Coordinated evolution of the hepatitis C virus

D. S. Campo*†, Z. Dimitrova*, R. J. Mitchell‡, J. Lara*, and Y. Khudyakov*†

*Molecular Epidemiology and Bioinformatics Laboratory, Division of Viral Hepatitis, Centers for Disease Control and Prevention, 1600 Clifton Road, Atlanta, GA 30333; and ‡Department of Genetics and Human Variation, School of Molecular Sciences, La Trobe University, Bundoora, Victoria 3086, Australia

Edited by Francisco J. Ayala, University of California, Irvine, CA, and approved April 23, 2008 (received for review February 22, 2008)

Hepatitis C virus is a genetically heterogeneous RNA virus that is a major cause of liver disease worldwide. Here, we show that, despite its extensive heterogeneity, the evolution of hepatitis C virus is primarily shaped by negative selection and that numerous coordinated substitutions in the polyprotein can be organized into a scale-free network whose degree of connections between sites follows a power-law distribution. This network shares all major properties with many complex biological and technological networks. The topological structure and hierarchical organization of this network suggest that a small number of amino acid sites exert extensive impact on hepatitis C virus evolution. Nonstructural proteins are enriched for negatively selected sites of high centrality, whereas structural proteins are enriched for positively selected sites located in the periphery of the network. The complex network of coordinated substitutions is an emergent property of genetic systems with implications for evolution, vaccine research, and drug development. In addition to such properties as polymorphism or strength of selection, the epistatic connectivity mapped in the network is important for typing individual sites, proteins, or entire genetic systems. The network topology may help devise molecular intervention strategies for disrupting viral functions or impeding compensatory changes for vaccine escape or drug resistance mutations. Also, it may be used to find new therapeutic targets, as suggested in this study for the NS4A protein, which plays an important role in the network.

complex systems | scale-free network | covariation | natural selection | epistasis

Hepatitis C virus (HCV) is a major cause of liver disease worldwide. The global prevalence of HCV infection is estimated to be 2.2%, representing 130 million people (1). HCV causes chronic infection in 70–85% of infected adults (2). There is no vaccine against HCV and current antiviral therapy is relatively toxic, being effective in 50–60% of patients treated (3). HCV is a single-stranded RNA virus of ≈9.4 kb belonging to the *Flaviviridae* family (4). The positive-sense genome of HCV contains one large ORF that encodes a polyprotein that can undergo proteolytic cleavage into 10 mature proteins (C-E1-E2-P7-NS2-NS3-NS4A-NS4B-NS5A-NS5B). The structural proteins, the core (C) and envelope glycoproteins E1 and E2, are present in the N-terminal part of the polyprotein and presumably self-assemble to form the virion. The nonstructural (NS) proteins have various functions and form the replication complex (5).

The HCV genome continually mutates during virus replication. Although a high rate of mutation significantly contributes to the enormous adaptability of RNA viruses, it also limits the size of viral genomes by causing error catastrophe (6). The small size of viral genomes imposes strong evolutionary constraints on their organization, as a result of which each genomic region may encode multiple and often conflicting functions. Such genomic organization requires a tight coordination between the properties of individual viral components, which are subjected to frequent structural changes imposed by numerous mutations. A large number of mutations in tightly organized RNA viral genomes should lead to frequent epistatic interactions between sites. Several experiments have provided strong evidence of extensive epistasis in RNA viruses, confirming that during viral evolution certain substitutions at different sites may occur in a coordinated manner (7–12). Experimental studies into the adaptation of genetically modified HCV genomes propagated in cell culture also have shown antagonistic epistatic interactions between deleterious mutations exhibited in the form of compensatory mutations (13–15). These results suggest the importance of long-range interactions among HCV constitutive amino acids that place additional constraints on HCV heterogeneity. Understanding the constraints that shape HCV evolution may bring forth new strategies for public health control and clinical treatment of HCV infections. In the present study, we explore coordinated variation among genomic sites under different forces of natural selection on a set of 114 full genome sequences belonging to the 1b subgenotype.

Results and Discussion

Positive and Negative Selection. Positive selection is the process where a new mutant has a fitness value higher than the average preexisting types in the population; thus, its frequency increases in the following generations (16). Different amino acid sites have diverse biological functions and, therefore, are subject to different types and intensities of selective forces operating on them. An excess of nonsynonymous (amino acid-replacing) substitutions is attributed to positive selection, whereas an excess of synonymous (silent) substitutions is considered to result from negative selection. We calculated maximum likelihood estimates of synonymous and nonsynonymous changes (dS/dN) in each of the 3,010 codon sites of HCV 1b. It was found that HCV 1b evolution is dominated by the effects of negative selection. The dS/dN ratio is very high for all proteins, with a substitution being 4.66 times more likely to be silent than to lead to amino acid change. A total of 1,905 codon sites show evidence of negative selection. There are 833 sites that have a dS/dN ratio not significantly different from 1 (they are referred to as neutral sites for the remainder of this article). Only 60 sites are identified as positively selected (supporting information (SI) Table S1). These positively selected sites may play an important role in adaptive evolution of HCV 1b strains. The other 212 sites are conserved. Thus, despite significant observed heterogeneity, changes in HCV genome are severely restricted by negative selection. To understand the underlying mechanisms of these restrictions and further investigate constraints on evolution of HCV 1b strains, we proceeded to analyze the potential coordinated variation between different sites of the HCV 1b polyprotein.

Coordinated Substitutions. Each amino acid in the alignment was transformed by using five factors or patterns of highly associated physicochemical variables and the correlation coefficients between all sites were calculated. This analysis was performed by

Author contributions: D.S.C., R.J.M., and Y.K. designed research; D.S.C. and Z.D. performed research; D.S.C., Z.D., and J.L. analyzed data; and D.S.C. and Y.K. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

†To whom correspondence may be addressed. E-mail: fyv6@cdc.gov or yek0@cdc.gov.

This article contains supporting information online at www.pnas.org/cgi/content/full/0801774105/DCSupplemental.

EVOLUTION

using 448 polymorphic amino acid sites (14.9% of the total polyprotein, see *Methods*) located in all 10 proteins. There are 4,323 pairs with a significant correlation at one or more factors ($P = 0.0001$) that were found among 414 polymorphic sites (92.4% of sites analyzed). The high level of coordination of substitutions in these closely related HCV sequences indicates profound epistasis that affects most of the sites. From the 4,323 significant correlations (links), 17% are links between sites inside the same protein (intraprotein) and 83% are links between sites of different proteins (interprotein). NS5A and NS5B exhibit the highest number of links (20.3% and 19.1% of all links, respectively), whereas P7 and C show the lowest number of links (1.8% and 1.9%, respectively) (see Fig. S1*A*). NS4A also shows a low number of links (3.13%), but the highest average number of links per polymorphic site and the highest level of intraprotein links (see Fig. S1*B*).

The set of covariable sites contains 241 neutral sites, 105 negatively and 58 positively selected sites. There are numerous connections between sites under different forces of natural selection. Many links (88.6%) involve neutral sites and 42.2% are between neutral sites alone, suggesting that changes at neutral positions are coordinated as well as at other positions, and that "neutral" changes are not independent of positively and negatively selected sites. There are 220 significant correlations between positively and negatively selected sites, indicating a high level of coordination between stabilizing and adaptive evolution. Positively selected sites have only 28 links between each other (1.7% of all possible links between them), which is 3.7 times less than between neutral sites and 2.6 times less than between negatively selected sites. Such distribution of links suggests that positively selected sites are more independent from each other than any other class of sites.

The high number of covariable sites found in HCV 1b suggests that these sites form a network of coordinated substitutions, which we have constructed. This network is an undirected and unweighted graph where a vertex represents a polymorphic HCV amino acid site and an edge (link) is a significant physicochemical correlation between two sites. Five sets of connected sites (components) are found, and one of them is a giant component that includes 404 sites (97.6% of the 414 vertices) and 4,317 links (99.9% of all of the significant links). We analyzed this component with the tools of network theory to understand the local and global topology of the HCV network of coordinated substitutions (hereafter called the HCV network).

HCV Network Is Scale-Free. The simplest local characteristic of a vertex is the degree, k, the total number of its links. In most complex networks, the degree follows a power-law distribution, $P(k) \approx k^{-\gamma}$ and such networks are called scale-free (17). The degree in the HCV network also follows a power-law distribution with $\gamma = 2.27$ (Fig. 1*A*). The average degree is 21.37 and the maximum is 115 at site 887 located in NS2 (NS2_887). Most of the sites in the HCV network have a very low number of links with other sites and a small fraction of sites (hubs) have a high number of links. Recent studies used information theory to identify nonconserved coevolving sites in multiple sequence alignments from a variety of protein families and found that coevolving sites in these alignments fall into two general categories (23, 24). One set is composed of sites that coevolve with only one or two other sites, often displaying direct amino acid side-chain interactions with their coevolving partner. The other set comprises sites that coevolve with many others and are frequently located in regions critical for protein function, such as active sites and surfaces involved in molecular interactions and recognition. The hubs of the HCV network are sites where amino acid changes affect many other sites in the HCV polyprotein and, therefore, could play a very important role in determining viral functions and HCV evolution.

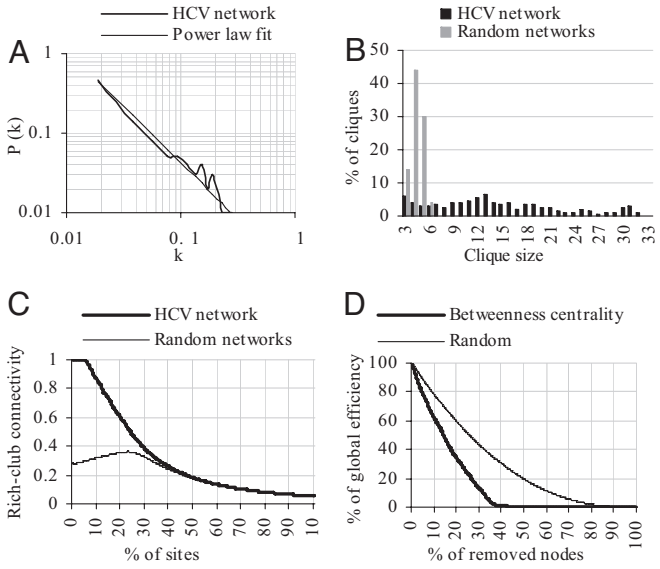

Fig. 1. Descriptive measures of the HCV network structure. (*A*) Degree distribution of the HCV network (solid line) and expected power law distribution (dashed line). (*B*) Clique distribution by size in the HCV network (black bars) and in 10,000 randomly rewired graphs (gray bars) that keep constant the number of vertices and their degree (19). (*C*) Rc of the HCV network (solid line) and the average of 10,000 randomly rewired graphs (dashed line) (20). (*D*) Structural robustness of the HCV network. The *x* axis is the percentage of removed nodes, and the *y* axis is the percentage of the original global efficiency of the network (21). The dashed line shows the random removal of nodes (average of 100 random sequences of node removal). The solid line shows the removal of nodes with the highest betweenness centrality (22). After each node removal, the betweenness centrality of each node was recalculated.

HCV Network Is a Small World. A ubiquitous characteristic of complex networks is the so-called small-world property. In simple terms, it means that there is a relatively short path between any two vertices despite the large size of the network (25). The HCV network has a low average shortest path between every pair of vertices (2.91) and a low diameter (maximum distance between two vertices being only 9). There is a high level of coordination between substitutions in the HCV 1b polyprotein, and very different regions of this polyprotein can be connected through a very small number of intermediary sites. The Clustering coefficient (26) (Cc) of each vertex was calculated. The Cc is the ratio of the number of links between the nearest neighbors of a vertex and the number of links they could have if they were a fully connected subgraph (clique). The HCV complex network has a high local structure, with an average Cc of 0.2894, more than five times higher than the average Cc of a random network with the same number of vertices and average degree (Cc = 0.0529). The number of cliques found in the HCV network was compared with the number of cliques found in randomly rewired graphs (Fig. 1*B*). There are many cliques of size 3 to 6 in the randomly rewired graphs, and bigger cliques are very rare. There is a significant enrichment of cliques in the HCV network, with 2,219 cliques of size 7 or more ($P = 0.00001$). The HCV network cliques can be viewed as tightly coordinated units of functionally or structurally connected sites.

HCV Network Has a Rich Club. Some complex networks show the so-called rich-club phenomenon where the vertices with a high degree (hubs) form a tightly interconnected community (20, 27). The rich-club coefficient (Rc) gives the ratio between the number of existing connections among vertices with high degree and the number of connections that they could have if they were

a clique (28). Fig. 1*C* illustrates that the HCV network shows a significant rich-club phenomenon. We can identify a set of 23 hubs that form a clique involved in 24.2% of all links in the network. This set of hubs has an Rc value that is 3.38 times higher than expected in random networks with the same number of vertices and degree distribution ($P = 0.0001$). The sites in the rich-club are pivotal to the HCV network and can be an important starting point for the functional mapping of inter- and intraprotein relationships.

Topological Robustness of the HCV Network. The computational analysis of complex networks indicates a strong correlation between robustness and network topology (25). In particular, scale-free networks display an unexpected degree of robustness (29). To evaluate topological robustness of the HCV network, we have analyzed changes in global connection efficiency by using random removal of nodes and removal of nodes with the highest centrality (Fig. 1*D*). Similar to many other scale-free networks, the HCV network was found to be very robust to random node failure but vulnerable to attacks at hubs with high centrality. The removal of only 14.4% of high-centrality nodes decreases the efficiency of the network to half its original value. The importance of some sites in maintaining network connectivity offers an opportunity for the development of new methods to reduce HCV viability.

Hierarchical Structure of the HCV Network. The HCV network is assortative. The correlation between the degrees of vertices connected by a link is $r = 0.4072$ ($P = 0.0001$). This means that extensively connected sites in the HCV network attach preferentially to other extensively connected sites whereas scarcely connected sites attach preferentially to other scarcely connected sites. To further analyze the HCV network organization, we used the *k*-shell decomposition method to disentangle the hierarchical structure of the network (30). The process is started by removing all vertices with only one connection ($k = 1$), repeating until no such vertices remain, and assigning all the removed vertices to the 1-shell. The process continues by increasing k until all vertices in the graph have been assigned to one of the shells. *k*-shell decomposition can uncover several topological and hierarchical properties of large-scale networks in a two-dimensional layout (Fig. 2). We can divide the HCV network in two subgraphs: a nucleus and crust. At the highest shell (no. 36) we find the nucleus of the network. It comprises 45 almost fully interconnected sites (Cc = 0.9592) that include all 23 members of the rich club (see Fig. S2). The nucleus contains highly connected sites whose links are globally distributed. These sites are involved in 40% of all links of the network, of which 54.7% are links to other sites inside of the nucleus and 45.3% to sites in the crust. The crust is defined as the union of all shells with value 1 to 35, which involve 359 sites (89% of all sites in the network). Interestingly, even though the nucleus is responsible for the global connectivity, if we remove it from the network, the crust still forms a very large connected component of 97.8% of the remaining vertices.

The HCV sites are not uniformly distributed among *k*-shells of the network, with almost 54.4% of sites being in shells 1–7 and 11.1% in the nucleus (see Fig. S3 *A* and *B*). The first shell contains 14.4% of all sites. Shells 8–35 are poorly populated. Each one of these shells contains on average only 1.2% of all sites. The sites from outer shells should have a limited global effect on the network given their low number of links. Sites in the nucleus, however, should exert a profound effect on the network because of their large number of links and high centrality. The structure of the network suggests that the nucleus plays an important global role in the coordination of substitutions in the HCV polyprotein and, consequently, in HCV functional organization and evolution.

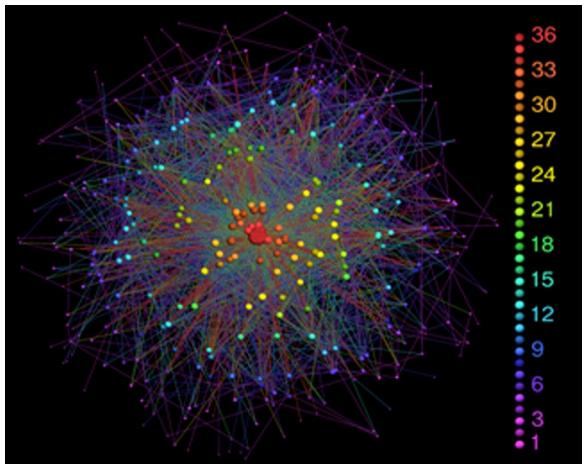

Fig. 2. The HCV network. A genomic network of HCV coordinated substitutions is shown, where a vertex is an amino acid site and a link between two vertices is a significant physicochemical correlation between two sites. The position of each vertex depends on its *k*-shell value and on the *k*-shell value of its neighbors. A color code allows for the identification of *k*-shell values, and the vertex's size is proportional to its degree. The *k*-shell decomposition and visualization was performed with LaNet-vi (31).

The hierarchical structure of the HCV network suggests disparity in the global effects of different sites on HCV evolution. Thus, negatively or positively selected sites found in the nucleus may have a more global effect on the HCV evolution than peripherally located sites under the same forces of natural selection. This distribution of sites in the network indicates that there are different types of positive and negative selection depending on the epistatic connectivity. The presence of neutral sites among the large hubs at the nucleus of the network suggests that the concept of neutrality cannot be applied with certainty in viral evolution. These sites seem to change randomly but their variation has a strong global effect on HCV evolution. All neutral sites in the nucleus are constrained by influences from sites under strong forces of positive and negative selection. We may speculate that neutral sites "buffer" these forces to sustain serviceable viral functions. The fact that ≈90% of neutral sites in the network have a straight link to positive and/or negative sites suggests that neutral sites may still be selected to mediate effects caused by substitutions at directly selected sites. It should be noted that there is a small fraction of neutral covariable sites that have no straight links to positive or negative sites, which implies that some neutral sites vary stochastically as a group. However, taking into consideration the small-world property, it seems that all neutral sites in the network are under some degree of influence from the directly selected sites.

The finding of positively selected sites in the nucleus is intriguing. There is a potential functional disparity between positively selected sites because a majority of these sites is located in the outer shells of the network and most probably have a more peripheral role in HCV adaptive evolution. The entire network contains 14.4% of positively selected sites. The nucleus, however, contains 3.2 times less positively selected sites than the outer shell, which seems to suggest a limited role of the nucleus in adaptive evolution. Only two positively selected sites (E1_210 and NS3_1384) are found in the nucleus of the HCV network. Both sites are large hubs that have links with 105 densely connected sites located in all 10 proteins. Assuming that the substitution paths of positively selected sites are driven by host-selective pressures, these two positively selected sites with a high centrality in the HCV network should play a prominent global role in HCV adaptive evolution.

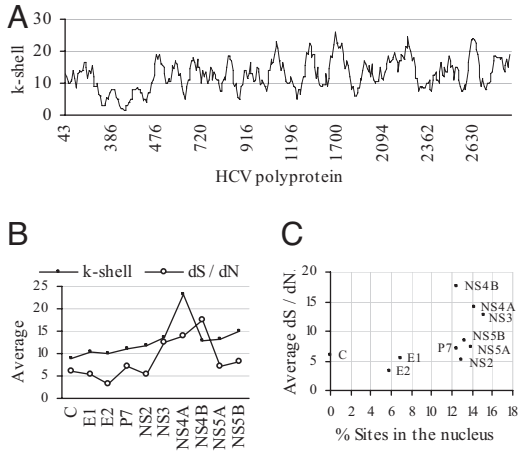

Fig. 3. HCV network and natural selection. (A) Sites that compose the HCV network and their *k*-shell value, which is shown as a moving average (window size = 9; step = 1). (*B*) Average values of *k*-shell for each protein (squares), calculated by using 404 sites in the network, and the average of the dS/dN ratio (circles) calculated over all of the sites of the protein. The correlation between the average *k*-shell and dS/dN values in HCV proteins is 0.5930 ($P = 0.00001$). (*C*) Scatter plot of HCV proteins. The *x* axis is the percentage of the sites of each protein that are located in the nucleus (*k*-shell, 36) and *y* axis the average dS/dN.

Network Inferences on the Global Role of HCV Proteins. Fig. 3*A* shows a moving average of the *k*-shell values of the 404 sites that make up the network. All proteins contain regions of low- and high-average *k*-shell value. Fig. 3*B* shows the average *k*-shell value for sites in each protein. The lowest *k*-shell values are found in the structural proteins (C, E1, and E2), indicating that, in general, their sites occupy peripheral positions in the network. The structural proteins also show the lowest percentages of sites located in the nucleus and the highest levels of positive selection (Fig. 3*B*). The highest *k*-shell value is found in NS4A; NS4A, NS3, and NS4B are the proteins with the highest levels of negative selection. Few sites in these two proteins are free to vary without deleterious consequences, thus decreasing the range of tolerated substitutions. Fig. 3*C* shows a scatter plot of HCV proteins, comparing the percentage of sites in the nucleus with the average dS/dN. The HCV structural and nonstructural proteins are clearly separated from each other indicating differences in natural selection forces and global network influences between these two groups of proteins.

Constraints on the Hypervariable Region 1 (HVR1). The region with the lowest average *k*-shell value is located in the C terminus of E1 and the N terminus of E2 including HVR1 (Fig. 3*A*). HVR1 contains the highest number of positively selected sites (11 sites), confirming its important role in HCV adaptive evolution and survival. However, the presence of five negatively selected sites is consistent with strong selective constraints previously found on HVR1 heterogeneity (32–34). There are 135 links involving 22 sites in the HVR1 and 86 sites distributed in all HCV proteins. The sites in the HVR1 are mostly in the periphery of the network (average *k*-shell value, 5.1818; standard deviation, 5.7622) suggesting a very limited global influence of these mutations on HCV evolution. However, there are 54 links between HVR1 sites and sites in the nucleus of the network. The connections with sites of high centrality are indicative of additional strong constraints on HVR1 evolution imposed by many highly connected sites located in different regions of the HCV polyprotein. There are five links involving eight different sites inside the HVR1 (see Fig. S4). Changes in the HVR1 are correlated in a way that contributes to the invariance of the integral physicochemical

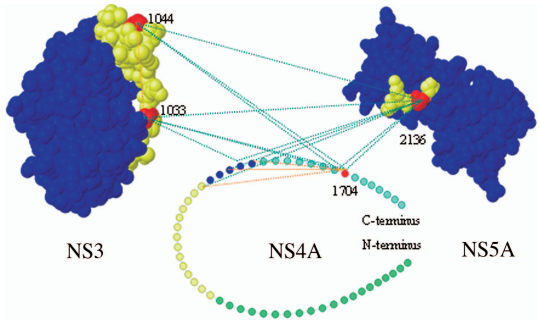

Fig. 4. Hypothetical molecular interactions between NS3, NS4A, and NS5A. The links between sites are shown as green lines (interprotein) or orange lines (intraprotein). Sites located in the nucleus of the HCV network are shown in red. NS3, 3D structure of the NS3 protease (sites 1029 to 1205). The first 28 N-terminal aa involved in the molecular interaction with its NS4A cofactor are shown in yellow and red. NS5A, 3D structure of NS5A (sites 2008 to 2170). The sites involved in the molecular interaction with NS4A (2135 to2139) are shown in yellow and red. NS4A, All 54 sites of NS4A are shown as a contiguous chain divided in four domains: the membrane anchor (green), the NS3 cofactor (yellow), the kink (blue), and the acidic domain (cyan).

characteristics of HVR1, keeping constant the polarity ($P = 0.0001$) and size of the segment ($P = 0.0001$). The high number of coordinated substitutions and their contribution to the integrity of these physicochemical properties provide an additional proof to conservation of conformational motifs in the HVR1, for which there is some experimental evidence (35).

High Centrality of NS4A. NS4A is a key regulator of the essential serine protease and RNA helicase activities of the NS3–4A complex (36) and a determinant of NS5A phosphorylation and HCV replicase activity (37). Our study identifies NS4A as the protein with the highest average number of links per polymorphic site, intraprotein links, and average centrality in the network, and the second-highest level of negative selection. These results suggest an important role for this small protein in the HCV life cycle, revealing that it is an ideal target molecule for HCV therapy. The central position of NS4A in the network encourages a fine mapping of some of its links with other nonstructural proteins. Fig. 4 shows the first 28 N-terminal aa of NS3 involved in the molecular interaction with its NS4A cofactor, positions 2135–2139 of NS5A, which are important for NS4A-dependent phosphorylation (38). Some of the sites of NS3 and NS5A involved in these molecular interactions also have links with some of the sites of NS4A, forming a subnetwork of coordinated substitutions. This subnetwork is in agreement with the results of a recent experimental study (15) that found that substitutions in NS4A lead to decreased replication and NS5A hyperphosphorylation and that compensatory mutations in NS3 and NS5A suppress these defects.

Conclusions. One of the main reasons for the popularity of complex networks is their flexibility and generality in representing virtually any natural structure. Each complex network presents specific topological features that characterize its connectivity and highly influence the dynamics of processes executed on that network. The HCV network of coordinated substitutions has a topology with some key characteristics also identified in biological, technological, and social complex networks, enhancing the sense of unity underlying the structure and dynamics of optimized systems.

In addition to genetic heterogeneity and natural selection, epistatic connectivity mapped in the scale-free network is an important property for the characterization of sites or proteins and may be used for typing genetic systems. We have found that

each HCV protein has regions of low and high centrality in the network. However, a general trend is that the sites of nonstructural proteins are located in more central regions of the HCV network than the sites of structural proteins. These findings suggest that, as a result of intragenomic coordination of substitutions over all proteins through their association to the nucleus, the coordinated changes in some critical sites of the viral replication complex may be responsible for directing long-term HCV evolution. Considering that the sites of structural proteins have a high level of positive selection and are preferentially located toward the periphery of the HCV network, substitutions in structural proteins probably have a low probability of directing viral evolution in the long term but may be responsible for fast adaptation to the host.

The complex coordination of substitutions shown in this viral network is an emergent property of genetic systems with many implications for the understanding of their robustness and evolvability. Our analysis suggests that the network is very tolerant to random point mutations. A very small number of high-degree positions exert a strong global impact on the state of the entire genetic system and mutations at such positions may direct viral evolution to a different path, for example, under immune selection pressure. Therefore, these important network positions may be used as targets for vaccines or drugs. Mapping epistatic connectivity between sites offers an insight into the mechanics of compensatory changes and may help in devising molecular strategies to effectively disrupt a large number of viral functions and/or hamper occurrence of viral compensatory mutations that may moderate fitness cost for vaccine escape or drug resistance mutations. Additionally, such a network analysis may help discover new targets for therapeutic interventions, similar to the HCV NS4A protein, the network importance of which was shown in this study.

Methods

Sequences, Alignment, and Selection. Please see *SI Text* for the GenBank accession numbers of the 114 HCV 1b complete genome sequences used in this research. The HCV sequences were aligned by using ClustalW (39). The detection of selection was performed with the Single-Likelihood Ancestor Counting (SLAC) algorithm implemented in the program HyPhy (40).

Coordinated Substitutions. Each amino acid in the 114 HCV sequences was transformed into a string of values at five physicochemical factors created by factor analysis of 494 amino acid properties (41, 42). A modified version of a recent algorithm (43) was used to calculate the correlation between the physicochemical values of HCV amino acid sites. To test whether a correlation value reflects a significant association (possibly because of structural and functional constraints), or results from evolutionary history and stochastic events (background covariation) we followed several guidelines. In brief, we conducted a permutation test by using 10,000 randomizations, the phylogenetic sequence weighting (44, 45), a background correlation threshold (24), and False Discovery Rate calculations (46). The analysis was performed by using only 448 polymorphic amino acid sites with an entropy >0.2370 (10% of the highest amino acid entropy found in HCV). The 448 polymorphic sites include 60 (13.39%) positively selected, 105 (23.44%) negatively selected, and 283 (63.17%) neutral sites. Although the use of highly polymorphic sites potentially increases the probability of finding truly covariable sites (23, 24), it also disproportionately reduces the number of negatively selected sites included in the analysis because of their limited heterogeneity. Nevertheless, because our goal was to identify the phenomenon of substitution coordination, such a conservative approach to site selection seems justified.

Network. All network analyses of this study were performed by using MATLAB (47) unless stated otherwise. We calculated the degree distribution of the network (18), Cc of each vertex (26), the significance of each clique (19), Rc (28) and its significance (20), the local and global efficiency of the network (21) and the betweenness centrality of each node (22). The *k*-shell decomposition and visualization was performed by using LaNet-vi (31).

Detailed description of the methods can be found in *SI Text*. The following files are available on request: the multiple sequence alignment, the SLAC results for 3,010 codons, and the complete list of links, including their correlation values for each physicochemical factor.

ACKNOWLEDGMENTS. We thank Dr. Chong-Gee Teo (Division of Viral Hepatitis, Centers for Disease Control and Prevention, Atlanta), Elizabeth Neuhaus (Division of Scientific Resources, Centers for Disease Control and Prevention, Atlanta), Dr. David Anderson (Macfarlane Burnet Institute for Medical Research, Melbourne), and two anonymous reviewers for encouraging and insightful comments.

1. Alter M (2007) Epidemiology of hepatitis C virus infection. *World J Gastroenterol* 13:2436–2441.
2. Alberti A, Chemello L, Benvegnu L (1999) Natural history of hepatitis C. *J Hepatol* 31:17–24.
3. Bowen D, Walker C (2005) Adaptive immune responses in acute and chronic hepatitis C virus infection. *Nature* 436:946–952.
4. Choo Q, *et al.* (1989) Isolation Of A Cdna clone derived from a bloodborne non-A, non-B viral hepatitis genome. *Science* 244:359–362.
5. Roingeard P, Hourioux C, Blanchard E, Brand D, Ait-Goughoulte M (2004) Hepatitis C virus ultrastructure and morphogenesis. *Biol Cell* 96:103–108.
6. Holmes E (2003) Error thresholds and the constraints to RNA virus evolution. *Trends Microbiol* 11:543–546.
7. Burch C, Chao L (2004) Epistasis and its relationship to canalization in the RNA virus phi 6. *Genetics* 167:559–567.
8. Bonhoeffer S, Chappey C, Parkin N, Whitcomb J, Petropoulos C (2004) Evidence for positive epistasis in HIV-1. *Science* 306:1547–1550.
9. Sanjuan R, Moya A, Elena S (2004) The contribution of epistasis to the architecture of fitness in an RNA virus. *Proc Natl Acad Sci USA* 101:15376–15379.
10. Poon A, Chao L (2005) The rate of compensatory mutation in the DNA bacteriophage phiX174. *Genetics* 170:989–999.
11. Mateo R, Mateu M (2007) Deterministic, compensatory mutational events in the capsid of foot-and-mouth disease virus in response to the introduction of mutations found in viruses from persistent infections. *J Virol* 81:1879–1887.
12. Garriga CP-E, *et al.* (2007) Mutational patterns and correlated amino acid substitutions in the HIV-1 protease after virological failure to nelfinavir- and lopinavir/ritonavir-based treatments. *J Med Virol* 79:1617–1628.
13. Murray C, Jones C, Tassello J, Rice C (2007) Alanine scanning of the hepatitis C virus core protein reveals numerous residues essential for production of infectious virus. *J Virol* 81:10220–10231.
14. Yi M, Ma Y, Yates J, Lemon S (2007) Compensatory mutations in E1, p7, NS2, and NS3 enhance yields of cell culture-infectious intergenotypic chimeric hepatitis C virus. *J Virol* 81:629–638.
15. Lindenbach B, *et al.* (2007) The C terminus of hepatitis C virus NS4A encodes an electrostatic switch that regulates NS5A hyperphosphorylation and viral replication. *J Virol* 81:8905–8918.
16. Suzuki Y, Gojobori T (1999) A method for detecting positive selection at single amino acid sites. *Mol Biol Evol* 16:1315–1328.
17. Barabasi A, Albert R (1997) Emergence of scaling in random networks. *Science* 286:509–512.
18. Clauset A, Shalizi CR, Newman MEJ (2007) Power-law distributions in empirical data, arXiv:0706.1062v1 [physics.data-an].
19. Maslov S, Sneppen K (2002) Specificity and stability in topology of protein networks. *Science* 296:910–913.
20. McAuley J, Costa L, Caetano T (2007) The rich-club phenomenon across complex network hierarchies. *Appl Phys Lett* 91:084103.
21. Latora V, Marchiori M (2001) Efficient behavior of small-world networks. *Phys Rev Lett* 87:1–14.
22. Freeman L (1977) A set of measures of centrality based on betweenness. *Sociometry* 40:35–41.
23. Gloor G, Martin L, Wahl L, Dunn S (2005) Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. *Biochemistry* 44:156–165.
24. Martin L, Gloor G, Dunn S, Wahl L (2005) Using information theory to search for coevolving residues in proteins. *Bioinformatics* 21:4116–4124.
25. Albert R, Barabasi A (2002) Statistical mechanics of complex networks. *Rev Modern Phys* 74, arXiv:cond-mat/0106096v1 [cond-mat.stat-mech].
26. Watts D, Strogatz S (1998) Collective dynamics of small-world networks. *Nature* 393:440–442.
27. Colizza V, Flammini V, Serrano M, Vespignani A (2006) Detecting rich-club ordering in complex networks. *Nat Phys* 2:110–115.
28. Zhou S, Mondragon R (2004) The rich-club phenomenon in the internet topology. *IEEE Commun Lett* 8:180–182.
29. Albert R, Jeong H, Barabasi A (2000) Error and attack tolerance of complex networks. *Nature* 406:378–382.
30. Carmi S, Havlin S, Kirkpatrick S, Shavitt Y, Shir E (2007) A model of Internet topology using k-shell decomposition. *Proc Natl Acad Sci USA* 104:11150–11154.
31. Alvarez-Hamelin I, Dall'Asta L, Vespignani A (2006) k-core decomposition: A tool for the visualization of large scale networks. *Adv Neural Inform Processing Syst* 18, arXiv:cs/0504107v2 [cs.NI].
32. McAllister J, *et al.* (1998) Long-term evolution of the hypervariable region of hepatitis C virus in a common-source-infected cohort. *J Virol* 72:4893–4905.

33. Smith D (1999) Evolution of the hypervariable region of hepatitis C virus. *J Viral Hepat* 6:41–46.
34. Sheridan I, Pybus O, Holmes E, Klenerman P (2004) High-resolution phylogenetic analysis of hepatitis C virus adaptation and its relationship to disease progression. *J Virol* 78:3447–3454.
35. Mondelli M, *et al.* (2001) Hypervariable region 1 of hepatitis C virus: Immunological decoy or biologically relevant domain? *Antiviral Res* 52:153–159.
36. Failla C, Tomei L, De Francesco R (1994) Both NS3 and NS4A are required for proteolytic processing of hepatitis C virus nonstructural proteins. *J Virol* 68:3753–3760.
37. Tanji Y, Kaneko T, Satoh S, Shimotohno K (1995) Phosphorylation of hepatitis C virus-encoded nonstructural protein NS5A. *J Virol* 69:3980–3986.
38. Asabe S, *et al.* (1997) The N-terminal region of hepatitis C virus-encoded NS5A is important for NS4A-dependent phosphorylation. *J Virol* 71:790–796.
39. Thompson J, Higgins D, Gibson T (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22:4673–4680.
40. Kosakovsky S, Frost S, Muse S (2005) HyPhy: Hypothesis testing using phylogenies. *Bioinformatics* 21:676–679.
41. Atchley W, Zhao J, Fernandes A, Druke T (2005) Solving the protein sequence metric problem. *Proc Natl Acad Sci USA* 102:6395–6400.
42. Kawashima S, Kanehisa M (2000) AAindex: Amino acid index database. *Nucleic Acids Res* 28:374.
43. Afonnikov D, Kolchanov N (2004) CRASP: A program for analysis of coordinated substitutions in multiple alignments of protein sequences. *Nucleic Acids Res* 32:W64–W68.
44. Felsenstein J (1985) Phylogenies and the comparative method. *Am Nat* 125:1–15.
45. Vingron M, Sibbald P (1993) Weighting in sequence space: A comparison of methods in terms of generalized sequences. *Proc Natl Acad Sci USA* 90:8777–8781.
46. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc Ser B* 57:289–300.
47. MathWorks (2007) MATLAB (MathWorks, Natick, MA), Version R2007a.