MAJOR ARTICLE

Molecular Epidemiology and Forensic Genetics: Application to a Hepatitis C Virus Transmission Event at a Hemodialysis Unit

Fernando González-Candelas, María Alma Bracho, and Andrés Moya

Instituto Cavanilles de Biodiversidad y Biología Evolutiva, Universidad de Valencia, Valencia, Spain

Molecular phylogenetic analyses are frequently used in epidemiologic testing, although only occasionally in forensics. Their acceptability is hampered by a lack of statistical confidence in the conclusions. However, maximum likelihood testing provides a sound statistical framework for the testing of phylogenetic hypotheses relevant for forensic analysis. We present the results of applying this method to a small hepatitis C outbreak produced in a hospital hemodialysis unit that involved 6 patients. Polymerase chain reaction products from a 472-nt fragment of the E1–E2 region, including the hypervariable region, HVR-1, of the hepatitis C virus genome were cloned, and an average of 10 clones/patient and from 11 additional control patients were sequenced. The method allows a statistical evaluation that the likelihood of each sample belonging or not to a given group, a question of relevance in many forensic and epidemiological analyses of molecular sequences.

Molecular phylogenetics have been frequently applied to epidemiological studies, although only occasionally to forensic analyses [1–4]. In a pioneer study [1], a molecular phylogenetic analysis of the *env* region of human immunodeficiency virus (HIV) was used to establish support for transmission of the virus by a practicing dentist to some of his patients. However, this conclusion was questioned on the grounds of the inadequacy of the evolutionary model applied in the analysis of the data [5], although subsequent analyses with different evolutionary models and phylogenetic reconstruction methods provided further support for the original conclusion [6, 7].

There are several statistical methods available for testing a specific evolutionary hypothesis. Although resampling methods such as bootstrap testing [8, 9] have become the most popular, more rigorous and readily interpretable methods under a statistical framework are available [10]. The need for an individual evaluation of relatedness or pertinence to a specified group appears traditionally in a forensic context but is also frequent in epidemiological studies—for instance, when an outbreak of a relatively prevalent disease affects several individuals who share several risks and there is a need to identify those who were infected from a common source. This is most necessary for diseases with a prolonged asymptomatic period and relatively frequent nosocomial transmission, such as hepatitis B and C virus infection.

Within the realm of forensic analysis, as depicted by Evett and Weir [11], it seems evident that maximum likelihood testing should be the method of choice for evaluating competing phylogenetic hypotheses that are linked to other nonmolecular or genetic evidence and for providing quantitative criteria for deciding between alternative possibilities to the jury or those in charge of making a judicial decision. There are 3 main reasons for this assertion.

Received 8 July 2002; revised 15 October 2002; electronically published 8 January 2003.

Financial support: Andorran Government (Ministeri de Salut i Benestar), Spanish Ministerio de Educación y Ciencia, Plan Nacional I+D (project 1FD97-2328), and Ministerio de Ciencia y Tecnología (grant BMC2001-3096).

Reprints or correspondence: Dr. Fernando González-Candelas, Instituto Cavanilles de Biodiversidad y Biología Evolutiva, Universidad de Valencia, Apartado Oficial 22085, 46071 Valencia, Spain (fernando.gonzalez@uv.es).

The Journal of Infectious Diseases 2003;187:352–8

0022-1899/2003/18703-0002$15.00

First, maximum liklihood testing provides a direct way of computing the odds ratio (OR) between the prosecution and the defense propositions (H_p and H_d, respectively), given the evidence *(E)* and other independent sources of evidence *(I),*

$$\frac{\Pr(H_p \mid E,I)}{\Pr(H_d \mid E,I)} = \frac{\Pr(E \mid H_p,I)}{\Pr(E \mid H_d,I)} \times \frac{\Pr(H_p \mid I)}{\Pr(H_d \mid I)},$$

by multiplying the prior odds,

$$\frac{\Pr(E \mid H_p,I)}{\Pr(E \mid H_d,I)},$$

by the ratio of 2 probabilities, the likelihood ratio (LR),

$$\mathrm{LR} = \frac{\Pr(E \mid H_p,I)}{\Pr(E \mid H_d,I)}.$$

Furthermore, this relationship underscores the scientist's role in forensic analysis as providing a quantitative answer to the question, "What is the probability of the evidence, given the proposition (of the prosecution or the defense)?" rather than, "What is the probability of the proposition, given the evidence?" [11]. Second, this method allows an evaluation of the evidence (in this context, the molecular data, and of the corresponding phylogenetic reconstruction) on an individual basis, because the phylogenetic positioning of all the sequences derived from a single source can be tested, rather than providing a joint evaluation of a set of several individuals, as in bootstrap-based statistical tests. Third, it provides a straightforward way of separating the contribution of the evidence evaluated by the scientist from those of other evidences incorporated into the judicial cause. All are necessary for deciding the verdict, but the scientist is in charge only of his or her part, and, within this framework, it is possible to provide these data in a quantitative form.

Hepatitis C virus (HCV), the primary etiologic agent of parenterally transmitted non-A, non-B hepatitis, is a major cause of acute and chronic hepatitis and cirrhosis worldwide. The most efficient transmission of HCV is associated with percutaneous exposures to blood [12]. Hence, despite the testing of blood donors for anti-HCV antibodies and the adoption of rigorous preventive measures [13, 14], hemodialysis units still represent one of the main sources of nosocomial infection with HCV [4, 15, 16].

However, apart from the multiple personal and clinical consequences of these infections, there is an epidemiological, and often also a forensic, component that deserves very close scrutiny. This is the ascertainment of the most likely source of infection when there is a common link among several patients, such as attendance at a particular hemodialysis unit, but there are also other common risks of infection for some or even all the affected patients—for instance, being intravenous drug users or transfused patients. Furthermore, hepatitis C is a rather prevalent disease [17, 18], with most countries reporting prevalences of 1%–4%, and a fraction similar to that in the general population is expected to share the risk associated with any other source considered.

A relevant feature of HCV, common to all RNA viruses, is its extremely high sequence variability, such that isolates from a single patient are not identical but differ to a certain extent depending on the genome region being compared [3, 19]. Hence, there is not a single genome sequence characterizing the population but a swarm of more or less related sequences, sometimes referred to as a "viral quasispecies" [20]. Therefore, the relationship among viral isolates, from the same or different patients cannot be ascertained on the basis of a single sequence. Rather, a number of independent sequences from each patient must be analyzed and their relationship established by use of appropriate methods, such as those provided by molecular phylogenetics and population genetics.

In 1999, several individuals with positive results of testing for anti-HCV antibodies were detected among attendees of the hemodialysis unit at the Hospital de Nostra Senyora de Meritxell (Andorra), some of whom also tested positive for presence of HCV RNA in their blood by polymerase chain reaction (PCR) assays. As were many other users of this unit, these were transient unit attendees, with permanent residence in different Spanish localities. After the clinical and epidemiological analysis of those patients, it became evident that the most likely source of infection was found at the aforementioned unit. Hence, it became necessary to confirm the epidemiological analysis and to establish which patients were actually related to this source.

Our goal in the present article is to apply molecular phylogenetic tools to establish the relatedness among virus variants and to test the inferred relationships by a rigorous statistical procedure, maximum likelihood, thus assigning a quantitative estimate on the reliability of those relationships. This analysis, complemented with an epidemiological study on common links and risks of the infected patients, will establish, on an individual basis, the probability of association to a certain source of infection and, in consequence, can be integrated into a forensic report if needed.

PATIENTS, MATERIALS, AND METHODS

Patients. Six patients attending the Hemodialysis Unit of Hospital Nostra Senyora de Meritxell were apparently infected with HCV genotype 1b. As population controls, the study included 2 patients undergoing hemodialysis from the same hospital not connected to the outbreak and 9 further patients from the nearest reference hospital in Spain (Hospital General Universitari Vall d'Hebron, Barcelona). All 11 control patients were

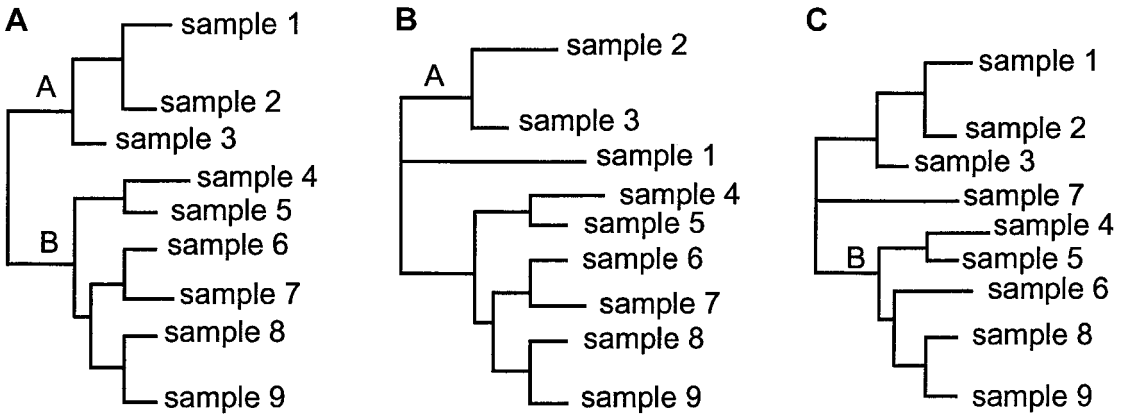

Figure 1. Schematic representation for the maximum-likelihood testing of alternative phylogenetic hypotheses. *A,* Initial topology that was obtained without any a priori hypothesis. Clades derived from A and B represent outbreak and control samples, respectively. *B,* Alternative topology to test whether sample 1 is actually more closely related to control samples than to outbreak samples. *C,* Alternative topology to test whether sample 7 is actually more closely related to outbreak samples than to control samples.

previously tested to be HCV 1b positive. The handling of samples followed the standards of the European Federation for Immunogenetics for nucleic acid analysis and PCR amplifications, with special care taken in avoiding PCR contamination and false positive results [21].

RNA extraction, reverse transcription (RT), and cloning. Viral RNA was extracted from 140 μL of serum using the QIAamp Viral RNA Kit (Qiagen). RT was performed on a 20-μL volume containing 10 μL of eluted RNA, 4 μL of 5× RT buffer, 500 μ*M* of each deoxynucleotide, 1 μ*M* antisense primer (5′-GGYGSGTARTGCCARCARTA-3′), 100 U of Moloney murine leukemia virus RT (USB), and 20 U of RNaseOUT (Gibco BRL). The reaction was incubated at 42°C for 45 min, followed by 3 min at 95°C.

The first amplification was performed in a 100-μL volume containing 10 μL of the rRT product, 10 μL of 10× PCR buffer, 200 μ*M* each dNTP, 400 n*M* each primer (sense, 5′-CGCATGGCYTGGGAYATGAT-3′; antisense, 5′-GGYGSGTARTGCCARCARTA-3′), and 2.5 U of *Pfu* DNA polymerase (Stratagene). In case patients for whom no amplicon was detected, it was necessary to perform a second PCR with a nested sense primer (5′-GGGATATGATRATGAAYTGGTC-3′). In all case patients, PCR was performed in a Perkin Elmer 2400 thermal cycler with the following thermal profile: 94°C for 3 min; 5 cycles at 94°C for 30 s, 55°C for 30 s, and 72°C for 3 min; 35 cycles at 94°C for 30 s, 52°C for 30 s, and 72°C for 3 min; and a final extension at 72°C for 10 s. A single 472-nt amplified product was observed after electrophoresis on a 1.4% agarose gel stained with ethidium bromide.

Amplification products were directly cloned in *Eco* RV-digested pBluescript II SK$^+$ phagemid (Stratagene). Plasmid DNA was purified with CONCERT Rapid Plasmid Purification Systems (GibcoBRL). Recombinant clones were sequenced by use of KS and SK primers (Stratagene) and the ABI PRISM d-Rhodamine Terminator Cycle Sequencing Ready Reaction Kit in an ABI 377 XL automated sequencer (Applied Biosystems). Sequences were verified, and both strands were assembled using the Staden package [22].

Phylogenetic analysis. Sequence alignments were obtained using CLUSTALW [23]. The neighbor-joining algorithm [24] applied on the pairwise nucleotide divergence matrix using the Kimura 2-parameter (K2P) model [25] was used to obtain phylogenetic trees. Further refinement of the evolutionary model was obtained by evaluating the likelihood of increasingly complex models (Jukes-Cantor [JC] [26], K2P [25], Felsenstein 1981 [27], and Hasegawa-Kishino-Yano 1985 [HKY85] [28], with constant and variable rates among sites using DNArates (available at http://geta.life.uiuc.edu/~gary/programs/DNArates.html). Likelihoods of the different phylogenetic reconstructions were computed using the program FASTDNAML [29].

Statistical analysis of competing hypothesis. An alternative hypothesis for relationships among viral sequences is represented by alternative phylogenetic reconstructions in which sequences derived from a given patient are more closely related to a different group from that originally established. These can be considered as the null and alternative hypotheses in the forensic framework depicted previously. Figure 1 presents a schematic representation of the alternative hypotheses considered when testing whether a sequence, or group of sequences, initially included in the outbreak (sample 1 in figure 1*A*) is more closely related to control samples (figure 1*B*) and whether a sequence initially assigned to the general population (sample 7 in figure 1*A*) is actually more closely related to those sharing a common source (figure 1*C*). In both cases, the topologies for the alternative hypotheses are specified without branch lengths, thus allowing the phylogenetic reconstruction program FASTDNAML to optimize them. A nice consequence of this procedure is that the topologies cor-

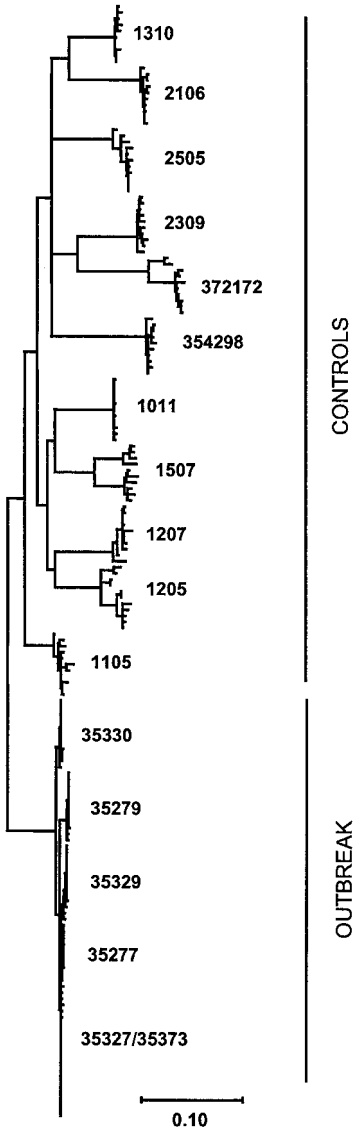

Figure 2. Phylogenetic tree for sequences of the E1–E2 region of hepatitis C virus. All control patients had clearly separate clades, whereas outbreak patients 35327 and 35373 shared several sequence clones, thus appearing in a unique clade.

responding to both alternative hypotheses become identical when branch lengths are not specified.

RESULTS

The HCV genome region analyzed encompasses the E1 and E2 genes, including the hypervariable region (HVR-1) in E2. This region shows the highest rate of evolution in the HCV genome (0.7–15.7 substitutions/site/year), with HVR-1 having the greatest potential for variation [30–32].

Figure 2 shows the phylogenetic tree obtained by the neighbor-joining algorithm using the K2P model. A clear separation between sequences from the outbreak and those from control patients can be observed. Monophyly of the different sequences from each patient was obtained in all but 1 case. The exception corresponded to patients 35327 and 35373, both included in the outbreak, who were indistinguishable from each other because they shared several identical sequences and the remaining sequences differed in only 1 or 2 positions. These were also very closely related to sequences from patients 35329 and 35277, also from the outbreak.

A remarkable feature from the analysis is the very low genetic variability among sequences from each patient in the outbreak compared with control patients. A summary of the intrapatient variability is shown in table 1. All the measurements of genetic variation reported show a much larger variability among control patients than among those from the outbreak, thus suggesting a recent infection by HCV in the outbreak patients. This recent origin of infection can also be postulated for patient 1011, a control patient from Barcelona, whose variability values are somewhat intermediate between the control and the outbreak patients.

Despite the well-supported relationship among the outbreak sequences, the phylogenetic analysis developed so far can only provide a global evaluation of their common origin. For most purposes this is enough, but things are not so clear when an individualized evaluation is needed. This evaluation is appro-

Table 1. Summary of genetic variability among sequences of the hepatitis C virus E1–E2 region from each patient.

Patient	No. of clones	*S*	No. of haplotypes	*D*	π
Control					
1011	11	8	7	0.818	0.00308
1105	11	27	11	1.000	0.01522
1205	11	47	10	0.982	0.03567
1207	10	29	10	1.000	0.01521
1310	10	16	9	0.978	0.00880
1507	10	66	10	1.000	0.05824
2106	9	17	9	1.000	0.01071
2309	10	24	10	1.000	0.01502
2505	11	26	9	0.945	0.01448
354298	10	22	10	1.000	0.01177
372172	10	46	10	1.000	0.03249
Outbreak					
35277	10	1	2	0.356	0.00075
35279	12	4	4	0.455	0.00170
35327	10	3	3	0.378	0.00127
35329	13	6	5	0.538	0.00272
35330	12	4	4	0.561	0.00193
35373	12	5	5	0.576	0.00177

NOTE. *D*, haplotype diversity; *S*, no. of polymorphic sites; π, nucleotide diversity (nucleotide substitutions/site).

Table 2. Likelihood ratio test results for outbreak patients.

Patient	ln*L*	Δln*L*	Probability
35277	−4706.581	−21.291	5.668×10^{-10}
35279	−4699.187	−13.898	9.200×10^{-7}
35327	−4713.804	−28.514	4.134×10^{-13}
35329	−4708.155	−22.865	1.174×10^{-10}
35330	−4689.374	−4.084	1.680×10^{-2}
35373	−4713.804	−28.514	4.134×10^{-13}

priately done under a likelihood framework, as described above. To obtain accurate estimates of the actual number of nucleotide substitutions, which is essential for all methods of phylogenetic reconstruction and for the computation of the likelihood of every phylogenetic alternative, several factors must be taken into account, most notably the frequency of the 4 bases and its equilibrium state among the sequences analyzed and the variation in evolutionary rates among different positions in the alignment. Although possible, it is computationally very demanding to estimate all these parameters simultaneously with the phylogenetic reconstruction using a large number of sequences, as in this case. However, it is possible to obtain a very good estimate of the relevant parameters by using an approximate phylogeny obtained from a simple evolutionary model [33, 34]. Therefore, we used the neighbor-joining tree described above as the initial tree for deriving the rate of evolution at each nucleotide position, using the DNArates program.

Except for the JC model, combinations of 3 parameters were tested for each model (only 2 for K2P), the nucleotide frequencies (all equal to 0.25 or the empirical estimates), the transition/transversion ratio, and the rate categories (the same for all positions or 8 different categories). The substitution model with the highest log-likelihood value was HKY85 with empirical frequencies of the bases (A = 0.199, C = 0.302, G = 0.292, and T = 0.206), 8 different categories of evolutionary rates, and a ratio of transition to transversion equal to 3.2. The relative probabilities of substitution for each category were 0.317 (226 positions), 0.547 (28), 0.936 (39), 1.601 (45), 2.739 (72), 4.685 (42), 8.015 (14), and 10.338 (6). In all cases there is a very large improvement in the adequacy of the model when evolutionary rates are allowed to vary among positions, and this is reflected in increased log-likelihood values. Also, base composition is not very different from the equidistribution of the 4 bases, which explains the high log-likelihood value for the K2P model (data not shown).

Once the most adequate evolution model was determined, we proceeded to evaluate the probability of the inclusion of each patient in the group defined as the outbreak. In this case, the definition of the outbreak seems very easy, given the close relationship among the sequences derived from the patients, but this is not necessarily always true. As described above, we tested the alternative hypothesis of all the sequences derived from each patient were closer, in the phylogenetic reconstruction, to the remaining control sequences than to those in the monophyletic group defining the outbreak. The results are expressed as the LR between this and the null hypothesis—that is, that those sequences are actually related to the outbreak. Table 2 shows the results of the test for the 6 outbreak patients. In all situations, the derived probability coefficient for the LR (the inverse of the value shown in the probability column in table 2) will significantly increase the probability of considering the corresponding patient included in the outbreak under a statistical forensic setting (see Materials and Methods). These values range from 0.0168 to 4.134×10^{-13}, but they represent minimum values (i.e., most conservative) under the testing scheme proposed.

A similar, complementary analysis was performed with the control patients, considering the alternative that they could belong to the cluster defined by the outbreak. The corresponding results are shown in table 3. The probability values associated to the LRs range from 0.216 to 2.713×10^{-10}. Again, these represent minimum values of the probability to be multiplied by the ratio derived from other evidences and cannot be taken as absolute estimates of the probability of any given hypothesis.

DISCUSSION

We have proposed a new method for incorporating the statistical evaluation of alternative phylogenetic hypothesis into the forensic evaluation of a nosocomial infection by HCV. Although phylogenetic evidence has occasionally been considered in court [1] and has provided support in epidemiological studies [2, 4], this represents the first case in which molecular phylogenetics has been used to single out the likelihood of a patient sharing

Table 3. Likelihood ratio test results for control patients.

Patient	ln*L*	Δln*L*	Probability
354298	−4698.263	−12.974	2.320×10^{-6}
372172	4694.256	−8.966	1.276×10^{-4}
1205	−4697.702	−12.413	4.065×10^{-6}
1207	−4696.449	−11.160	1.423×10^{-5}
1310	−4707.317	−22.028	2.713×10^{-10}
1507	−4695.678	−10.389	3.078×10^{-5}
2106	−4701.668	−16.378	7.700×10^{-8}
2309	−4692.125	−6.835	1.075×10^{-3}
2505	−4696.431	−11.142	1.449×10^{-5}
1101	−4685.290	−11.641	8.796×10^{-6}
1105	−4686.822	−1.532	0.216

the source of virus infection with other infected patients, instead of considering the joint evaluation of the existence of a monophyletic clade encompassing several patients related to a common source. A precedent for the application of likelihood testing to deciding between 2 alternative sources of infection for a patient with HIV can be found in Holmes et al. [35]. This procedure can lead to uncertain results when, because of the extreme variability of some RNA virus, some clone or clones isolated from a single or different patients fail to group in a monophyletic clade, with the remaining sequences considered to be part of the common outbreak. In this case, the usual procedures for statistical evaluation of common ancestry, such as the bootstrap support for the relevant node defining the monophyletic clade, become useless. The difference is also especially relevant with court plaintiffs, for whom individual demands have to be considered separately.

Maximum likelihood testing is currently the best statistical method for the evaluation of competing hypotheses, and it can be readily applied to forensic analysis. Furthermore, maximum likelihood testing is also used to derive phylogenetic trees under specific patterns of substitution. However, despite recent developments [36, 37], its application for large data sets is hampered by computational restrictions. Ideally, maximum likelihood should be used during the whole process, starting with the choice of the most likely model of nucleotide substitution [38] and then using this with PAUP [39], PAML [40], or TREE-PUZZLE [37] to infer the most likely tree. Alternatively, Monte Carlo–Markov chain-based methods [41] could also be used to obtain the initial phylogenetic tree. All these implementations also allow the computation of the likelihood of alternative, user-defined trees, hence providing the means to test phylogenetic hypothesis. Apart from the LR applied in the forensic framework of the present work, Kishino-Hasegawa [42] or Shimodaira-Hasegawa [43] tests could be used for testing between the different hypotheses. Nevertheless, we must emphasize that, in the forensic context that frames our analysis, we are not interested in deciding which hypothesis is most likely or to be preferred. Rather, we are assigning a factor, the LR, to transform the prior probability of the 2 alternative hypotheses (usually, the prosecution and the defense propositions). If both hypotheses have similar likelihoods (i.e., the alternative hypothesis is not significantly better than the null hypothesis, given the empirical data), the analysis does not change the prior value. On the contrary, the more significantly better explanation of the data provided by the alternative hypothesis, the larger the LR, and the more significant the alteration of the alternative hypothesis will be introduced. The final evaluation depends both on the phylogenetic analysis and on other sources of evidence in each case. This has to be borne in mind continuously. For instance, control patient 1105 presents a LR ($\triangle \ln L = -1.532$; $P = .216$) that places him very close to the outbreak. However, this is only so if there is other evidence that could place him in this realm (i.e., if he had attended the aforementioned hemodialysis unit, which he did not). Also, this low ratio value is a direct consequence of the conservativeness of the procedure. The phylogenetic tree shown in figure 2 represents this patient at the base of the control patients' cluster, closest to the outbreak patients. The alternative hypothesis, depicted in figure 1, is constructed by placing all the sequences from each patient in a separate cluster at the node that separates outbreak and control samples in the original phylogenetic tree. For patient 1105, this corresponds to shifting its position to the precedent node in the tree, which results in a very low change in the total likelihood estimate.

Although in the specific instance considered in the present analysis, all the individuals initially considered as belonging to a single outbreak turned out to be so (supported by our molecular analysis), there are many situations in which this is not the case, and, among an unknown number of individuals sharing an identified risk factor, such as the presence and use of a contaminated equipment, there might be a fraction for whom other sources of infection are actually responsible for their condition. When this situation affects a large number of patients and social and/or economic questions interfere with the clinical and epidemiological enquiry, it is very difficult to differentiate between patients who share a common variant and those who simply share a common risk with the former but who have been infected by different means. The method of analysis we have proposed on the basis of the study of as large a number of highly variable sequences per patient as possible, the inclusion of unrelated population controls, and detailed phylogenetic and statistical analysis of both kinds of alternative hypothesis will allow the determination of which patients belong to which group and, furthermore, make it possible to set a quantitative estimate of the reliability of the assignation for each patient.

There are 2 further points to be stressed in the molecular epidemiological analysis of highly variable sequences such as those of RNA viruses. First, it is necessary to include control samples from the source population. In the present study, this was not feasible—most HCV-infected patients attending the hemodialysis unit in the Andorran hospital are not Andorran residents but actually are visitors mostly from different Spanish cities, with a high proportion from Catalonia. As a consequence, we used as controls individuals from the reference hospital for hepatitis C in this Spanish region. Second, the choice of genomic region has to be based on the evolutionary divergence levels for the problem under consideration. In this case, all the patients had been recently infected, and only a rapidly evolving region could provide enough variability for the analysis. In other cases, a slower region in the virus genome might also be adequate.

Acknowledgments

We thank M. Torres and N. Jiménez for technical support; the Servicio Central de Soporte a la Investigación Experimental–Servicio de Secuenciación, Universitat de Valencia, for sequencing; and M. Coll, A. Pérez, and J. I. Esteban for providing the samples used in the study.

References

1. Ou CY, Ciesielski CA, Myers G, et al. Molecular epidemiology of HIV transmission in a dental practice. Science **1992**; 256:1165–71.
2. Power JP, Lawlor E, Davidson F, Holmes EC, Yap PL, Simmonds P. Molecular epidemiology of an outbreak of infection with hepatitis C virus in recipients of anti-D immunoglobulin. Lancet **1995**; 345:1211–3.
3. Esteban JI, Martell M, Carman W, Gómez J. The impact of rapid evolution of the hepatitis viruses. In: Domingo E, Holland JJ, Webster JR, eds. Origin and evolution of viruses. San Diego: Academic Press, **1999**:345–76.
4. Izopet J, Pasquier C, Sandres K, Puel J, Rostaing L. Molecular evidence for nosocomial transmission of hepatitis C virus in a French hemodialysis unit. J Med Virol **1999**; 58:139–44.
5. DeBry RW, Abele LG, Weiss SH, et al. Dental HIV transmission? Nature **1993**; 361:691.
6. Hillis DM, Huelsenbeck JP. Support for dental HIV transmission. Nature **1994**; 369:24–5.
7. Crandall KA. Intraspecific phylogenetics: support for dental transmission of human immunodeficiency virus. J Virol **1995**; 69:2351–6.
8. Efron B. The jacknife, the bootstrap and other resampling plans. In: Social industry and applied mathematics. Monograph 38 in CBMS-NSF Regional Conference series in applied mathematics. Philadelphia: Society for Industrial and Applied Mathematics, **1982**.
9. Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. Evolution **1985**; 39:783–91.
10. Hillis DM, Huelsenbeck JP, Cunningham CW. Application and accuracy of molecular phylogenies. Science **1994**; 264:671–7.
11. Evett IW, Weir BS. Interpreting DNA evidence. Sunderland, MA: Sinauer, **1998**.
12. Alter MJ. The detection, transmission, and outcome of hepatitis C virus infection. Infect Agents Dis **1993**; 2:155–66.
13. Irie Y, Hayashi H, Yokozeki K, Kashima T, Okuda K. Hepatitis C infection unrelated to blood transfusion in hemodialysis patients. J Hepatol **1994**; 20:557–9.
14. Okuda K, Hayashi H, Kobayashi M, Irie Y. Mode of hepatitis C infection not associated with blood transfusion among chronic hemodialysis patients. J Hepatol **1995**; 23:28–31.
15. Jadoul M, Cornu C, Ypersele De Strihou C, Group UC. Incidence and risk factors for hepatitis C seroconversion in hemodialysis: a prospective study. Kidney Int **1993**; 44:1322–6.
16. Simon N, Courouce AM, Lemarrec N, Trepo C, Ducamp S. A twelve year natural history of hepatitis C virus infection in hemodialyzed patients. Kidney Int **1994**; 46:504–11.
17. Cohen J. The scientific challenge of hepatitis C. Science **1999**; 285:26–30.
18. Poynard T, Ratziu V, Benhamou Y, Opolon P, Cacoub P, Bedossa P. Natural history of HCV infection. Baillieres Best Pract Res Clin Gastroenterol **2000**; 14:211–28.
19. Lu M, Kruppenbacher J, Roggendorf M. The importance of the quasispecies nature of hepatitis C virus (HCV) for the evolution of HCV populations in patients: study on a single source outbreak of HCV infection. Arch Virol **2000**; 145:2201–10.
20. Domingo E, Escarmís C, Sevilla N, et al. Basic concepts in RNA virus evolution. FASEB J **1996**; 10:859–64.
21. Kwok S, Higushi R. Avoiding false positives with PCR. Nature **1989**; 339:237–8.
22. Staden R, Beal KF, Bonfield JK. The Staden package. Methods Mol Biol **2000**; 132:115–30.
23. Thompson JD, Higgins DG, Gibson TJ. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res **1994**; 22:4673–80.
24. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol **1987**; 4:406–25.
25. Kimura M. A simple method for estimating rate of base substitution through comparative studies of nucleotide sequences. J Mol Evol **1980**; 16:111–20.
26. Jukes TH, Cantor CR. Evolution of protein molecules. In: Munro HN, ed. Mammalian protein evolution. New York: Academic Press, **1969**: 21–132.
27. Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol **1981**; 17:368–76.
28. Hasegawa M, Kishino H, Yano T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol **1985**; 22: 160–74.
29. Olsen GJ, Matsuda H, Hagstrom R, Overbeek R. FASTDNAML: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. Comput Appl Biosci **1994**; 10:41–8.
30. Maertens G, Stuyver L. Genotypes and genetic variation of hepatitis C virus. In: Harrison TJ, Zuckerman AJ, eds. The molecular medicine of viral hepatitis. New York: Wiley, **1997**:183–233.
31. Ogata N, Alter HJ, Miller RH, Purcell RH. Nucleotide sequence and mutation rate of the H strain of hepatitis C virus. Proc Natl Acad Sci USA **1991**; 88:3392–6.
32. Okamoto H, Kojima M, Okada S. Genetic drift of hepatitis C virus during an 8.2-year infection in a chimpanzee: variability and stability. Virology **1992**; 190:894–9.
33. Yang Z, Goldman N, Friday A. Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. Syst Biol **1995**; 44: 384–99.
34. Anderson JP, Rodrigo AG, Learn GH, et al. Substitution model of sequence evolution for the human immunodeficiency virus type 1 subtype B gp120 gene over the C2-V5 region. J Mol Evol **2001**; 53:55–62.
35. Holmes EC, Zhang LQ, Simmonds P, Smith Rogers A, Leigh Brown AJ. Molecular investigation of human immunodeficiency virus (HIV) infection in a patient of an HIV-infected surgeon. J Infect Dis **1993**; 167:1411–4.
36. Brauer MJ, Holder MT, Dries LA, Zwickl DJ, Lewis PO, Hillis DM. Genetic algorithms and parallel processing in maximum-likelihood phylogeny inference. Mol Biol Evol **2002**; 19:1717–26.
37. Schmidt HA, Strimmer K, Vingron M, von Haeseler A. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics **2002**; 18:502–4.
38. Posada D, Crandall KA. MODELTEST: testing the model of DNA substitution. Bioinformatics **1998**; 14:817–8.
39. Swofford DL. PAUP*. Phylogenetic inference using parsimony (*and other methods). Version 4.0b1. Sunderland, MA: Sinauer, **1998**.
40. Yang Z. Phylogenetic analysis by maximum likelihood (PAML). Version 3.0c. London: University College, **2001**.
41. Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP. Bayesian inference of phylogeny and its impact on evolutionary biology. Science **2001**; 294:2310–4.
42. Kishino H, Hasegawa M. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J Mol Evol **1989**; 29:170–9.
43. Shimodaira H, Hasegawa M. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol **1999**; 16: 1114–6.