the two major contributors of sites into the HCV polyprotein BN (21% and 18% of amino acid sites, respectively). The E1, E2, and NS5A regions are also major contributors of links into the network (25.3%, 42%, and 26.8% of all links, respectively). The majority of links are between proteins, with only 17.5% of all links being within individual protein regions. Among all E2 links, 18.9% are among E2 sites, whereas all other proteins contain only 1.4% to 10.5% of intraprotein links. Owing to the large number of polymorphic sites contributing to the network, the E2 and NS5A proteins are extensively connected to each other and to all other proteins. As shown in Fig. 3, ~20% of all sites from these two proteins have direct links to NS5A, and ~35% of all NS5A sites have direct links to E2 in the polyprotein BN, indicating a significant coordination of substitutions between these two proteins.

Despite generating many connections ($n = 554$) and contributing many sites ($n = 118$) to the polyprotein BN (Table 1), E2 does not have direct links to therapy outcome. Only six sites form such direct connections, with two sites (at positions 864 and 934) being from NS2, a single site (at position 1841) from NS4B, two sites (at positions 2280 and 2283) from NS5A, and a single site (at position 2633) from NS5B.

The core protein contributes only 11 sites (1.8% of all sites) but 136 links (10.3% of all links) to the polyprotein BN, with each site being connected to ~13 other sites, which is ~3 to 8 times more than the individual sites from any other HCV protein (Table 1). The E1 sites contain 4.5 connections on average, while sites of all other proteins are linked on average to 1.5 to 2.7 other sites. The essential difference is in the directionality of links among proteins. Two proteins, core and E1, located at the N terminus of the HCV polyprotein, have 92% and 69.7% of their links directed outside, respectively, suggesting their important causal role in defining states of many polyprotein sites connected to these two proteins.

**TABLE 1. Properties** of the HCV polyprotein BN

<table>
<thead>
<tr>
<th>Region</th>
<th>No. of sites</th>
<th>Outdegree</th>
<th>Indegree</th>
<th>Max. outdegree</th>
<th>Avg outdegree</th>
<th>Max. indegree</th>
<th>Avg indegree</th>
<th>No of links</th>
<th>To response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core</td>
<td>11</td>
<td>136</td>
<td>11</td>
<td>30</td>
<td>12.36</td>
<td>3</td>
<td>1.22</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>E1</td>
<td>65</td>
<td>248</td>
<td>113</td>
<td>24</td>
<td>6.53</td>
<td>7</td>
<td>2.17</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>E2</td>
<td>118</td>
<td>239</td>
<td>419</td>
<td>16</td>
<td>2.99</td>
<td>9</td>
<td>3.58</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>P7</td>
<td>17</td>
<td>45</td>
<td>39</td>
<td>14</td>
<td>4.09</td>
<td>6</td>
<td>2.29</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>NS2</td>
<td>67</td>
<td>151</td>
<td>154</td>
<td>19</td>
<td>3.51</td>
<td>7</td>
<td>2.30</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>NS3</td>
<td>73</td>
<td>135</td>
<td>144</td>
<td>17</td>
<td>3.00</td>
<td>7</td>
<td>2.03</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>NS4A</td>
<td>8</td>
<td>19</td>
<td>17</td>
<td>10</td>
<td>6.33</td>
<td>5</td>
<td>2.13</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>NS4B</td>
<td>30</td>
<td>71</td>
<td>57</td>
<td>12</td>
<td>3.38</td>
<td>7</td>
<td>1.97</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>NS5A</td>
<td>100</td>
<td>166</td>
<td>216</td>
<td>27</td>
<td>3.00</td>
<td>7</td>
<td>2.24</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>NS5B</td>
<td>72</td>
<td>108</td>
<td>142</td>
<td>12</td>
<td>2.55</td>
<td>6</td>
<td>2.17</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Polyprotein</td>
<td>551</td>
<td>1318^b</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>30</td>
<td>3.77</td>
</tr>
</tbody>
</table>

^a Properties correspond to a polyprotein BN inferred for the Viralhep-C HCV variants.

^b The number of incoming links to any given node was constrained to a maximum of 10. The node count represents the total number of polymorphic sites from each polyprotein region that contributed to the BN. Outdegree, total number of outgoing links from all sites in each protein region. Indegree, total number of incoming links to sites in each region. Max. outdegree, maximum number of outgoing links from any one site from the respective protein region. To response, number of links with direct relationship to the outcome in the BN.

^c Total number of links in the network.
other proteins have almost equal measures of incoming (in-degree) and outgoing (out-degree) links.

Many essential properties of the polyprotein BN constructed using the 40 Virahep-C sequences, except for linkage to therapy, were observed with another BN constructed using HCV genotype 1a full-length genome sequences obtained from GenBank (n = 298). As shown in Fig. 3 and 4, the GenBank BN and Virahep-C BN have similar distributions of links, and interrelationships among individual proteins are highly correlated (r = 0.99, P < 0.0001), indicating that the overall coordination among substitutions in the HCV genotype 1a data set has been adequately represented by the Virahep-C sequences used in this study. However, variations in the number of polymorphic sites are observed between the GenBank BN (n =
and polyprotein BN (n = 551). Despite the greater number of polymorphic sites in the GenBank sequences, the Virahep-C sequences contain 25 unique polymorphic sites distributed among all but core proteins: at positions 230, 349, and 381 in E1; 385, 582, 631, and 742 in E2; 768 in P7; 826 and 926 in NS2; 1385, 1461, 1520, 1528, 1565, and 1592 in NS3; 1681 in NS4A; 1805, 1820, and 1846 in NS4B; 2003, 2049, and 2343 in NS5A; and 2500 and 2548 in NS5B. These findings—in conjunction with the observation of the 1.7-fold increase in the number of links between sites in E1 and E2 and the 2-fold increase between sites in E2 and NS5A in the Virahep-C BN compared to the GenBank BN (Fig. 3)—suggest the treatment-specific variations in coordination of substitutions at the genomic sites in the UR/NR HCV strains.

Protein sites relevant to therapy outcome. Observation of a significant interconnection and coordination among HCV proteins suggests that all proteins contribute to determining the UR/NR outcome. To analyze these contributions in more detail, BNs were constructed for the individual gene products. Extensive dependencies between sites and association to therapy outcome were found in all individual polyprotein regions, albeit to different degrees. The E2 and NS5A regions were found to form a more dense set of links than other regions of the polyprotein (Table 2).

Although many polymorphic sites were found to be interlinked in the model shown in Fig. 1, indicating a significant coordination of heterogeneity along the HCV polyprotein, there are a large number of sites with very few links, suggesting their marginal contribution to the polyprotein BN. To evaluate which proteins and amino acid sites were most associated with the outcome, we conducted feature selection experiments. By using a naïve Bayesian network with feature selection, the E2 and NS5A polyprotein regions were found to contribute the greatest number of sites relevant to outcome (27.5% and 26.3%, respectively) (Fig. 5). Similar results were observed with four filtering methods for feature selection (Fig. 6). Each of the feature selection techniques extracted a certain number of the most relevant sites. A greater proportion of sites were selected from E2 and NS5A as relevant to the outcome (Fig. 6). The NS5A region consistently contributes a large number of relevant sites with all four feature selection techniques. Depending on the technique, 14.3% to 32% and 24.0% to 44% of amino acid sites were, respectively, selected from E2 and NS5A as contributing to the outcome. All of the techniques used selected significantly overlapping sets of the relevant amino acid sites from all proteins, albeit with variations in ranking among the selected sites (see Table S2 in the supplemental material). A set of sites selected using one of the techniques is shown in Table 3.

Relationships between variables in a BN may be interpreted as being causal (22), which can be applied to detect relevance of a variable to define a target feature, in this case, therapy outcome. Analysis of the strength of influence measured as the Kullback-Leibler divergence (69) between the joint probability distribution with and without the arc shows that sites from the E2 and NS5A proteins have the strongest overall influences on

![FIG. 4. Relative strength and direction of links associated with individual HCV proteins in the Virahep-C BN (A) and GenBank BN (B). The total strength of all outgoing links (blue bars), incoming links (red bars), and the global strength (green bars) are shown for each protein.](image-url)
outcome (Fig. 7). Additionally, analysis of contribution of individual sites to the UR/NR outcome was conducted using a ratio of the mutual information calculated for each site and the outcome over the maximal mutual information (MI) \( \frac{\text{MI}}{H_1} = 0.3951, \ P = 0.0001 \) was calculated for site 2283 in the NS5A protein. Using this ratio as a measure of the relative significance of each site for determining outcome, 25 sites were identified in six proteins with values for this ratio being \( >0.5 \) (Fig. 8). Among these sites were one site at position 242 in E1, eight sites at positions 391, 394, 397, 400, 401, 434, 528, and 655 in E2, two sites at positions 753 and 790 in P7, one site at position 941 in NS2, nine sites at positions 2153, 2198, 2280, 2288, 2320, 2339, 2375, 2376, and 2413 in NS5A, and four sites at positions 2633, 2730, 2747, and 2755 in NS5B. The important observation from this analysis is that E2 and NS5A together contain \( \sim 70\% \) of these highly relevant sites. Another interesting observation is that hypervariable region 1 (HVR1) contributes five of eight relevant sites in E2, thus suggesting that HVR1 heterogeneity is associated with HCV evolution toward the IFN-RBV resistance.

Association of protein physicochemical properties with IFN-RBV resistance. The observation of coordinated substitutions in all HCV proteins suggests extensive interrelationships among phenotypic traits encoded by these proteins and an important role of these interrelationships in defining HCV evolution toward IFN-RBV resistance. Although not clearly determined, these phenotypic traits can be further analyzed using amino acid physicochemical properties as a quantitative approximation to phenotype. The factors affecting sequence variation and diversity should be also reflected in the physicochemical properties of the HCV polyprotein. Herein, the physicochemical properties of the HCV polyprotein were analyzed using amino acid properties calculated for each position in the polyprotein sequence.

**TABLE 3.** Correlation-based feature selection (CFS) of HCV sites relevant to UR/NR outcomes\(^a\)

<table>
<thead>
<tr>
<th>Feature selection</th>
<th>No. of relevant sites</th>
<th>Polyprotein positions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core</td>
<td>7</td>
<td>29, 48, 75, 81, 106, 147, 161</td>
</tr>
<tr>
<td>E2</td>
<td>15</td>
<td>394, 397, 434, 478, 480, 490, 498, 524, 528, 534, 591, 595, 625, 655, 668</td>
</tr>
<tr>
<td>P7</td>
<td>8</td>
<td>762, 763, 767, 768, 770, 777, 789, 790</td>
</tr>
<tr>
<td>NS2</td>
<td>14</td>
<td>814, 824, 841, 843, 873, 934, 938, 941, 957, 958, 962, 982, 1017, 1021</td>
</tr>
<tr>
<td>NS3</td>
<td>23</td>
<td>1068, 1087, 1088, 1090, 1115, 1124, 1145, 1148, 1196, 1200, 1239, 1266, 1306, 1366, 1398, 1405, 1409, 1412, 1417, 1428, 1444, 1461, 1592</td>
</tr>
<tr>
<td>NS4A</td>
<td>5</td>
<td>1681, 1686, 1687, 1693, 1700</td>
</tr>
<tr>
<td>NS4B</td>
<td>8</td>
<td>1737, 1753, 1759, 1804, 1816, 1841, 1941, 1968</td>
</tr>
<tr>
<td>NS5A</td>
<td>9</td>
<td>2024, 2043, 2283, 2320, 2366, 2376, 2411, 2414</td>
</tr>
<tr>
<td>NS5B</td>
<td>10</td>
<td>2501, 2530, 2582, 2629, 2633, 2730, 2747, 2755, 2796, 2825</td>
</tr>
<tr>
<td>Polyprotein</td>
<td>14</td>
<td>524, 790, 1090, 1409, 1592, 2024, 2280, 2283, 2366, 2376, 2414, 2530, 2633, 2950</td>
</tr>
</tbody>
</table>

\(^a\) List of subset of amino acid sites relevant to outcome prediction. Subsets of sites from each region were determined by filtering out less-predictive sites. CFS was applied to data sets: 10 data sets representing sites from each individual gene product and the therapy outcome and 1 data set representing the full-length HCV polyprotein sequences and associated therapy outcome. Site subsets listed for the E2 and NS5A proteins were used in E2-BNC and NS5A-BNC (Fig. 10).
The chemical space dispersion of the HCV variants from the UR/NR Virahep-C cases (18) was examined using a linear projection technique (32). The analysis was conducted using polymorphic sites of the HCV polyprotein or individual gene products (see Table S1 in the supplemental material). The polymorphic sites from each protein were converted into vectors of amino acid physicochemical properties (6). For each protein, these vectors were used to generate a multidimensional physicochemical space and project this space into the optimized linear two-dimensional (2D) spaces. The probabilistic mapping of NR and UR outcomes in these 2D physicochemical spaces is shown in Fig. 9.

This analysis showed that the probability of outcomes mapped in the optimized 2D physicochemical spaces of the polyprotein, E2, and NS5A was distributed in the least convoluted way, providing almost equal representations of UR and NR (Fig. 9). These observations suggest that the physicochemical properties of all HCV proteins are related to outcome, albeit to various degrees.

**Strong association of E2 and NS5A with IFN-RBV resistance.** The results shown above strongly suggest that the IFN-RBV resistance is encoded in many regions of the HCV polyprotein, with E2 and NS5A being strongly linked to this resistance. To further investigate the strength of association of the IFN-RBV resistance with variation in the E2 and NS5A primary structure, BN classifiers (BNCs) were developed using polymorphic sites from these two proteins. The accuracy of performance of the models was evaluated using the 10-fold CV protocol. The results of the evaluation are shown in Fig. 10. The E2 and NS5A BNCs constructed using all polymorphic sites were found to be 82.5% and 90% accurate in the prediction of outcomes in the 10-fold CV, respectively. BNCs constructed using 15 sites selected from E2 and 9 sites selected from NS5A (Table 3) improved accuracies to 85% and 97.5%, respectively, while the randomized data sets produced BNCs showing accuracies of only 35% to 47.5% (Fig. 10). Thus, although the networks of sites from both proteins have a strong
association with the IFN-RBV resistance, the NS5A BNCs significantly outperformed the E2 BNCs in the CV experiments.

**Prediction of UR/NR outcomes using NS5A.** A high accuracy of the BNC models described above suggests a strong association of coordinated substitutions in NS5A with evolution toward the IFN-RBV resistance. However, since these models were generated using only 40 sequences from 20 patients, it is critical to demonstrate that the interrelationships identified for these patients are representative of those for other patients. For this purpose, two predictive models were constructed using the same Virahep-C data set and tested using the HCV NS5A sequences from baseline specimens obtained from patients in the HALT-C study (131). Because no additional data were available for E2 from patients with NR and UR outcomes investigated in a single study, only the NS5A models were validated.

One model was constructed using physicochemical properties of five NS5A sites selected using a heuristic method (73). The secondary structure for sites at position 2153 (projection X167 in Fig. 11) and 2413 (X492), the electrostatic charge for site at position 2198 (X195), the polarity for site at position 2280 (X281), and the molecular volume or size for site at position 2320 (X328) were selected as the most relevant features for outcome in the Virahep-C data set. The LP model mapping UR and NR outcomes into the 2D space generated using linear projection from the 5D physicochemical space is shown in Fig. 11. Another model was constructed as a hybrid between the decision table and a naïve Bayes (DTBN)-based machine-learning technique (56) using 12 NS5A sites: nine shown in Table 3 and three additional sites, at positions 2153, 2198, and 2413, used in the linear projection approach.

After a 10-fold CV, both Virahep-C models were tested on the HALT-C data set with 6 NS5A sequences obtained from UR and 12 from NR patients. The hybrid DTBN model showed an overall accuracy of 72.2% and the linear projection model showed an overall accuracy of 83.3% of outcome prediction for the HALT-C patients (Table 4). This finding suggests that, although many sites along the entire HCV polyprotein are relevant to development of the IFN-RBV resistance, the small number of features from the NS5A protein alone may be sufficient for the prediction of therapy outcomes.

**DISCUSSION**

Two important features of HCV infection, persistence following primary infection and resistance to IFN-based therapy, have been related to the extensive HCV genetic variability (39, 41). Although HCV has developed a very efficient capacity to escape from adaptive (15, 35, 104, 128) and innate immune responses (12, 13, 85, 126), ~20% to 30% of all HCV infections are cleared by the host (23) and 50% to 70% of chronic infections can be successfully treated with IFN-RBV (48, 52, 55, 80). The variation in response to therapy among HCV strains remains poorly understood. However, differential sensitivity of HCV genotypes to IFN therapy (52, 80) suggests that viral genetic factors play an important role in determining therapy outcomes. Despite a low degree of response to treatment during chronic infection, 80% to 98% of patients with acute HCV infection can achieve complete virological response to IFN therapy (51, 62), suggesting that HCV acquires a significant degree of IFN resistance during chronic infection. Taken together, these observations indicate a strong connec-

![FIG. 10. 10-fold CV performance of the E2 BNC and NS5A BNC constructed using all polymorphic sites (black bar) and selected relevant sites (white bar). Results for BNCs with randomized labels are shown using patterned bars (black for all and white for selected sites).](image-url)

![FIG. 11. Projection of five selected physicochemical features of five NS5A sites from the HALT-C sequence data set onto the physicochemical space-based model derived from the Virahep-C sequence data set. Lines originating from the center of the graph are projections of five physicochemical features. Circles in the graph map the UR/NR outcomes of therapy for Virahep-C (unfilled circles) and HALT-C (filled circles). For color coding, see legend to Fig. 9.)](image-url)

**TABLE 4. Validation of the NS5A Virahep-C models using the HALT-C NS5A sequences**

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Evaluation</th>
<th>Validation (% accuracy)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Overall</td>
<td>NR class</td>
</tr>
<tr>
<td>BNC</td>
<td>97.5b</td>
<td>72.2</td>
</tr>
<tr>
<td>LP</td>
<td>95.2</td>
<td></td>
</tr>
</tbody>
</table>

b Shown are the overall prediction accuracies of the BNC (DTBN method) and physicochemical-based LP models using selected NS5A sites (see text for details).

c Average percent classification correctness in 10-fold CV.
tion between the intrahost HCV evolution and success of the IFN-RBV therapy.

In the current study, an integrative approach was implemented for the evolutionary analysis of the HCV genome. This approach was based on modeling interrelationships between polymorphic sites along the entire HCV polyprotein and relating the modeled coordination among amino acid substitutions to the UR/NR outcomes of therapy. Models constructed here showed an extensive interdependence of all polymorphic sites within the HCV polyprotein, suggesting a significant co-evolution among individual HCV proteins. The data indicate that all HCV proteins contain sites coordinating their polymorphism with sites in all other proteins (Fig. 2). A similar observation has been recently made using a correlation network analysis of the HCV genotype 1a full-genome sequences from untreated patients (17) and patients on therapy (7). Among all connections identified using the polyprotein BN in this study, only 17.5% were among sites within individual proteins. It is interesting to note that E2 shows the most extensive coordination among its sites, with all other proteins having ~2 to 13 times fewer connections among intraprotein sites than E2. With 82.5% of all connections in the network being among proteins, HCV evolution is evidently defined by coadaptation among many phenotypic traits encoded by different HCV proteins.

Although all HCV proteins contribute to the network, the topological properties of sites differed among proteins. The core protein contributes fewer sites (n = 11) per its size than any other HCV protein. However, each core site forms ~2 to 4 times more links in the network than any site from other proteins (Table 1). This protein has 12.4 times more outgoing than incoming links in the polyprotein BN, while the ratio between outgoing and incoming links for all other proteins varies from 0.8 to 2.2 (Table 1). Another important feature of core connectivity in the polyprotein BN is that 98.6% of all core links are with other proteins. The presence of only two intraprotein links (polyprotein positions 90 and 47) makes the core protein the least intracorrelated protein, indicating a minimal direct coordination among core polymorphic sites. Thus, the contribution of core to the network topology differs considerably from those of all other proteins, suggesting that this protein has a unique role in coordinating substitutions and defining heterogeneity at many sites of the HCV polyprotein.

This observation is in agreement with the multitude of functions performed by the core protein and emphasizes its important role in HCV infection. In addition to forming the nucleocapsid (105), this protein was shown to interfere with many cellular signaling pathways involved in apoptosis (134), transcription (60, 130), and transformation (21, 65, 102, 129). The core protein is also involved in lipid metabolism (10, 96). It inhibits the microsomal triglyceride transfer protein, binds to apolipoprotein AI, and induces accumulation of cytoplasmic lipid droplets (2). Core and NS5A are key factors for assembly of infectious particles. Both colocalize on the surface of lipid droplets, a proposed site for HCV particle assembly (4). With lipid droplets playing a crucial role in the assembly and release of infectious HCV particle (83), interactions involving domain 2 of core and domain 3 of NS5A (5, 14, 81, 82) are essential for virion production and, therefore, have a strong impact on infectivity and viral fitness. Mutation at position 147 in domain 2 of the core protein was found to affect adherence of core to lipid droplets and virus production (107). Our data show that this site has direct links in the polyprotein BN to sites in E1, E2, NS2, and domain 3 of NS5A. Another site, from domain 2 of core at position 161, linked to P7 in addition to these four proteins. All of these proteins play a role in the membrane-associated viral replication (86). These observations suggest coordination of heterogeneity across the HCV polyprotein related to viral production and the important role played by the core protein in this coordination.

Two proteins, E2 and NS5A, together contribute ~40% of all sites and ~62% of all links to the polyprotein BN and, therefore, essentially define the state of this entire network. In combination with E1, these three proteins contribute ~50% of all sites and ~77% of all links to the polyprotein BN. It is interesting that E2 and NS5A also mutually coordinate their heterogeneity (Fig. 3). Although coordination between sites from any two HCV proteins is a common feature of the polyprotein BN, this coordination is most extensive between sites of E2 and NS5A, owing to the large number of sites contributed by these two proteins to the network. Thus, the states of many sites in one of these two proteins reflect the states of many sites in the other protein, suggesting a high degree of coevolution between these two proteins. Additionally, it was observed that sites from E2 formed the strongest links with many other sites in the polyprotein as determined by CI testing (Fig. 1), among which were links between sites 482 and 642 in E2 (P = 2 × 10−10), 612 in E2 and 233 in E1 (P = 3 × 10−8), and 642 in E2 and 1756 in NS4B (P = 2 × 10−8). Taking into consideration that site 482 is from the CD81-binding region (45, 127) and site 612 from one of two E2 regions proposed to be involved in the viral fusion process (72, 91, 95), we speculate that the tight coordination between sites 482 and 642 as well as that between sites 612 and 233 is associated with viral entry.

Another important observation made in this study is that all HCV proteins have association with the UR/NR outcome of IFN-RBV therapy. Taking into consideration the aforementioned extensive linkage among polymorphic sites from different proteins, this observation, although not surprising, reveals that the HCV response to immunomodulatory therapy is a very complex trait involving numerous viral functions that require coordination. All networks constructed for individual proteins included the UR/NR outcome as a variable (Table 2). However, this observation cannot be unequivocally interpreted in terms of equal contribution of each protein to the IFN-RBV response. Nevertheless, it suggests that the genome-wide coordination among sites is important for this response, with some proteins possibly playing accessory roles and reflecting the IFN-RBV-related changes in other proteins that are mainly responsible for resistance. The analysis conducted here revealed that sites substantially associated with the outcome are scattered along the entire HCV polyprotein. Among the sites with relevant significance of >0.5 (Fig. 8) are sites in E1 (n = 1), E2 (n = 8), p7 (n = 2), NS2 (n = 1), NS5A (n = 9), and NS5B (n = 4). Two proteins, E2 and NS5A, shared 68% of these 25 sites, suggesting their strong connection to IFN-RBV resistance. E2, NS5A, and p7 have, respectively, 6.8%, 9.0%, and 11.7% of their polymorphic sites being highly rele-
vant to the therapy outcome, while all other proteins have only 1.5% to 3.1% of these sites.

One surprising finding was that five among the eight sites most relevant to therapy outcome are located in HVR1 of the E2 protein (aa 384 to 410), emphasizing a strong connection of HVR1 heterogeneity to IFN-RBV resistance. Association of HVR1 sites with outcomes of therapy can be also found in the correlation networks (7). However, the significance of these observations is not apparent. Analysis of HVR1 connectivity in the polyprotein BN showed that polymorphic HVR1 sites have a total of 140 links to all HCV proteins, with each HVR1 site being connected to three to nine sites in the HCV polyprotein. Such an extensive interdependence of HVR1 sites with many sites across the entire HCV polyprotein (Fig. 3), in conjunction with the earlier similar observations using network analysis (17), suggests that the HVR1 substitutions are not random and that HVR1 evolution is substantially coordinated with all HCV proteins. Coordination of HVR1 heterogeneity is especially noticeable with E1, E2, and NS5A, which share, respectively, 15%, 26.4%, and 14% of all HVR1 links in the polyprotein BN, while any other HCV protein shares 3.6% to 9.3% of HVR1 links.

HVR1 contains antigenic epitopes (66, 67, 112, 121) with HCV neutralizing activity (40). Rapid HVR1 evolution is associated with immune escape (70). However, the conservation of the HVR1 physicochemical properties and conformation (94) argues that this region is significantly functionally constrained despite its extensive heterogeneity. The observation that compensatory mutations in the ectodomain of E2 (46) and the I347L mutation in E1 compensate for HCV fusion impairment (9) in HCV mutants whose HVR1 have been excised suggests potential functional relationships of this region with other parts of the HCV genome. HVR1 was shown to be involved in the SR-B1-facilitated entry of HCV pseudoparticles in cell culture (11). It was suggested that HVR1 plays an important role in HCV entry by modulating receptor recognition and affects lipoprotein composition and infectivity of viral particles (9). HVR1 heterogeneity was also associated with the development of resistance to therapy (74, 87, 117, 118). We hypothesize that complex functional relationships of HVR1 are reflected in coordinated evolution with other HCV proteins and that HVR1 mirrors the evolution of the entire HCV genome, including evolution toward the IFN-RBV resistance.

There are many sites from different HCV proteins strongly linked to the IFN-RBV resistance (Fig. 1 and 8). However, consideration of individual sites allows only for the identification of connections to the therapy outcome in the form of a trend and does not have a strong predictive power. Correlation of the IFN-RBV therapy outcomes has been reported with site polymorphisms in the core (36), E2 (87, 106), and NS5A (88, 106) proteins. Although these observations revealed numerous associations between the HCV genetic polymorphism and evolution toward IFN-RBV resistance, these associations were never explored in terms of their interrelationships and formulated into an integrative model capable of revealing accurate quantitative connections between HCV genetic changes and therapy outcomes.

The current report presents several probabilistic models connecting the UR/NR outcome to coordinated changes at polymorphic sites across the entire HCV polyprotein as well as from individual HCV proteins. Analysis of individual sites without consideration of their relationships seems inefficient in detecting a reliable connection to the outcomes. Only 3 among 25 sites having the highest value of mutual information with the outcome (Fig. 8) were found to be directly linked to the outcome in the polyprotein BN (Fig. 2). The same 3 sites, 2280, 2283, and 2633, are among the 14 most relevant sites extracted from the HCV polyprotein using correlation-based feature selection (Table 3) and among 18 sites that have the strongest connections to outcome in the undirected dependence graph (Fig. 1). All computational techniques used in this study ranked the contribution of various sites differently. For example, only 12 sites were shared by 18 sites shown in Fig. 1 and 25 sites shown in Fig. 8. Although sites 2280 and 2283 from NS5A and site 2633 from NS5B were frequently identified as most relevant to the IFN-RBV response, analysis of states at these sites is not sufficient for an accurate prediction of the therapy outcome (data not shown). Such a prediction requires the use of a combination of sites selected for their collective contribution to the outcome.

For that purpose, we conducted a series of experiments for selection of site sets most relevant to the therapy outcome from the entire HCV polyprotein and individual proteins (Table 3). Two proteins, E2 and NS5A, were explored in detail. As mentioned earlier, both proteins have many polymorphic sites and contributed many links to the polyprotein BN. These two proteins consistently made substantial contributions of the most relevant sites identified using different feature selection techniques (Fig. 5 and 6). Probabilistic mapping of UR and NR outcomes in 2D physicochemical space showed an equally representative distribution of the outcome probabilities for E2, NS5A, and the polyprotein (Fig. 9). All these findings strongly suggest that these two proteins have a strong connection to therapy response and can be used for the accurate prediction of therapeutic outcomes. However, as can be seen in Fig. 10, the 10-fold CV experiments showed that the NS5A BN outperforms the E2 BN constructed using complete sets of polymorphic sites (82.5% versus 90% accuracy) or feature-selected sites (85% versus 97.5% accuracy). These results, taken together with the observation that NS5A contains two of six sites directly connected to the therapy outcome in the polyprotein BN while E2 has no direct links to the outcome, suggest that NS5A has a very strong relevance to evolution toward the IFN-RBV resistance.

Two sites, at positions 2376 and 2414 in NS5A, have experimentally been associated with the development of resistance to RBV (97). It is important to note that these two sites were consistently selected as being relevant to the therapy outcome (Table 3 and Fig. 8), indicating that the NS5A BN as well as polyprotein BN constructed using all or feature-selected sites includes links that reflect contribution of RBV to therapy. Site 2414 located in domain 3 of NS5A is linked to site 161 in domain 2 of core in the polyprotein BN. As mentioned earlier, both domains are involved in protein-protein interactions between these two proteins, association with lipid droplets, and assembly and release of viral particles (81, 83). There seems to be a linkage between coevolution of the core and NS5A proteins and RBV resistance, and this resistance is associated with interaction between these two proteins. The final validation of the two predictive NS5A Viralhep-C models using the HALT-C data strongly confirms a robust connection between coordina-
tion among the NS5A sites and IFN-RBV resistance. Additionally, it shows that a small number of features from NS5A alone may be sufficient for the prediction of therapy outcomes (Fig. 11). This finding suggests that analysis of a very few sites from a small HCV genomic region, such as NS5A, may be used for monitoring sensitivity to the IFN-RBV therapy.

A general interconnectivity among HCV proteins was comparable for the 40 Virahep-C sequences and the 298 HCV genotype 1a full-genome sequences obtained from GenBank (Fig. 3 and 4), indicating that the modeled coordination among substitutions is essentially similar for all HCV variants from treated and treatment-naive patients. This observation additionally suggests that the development of resistance during immunomodulatory therapy is generally shaped by selection pressures similar to the HCV evolution in untreated patients. However, there are some important differences between the polyprotein BNs generated using sequences from treated and treatment-naive patients. The GenBank sequences from untreated patients contain more polymorphic sites (treatment-naive patients. The GenBank sequences from untreated patients contain more polymorphic sites (n = 1,296) than the Virahep-C sequences (n = 551). Despite this fact, the Virahep-C sequences contain 25 polymorphic sites that are conserved in the GenBank sequences. These sites are distributed within E1 (n = 3), E2 (n = 4), P7 (n = 1), NS2 (n = 2), NS3 (n = 6), NS4A (n = 1), NS4B (n = 3), NS5A (n = 3), and NS5B (n = 2). Among them, sites at positions 230 in E1, 768 in P7, and 1461 and 1592 in NS3 are the most relevant to the IFN-RBV response (Table 3). Furthermore, the two BNs had topological differences in the number of interprotein links, most notably the 1.7- and 2-fold proportional increase in the number of links between E1 and E2 and between E2 and NS5A in the Virahep-C BN compared to those in the GenBank BN (Fig. 3). These observations suggest that despite the similarity of these two networks, there are distinct differences in coordination among substitutions in HCV from treated and treatment-naive patients.

IFN is a major component of innate immunity (19, 100). Several HCV proteins are involved in modulation of the host IFN response (12, 13, 85, 126). RBV used as a component of combined therapy seems to facilitate early response to IFN (43) rather than playing a strong independent role. Resistance to IFN is not clearly linked to any specific mutation within the HCV genome. As shown in this study, HCV adaptation to IFN is a complex trait encoded in the interrelationships among many sites along the entire HCV polyprotein. The extensive coevolution among HCV amino acid sites leads to a significant integration among the HCV IFN-response-related phenotypic traits. Each HCV protein contributes to the IFN resistance, albeit to a different degree. With E2 and NS5A contributing many polymorphic sites to the network and generating a broad epistatic connectivity to sites in other HCV proteins, intrahost HCV evolution toward the IFN resistance is essentially defined and, therefore, can be accurately predicted using a carefully selected combination of sites from these two proteins.

Treatment with IFN does not exert an unusual selection pressure on HCV, unlike treatment using direct-acting antiviral compounds, but rather generates an unusually strong selection pressure of the innate immune system. Thus, HCV strains capable of resisting or evolving toward resistance to immunomodulatory therapy are most efficient in overcoming the host immune system. With the entire HCV genome being responsible for the response to IFN, there is no single IFN resistance mutation. Once established, the wide-ranging epistatic connectivity among sites involved in the IFN response may not be rapidly reverted even with reduction of the selection pressure in the absence of treatment, thus locking the HCV genome into the state of resistance to IFN. Without being eliminated by IFN-RBV therapy, these variants can continue to circulate among humans. In contrast, IFN-RBV-sensitive strains are being removed from circulation. This consideration implies that the current widespread adoption of IFN-based therapy, although extremely beneficial for individual patients with SVR, may affect the composition of the circulating HCV population and enlarge the reservoir of IFN-resistant HCV, a potentially alarming public health issue that warrants a further investigation.

ACKNOWLEDGMENTS

We are grateful to Chong-Gee Teo for critical review and discussion of findings in this paper as well as to two anonymous reviewers for important comments.

This work was supported by CDC intramural funding.

REFERENCES

3. Reference deleted.


