BULLETIN OF THE POLISH ACADEMY OF SCIENCES
TECHNICAL SCIENCES
Vol. 53, No. 3, 2005

Computational methods in diagnostics of chronic hepatitis C

P. KĘDZIORA[2], M. FIGLEROWICZ[3], P. FORMANOWICZ[1,2], M. ALEJSKA[1], P. JACKOWIAK[1], N. MALINOWSKA[1], A. FRĄTCZAK[1], J. BŁAŻEWICZ[1,2], and M. FIGLEROWICZ[1*]

[1]Institute of Bioorganic Chemistry, Polish Academy of Sciences, 12/14 Noskowskiego Str., 61-704 Poznań, Poland
[2]Institute of Computing Science, Poznań University of Technology, 3A Piotrowo Str., 60-965 Poznań, Poland
[3]Department of Infectious Diseases, University of Medical Sciences, 27/33 Szpitalna Str., 60-572 Poznań, Poland

Abstract. Despite the considerable progress that has recently been made in medicine, the treatment of viral infections is still a problem remaining to be solved. This especially concerns infections caused by newly emerging patogenes such as: human immunodeficiency virus, hepatitis C virus or SARS-coronavirus. There are several lines of evidence that the unusual genetic polymorphism of these viruses is responsible for the observed therapeutic difficulties. In order to determine whether some parameters describing a very complex and variable viral population can be used as prognostic factors during antiviral treatment computational methods were applied. To this end, the structure of the viral population and virus evolution in the organisms of two patients suffering from chronic hepatitis C were analyzed. Here we demonstrated that phylogenetic trees and Hamming distances best reflect the differences between virus populations present in the organisms of patients who responded positively and negatively to the applied therapy. Interestingly, the obtained results suggest that based on the elaborated method of virus population analysis one can predict the final outcome of the treatment even before it has started.

Key words: RNA sequence analysis, phylogenetic trees, combinatorial methods, hepatitis C virus, chronic hepatitis C diagnostics, interferon alpha and ribavirin therapy.

1. Introduction

Studying nucleotide and amino acid sequences is one of the central issues in computational biology. By analyzing and comparing information encoded in DNA sequences one can gain more knowledge about the relation between different species, the genetic diversity and the evolution of living organisms. The latter process is especially difficult to investigate because of its unusually long duration. This general principle does not, however, apply to those viruses which, contrary to all other living creatures, use not DNA but RNA to store their genetic information. There are several lines of evidence that RNA-based viruses can rapidly evolve and adapt to changing environmental conditions [1]. Studies conducted during the last two decades demonstrated that an individual RNA virus does not form a homogenous population but circulates in a host organism as a pool of genetically distinct variants. To describe such a complex structure of virus population the concept of quasi-species was applied [2,3]. The virus quasi-species can be defined as a set of phylogenetically related variants, which are present in a single infected organism. One of the models of virus quasi-species assumes that it is a cloud in sequence space formed by all possible variants of a genomic sequence. For an average single stranded RNA virus, which genome is composed of 10000 nucleotides (nt), the cardinality of sequence space is equal to 4^{10000}. The cloud representing a quasi-species' distribution usually moves within the sequence space as long as new viral variants are produced and subjected to a continuous process of competition and selection. If a positive selection is applied, the process of differentiation of virus population is accelerated and certain mutations may become fixed. On the other hand, negative selection can reduce genetic polymorphism by eliminating some viral variants. Analyzing viral population developing in the infected organism one can find that it is usually dominated by one or several master variants, which are accompanied by substantial amounts of closely related ones that replicate almost as fast as the master.

Right now, it is becoming increasingly clear that the unusual genetic polymorphism of viruses might be responsible for most of the problems, which are encountered during the treatment of RNA virus infections [4]. This observation seems to be of special importance with regard to chronic infections such as those caused by hepatitis C virus (HCV). According to the WHO the number of people infected with HCV is continuously growing and has recently reached over 170 million worldwide – this problem also concerns c.a. 700 thousand Poles [5]. The clinical course of acute HCV infection is asymptomatic in about 70% of patients, and therefore the disease frequently remains unrecognized for many years. However, in 70–85% of the infected, the development of chronic hepatitis C (CHC) is observed [6–8].

Long – term observation of adults suffering from CHC demonstrated that about 20 years post-infection, cirrhosis develops in c.a. 20–30% of patients. After the next 5 years, 20% of them suffer from hepatic decompensation, and 10% from hepatocellular carcinoma [9,10]. Initially, interferon alpha (IFN), which displays antiviral and immunomodulating effects was the basic therapeutic used in CHC. In adults it gave a very poor result: only 10–20% eliminated the virus. Lately, a new method of CHC treatment involving two therapeutics,

*e-mail: marekf@ibch.poznan.pl

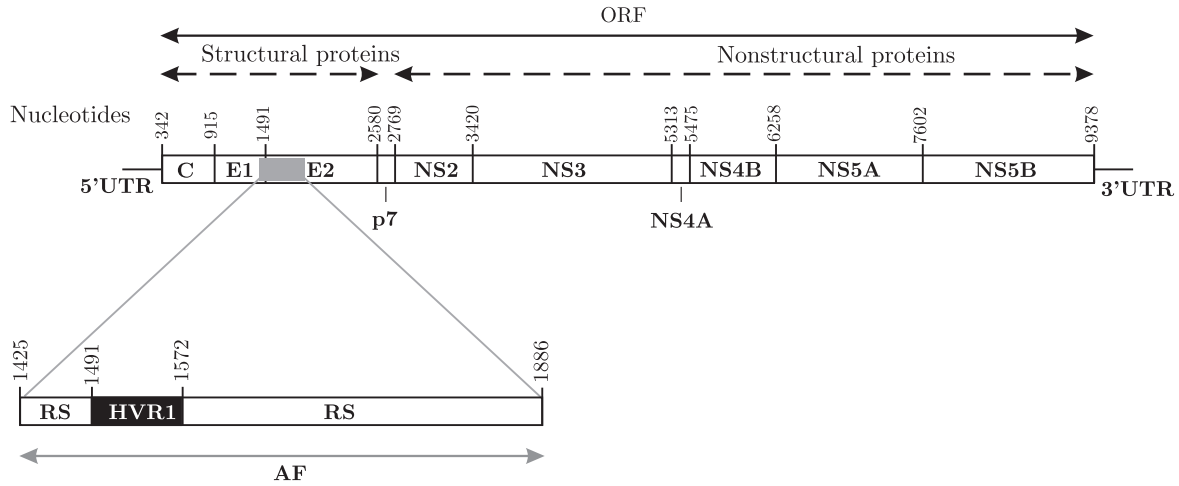

Fig. 1. Hepatitis C virus genome organization. HCV has a single stranded ∼9600 nucleotides long genomic RNA molecule. The genome contains one open reading frame (ORF, boxed and denoted by black solid arrow), which is flanked by 5' and 3' untranslated regions (UTRs, horizontal lines). The scheme shows the localization of individual segments coding viral structural and nonstructural proteins. Shaded box indicates the subsequence chosen for analysis (AF, grey solid arrow). The nucleotides spanning positions from 1425 to 1886 comprise relatively stable sequence (RS, white rectangle) and hypervariable region (HVR1, black box)

IFN and ribavirin, has been introduced. Ribavirin displays a wide spectrum of antiviral actions, including enhancing host immune response and increasing the rate of RNA mutations, in addition to its inhibitory effect on RNA synthesis [11,12]. The combination of IFN and ribavirin gives much better therapeutic effects than IFN monotherapy. A permanent elimination of HCV (sustained response – SR) is observed in 40–50% of the treated [13–15]. Interestingly, despite extensive investigations, no obvious correlation between the parameters characterizing the treated patient and the final outcome of the therapy has been established so far. The collected data suggested that the specific properties of the virus may actually be a major factor affecting the course of chronic HCV infection.

To better recognize this problem we decided to apply combinatorial methods for detailed analysis of the changes occurring in HCV population during the first two weeks of the IFN-ribavirin treatment. Thus, the major goal of the studies was to determine which parameters describing the virus population could be used to predict the final outcome of the antiviral therapy.

The organization of the paper is as follows. In the next section some basic properties of the hepatitis C virus are described. In Section 3 the analysis of the virus populations with the use of combinatorial approaches and involving phylogenetic trees, is discussed in details. The paper ends with conclusions given in Section 4.

2. Hepatitis C virus

The discovered in 1989 hepatitis C virus belongs to the family *Flaviviridae*. A more detailed analysis of the viral genome allows to distinguish six major HCV genotypes [16]. Viral particles are enveloped and comprise single stranded genomic RNA composed of about 9600 nucleotides. The genome contains one large open reading frame (ORF) encoding more than 3000 amino acids long polyprotein. The ORF is flanked by two untranslated regions (UTRs), which perform important regulatory functions during viral genome replication and translation (Fig. 1). The polyprotein precursor is co- and posttranslationally cleaved to yield ten proteins, divided according to their functions into two groups: structural and nonstructural (NS) ones. The former group includes proteins, which are key components of viral particle: core protein (C) and two envelope glycoproteins (E1 and E2). The latter consists of NS2, NS3, NS4A, NS4B, NS5A and NS5B, involved in HCV replication. The function of the protein p7 remains unknown [17,18].

The main source of HCV genetic variability are point mutations: single nucleotide substitutions, deletions or insertions. They are introduced to the viral genome during its replication by virus encoded RNA-dependent RNA polymerase. This process is highly error-prone: it was established, that the mutation frequency is approximately 10^{-4}–10^{-5} per incorporated nucleotide [19,20]. Considering the number of progeny virions produced daily in the infected organism (10^{12} HCV particles) and the length of HCV genomic RNA (10^4 nt), it was calculated that each individual nucleotide can be exchanged 10^5–10^8 times during one day of infection [4]. This result clearly shows how enormous the evolutionary potential of RNA viruses is. Each day huge number of new viral variants is generated. Because they are subjected to strong selection only those best fitted (well-replicating, infectious, nonneutralizable by the host immune system and resistant to antiviral drugs) can survive and spread.

Theoretically point mutations should be randomly distributed along the viral genome. However, experimental data show that some segments are more variable than others. In HCV, the 3' and 5' UTRs are most conserved (more than 90% of homology between viral isolates), because they are involved in the regulation of the viral life cycle [21–23]. In contrast,

Table 1
Brief characteristics of 2 patients subjected to combined interferon-ribavirin therapy, their age, sex and response to applied treatment (Ti denotes the i-th week of the therapy)

Patient	Age (years)	Sex	HCV RNA concentration in serum (copies/ml)				Result of the therapy
			T_0	T_{24}	T_{48}	T_{72}	
N	12	female	804 600	283 500	725 900	894 000	non response
P	11	female	337 500	not detected	not detected	not detected	sustained response

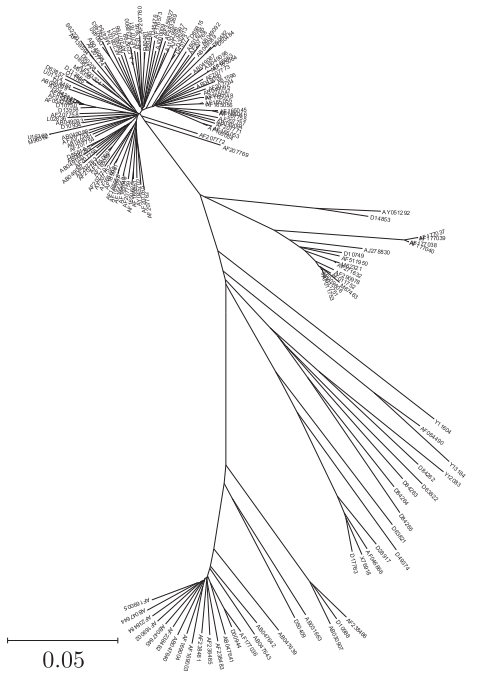

Fig. 2. Phylogenetic tree constructed using complete amino acid sequence encoded in the HCV genome

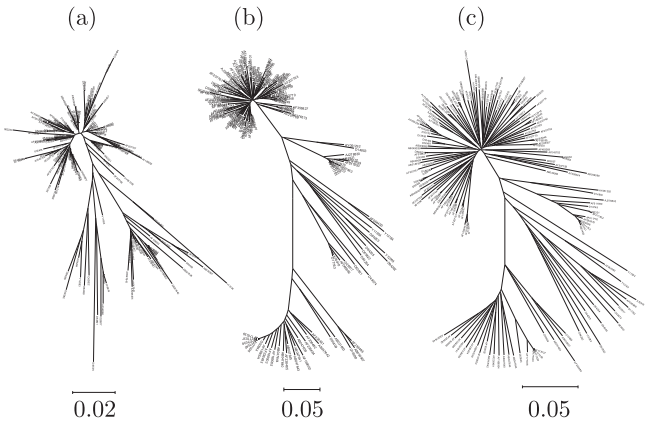

Fig. 3. Phylogenetic trees constructed using complete amino acid sequence of structural proteins: 3a – C protein, 3b – E1 protein, 3c – E2 protein

other parts of HCV genomic RNA are much more heterogeneous, especially those encoding envelope proteins E1 and E2. In the E2 coding sequence a 81 nt hypervariable region (HVR1) was distinguished [24].

3. The analysis of the virus population

3.1. The outline of the combinatorial analysis. The main goal of the undertaken HCV genome analysis is to determine which parameters characterizing viral population could be applied as prognostic factors during CHC treatment. To this end, two patients with a similar history of CHC development were chosen. Both of them had been infected with the same HCV genotype (1a). Duration of the CHC and level of virus accumulation were also similar (4 years and 10^5 copies of HCV RNA/ml, respectively). However, the response of the two patients to the IFN-ribavirin therapy was different. One of them responded positively (patient P with the sustained response), whereas the other failed to respond (patient N). According to commonly used standards, the response to antiviral therapy was considered positive if viral RNA was below the detection level at 24, 48 and 72 week of the treatment, and negative if viral RNA was present at these timepoints. Brief characteristics of the patients are shown in Table 1. In order to analyze viral population, HCV RNA was isolated from patients' blood before starting the IFN-ribavirin treatment (t_0) and after two weeks of the therapy (t_f). Then, 20 viral genomes were randomly selected from each sample. The studies were conducted with the permission of the Bioethical Commission of the University of Medical Sciences in Poznań (Permission No 712/02 of 05.09.2002). Since the sequencing and analysis of the whole HCV genome would be very difficult, a representative subsequence was chosen, according to the procedure described in the next subsection.

3.2. Choosing a subsequence of HCV for the analysis. Usually the evolutionary history of groups of species is studied not on the base of the whole genomes but on their short fragments. One of the motivations for such an approach is the intractability of the computational problem, which in practice leads to a rapid increase of the required computation time when the number of the analyzed sequences and their lengths increase. The other reason is a limited availability of the whole genome sequences. The interesting phenomenon is that the evolutionary history usually varies depending which genome segment has been selected. Therefore the first step in our research was determining which segment of the HCV genome can be used as the most representative one for the whole genome. The sequencing of 40 complete HCV genomes for each patient would be very time and cost consuming. Hence, it has been decided to select a region, which fulfils the following criteria: (i) can be sequenced in a single reaction – which means that it should be no more than 500 nucleotides long,

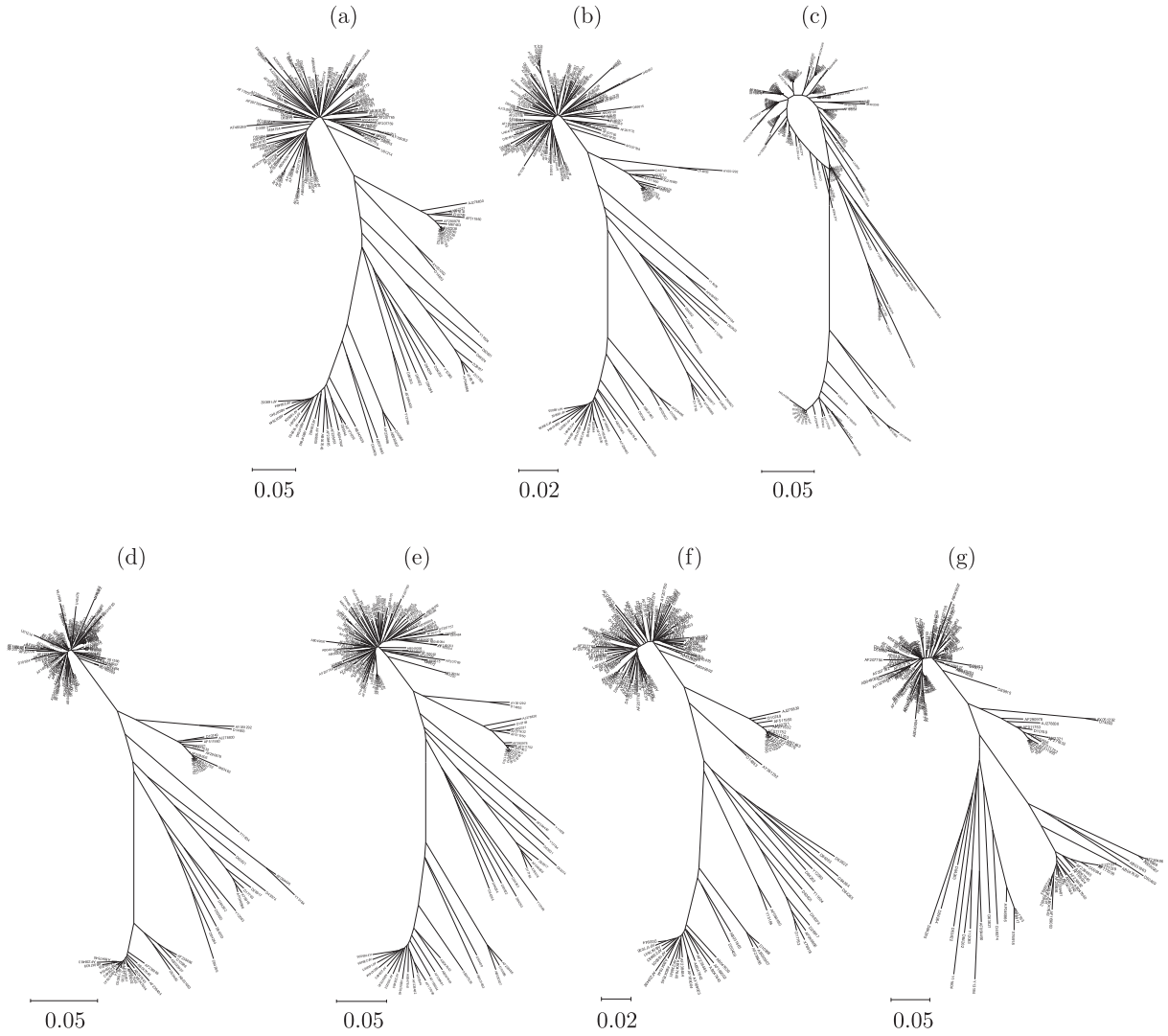

Fig. 4. Phylogenetic trees constructed using complete amino acid sequence of non-structural proteins: 4a – NS2 protein, 4b – NS3 protein, 4c – NS4A protein, 4d – NS4B protein, 4e – NS5A protein, 4f – NS5B protein, 4g – p7 protein

(ii) its evolution history is similar to that of the whole genome – phylogenetic tree constructed for this region is similar to the tree obtained for the entire genome, (iii) well reflects the heterogeneity of the HCV genome – in other words, it contains both variable and more stable regions. In order to choose the proper subsequence, over 160 HCV genomes taken from a database located at `http://s2as02.genes.nig.ac.jp.` were analyzed. The phylogenetic trees were built for the whole genomes and then for each virus protein separately. The trees were created using Neighbor Joining method implemented in MEGA package, ver. 3.0 [25]. The analysis was made on the basis of nucleotide as well as amino acid sequences, however, no significant differences between the two types of results have been found. The trees built for amino acid sequences are shown in Figs 2, 3 and 4, respectively, whole genome, structural and non-structural proteins. As one can notice, the trees for E1, E2, NS2, NS3, NS5A and NS5B proteins are quite similar to the tree obtained for the whole genome, whereas C, NS4A, NS4B and p7 are distinctly different. On the base of the comparison of the evolutionary trees, a subsequence encoding fragments of E1 and E2 proteins has been chosen for further analysis. Taking into account the length limitation (500 nt) and a variability of E1 and E2 coding regions, the sequence (called AF) located between positions 1425 and 1886 was selected as a representation for the entire HCV genome. AF spans HVR1 and more stable segments flanking this region (RS) (cf. Fig. 1).

3.3. Population of the virus. The next step in the analysis was the determination of the number of different variants of the virus RNA in the samples taken from the patients' organisms. This number has been determined on the base of three different subsequences, i.e.: (i) the whole subsequence determined in the previous subsection and denoted by AF; (ii) HVR1 being a part of the former subsequence, and (iii) the subsequence without HVR1, denoted by RS. The identification of the number of the variants has been made by the comparison of the analyzed sequences. In Figs 5, 6, and 7 multiplicities of each

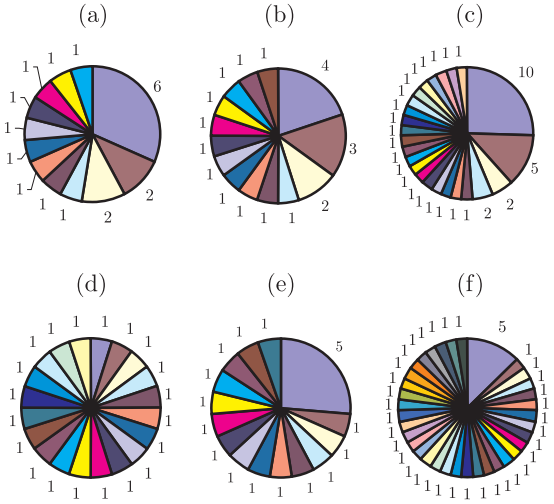

Fig. 5. The number of different variants of virus RNA in the population isolated from blood sample of patient N – determined for the entire considered region (AF): 5a – before starting the treatment (t_0); 5b – after two weeks of the therapy (t_f), 5c – the total number of the variants in the two populations of HCV (i.e. present in the patient's organism before the treatment and after two weeks of the therapy). The number of different variants of virus RNA in the population isolated from blood sample of patient P – determined for the entire considered region (AF): 5d – before starting the treatment (t_0); 5e – after two weeks of the therapy (t_f), 5f – the total number of the variants in the two populations of HCV (i.e. present in the patient's organism before the treatment and after two weeks of the therapy)

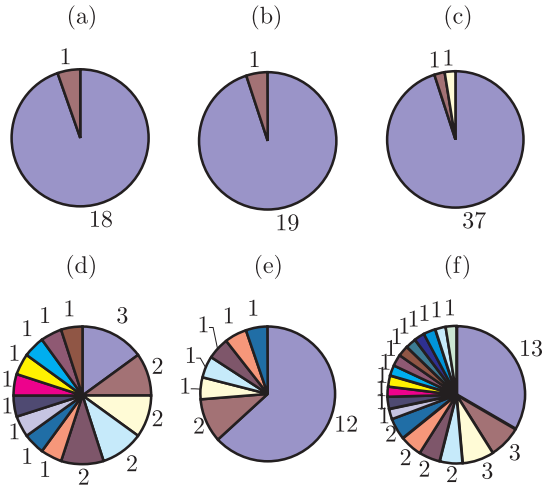

Fig. 6. The number of different variants of virus RNA in the sample of patient N – determined for HVR1 region: 6a – before starting the treatment; 6b – after two weeks of the therapy, 6c – the total number of the variants in the two populations of HCV (i.e. present in the patient's organism before the treatment and after two weeks of the therapy). The number of different variants of virus RNA in the sample of patient P – determined for HVR1 region: 6d – before starting the treatment; 6e – after two weeks of the therapy, 6f – the total number of the variants in the two populations of HCV (i.e. present in the patient's organism before the treatment and after two weeks of the therapy)

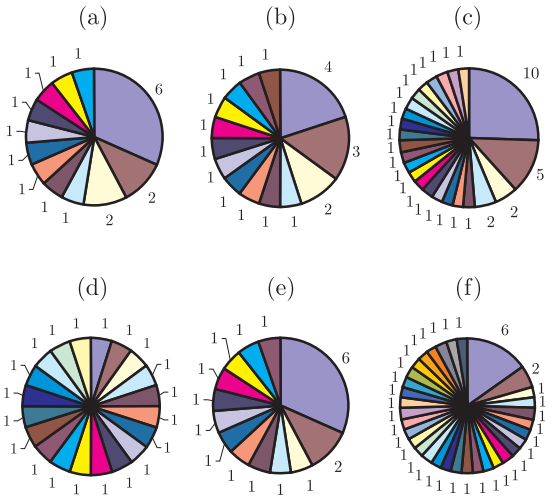

Fig. 7. The number of different variants of virus RNA in the sample of patient N – determined for the considered region without HVR1 (RS): 7a – before starting the treatment; 7b – after two weeks of the therapy, 7c – the total number of the variants in the two populations of HCV (i.e. present in the patient's organism before the treatment and after two weeks of the therapy). The number of identical clones of HCV in the sample of patient P – determined for the considered region without HVR1 (RS): 7d – before starting the treatment; 7e – after two weeks of the therapy, 7f – the total number of the variants in the two populations of HCV (i.e. present in the patient's organism before the treatment and after two weeks of the therapy)

of the detected variants are shown, where Fig. 5 presents the results obtained for the whole subsequence AF. In Fig. 6 the results for HVR1 are shown, while in Fig. 7 the numbers of identical clones for subsequence RS are presented.

As mentioned before, the populations have been analyzed at the beginning of the therapy and two weeks later. Finally, the total numbers of different sequence variants at these two moments have been determined.

The most obvious difference between the samples coming from the two patients can be observed in the case of the number of HVR1 variants (cf. Fig. 6). The samples taken from patient N consist of only three different sequences in this region. Sample taken from patient P at the beginning of the therapy contains 14 variants. This number has been reduced after two weeks of the treatment and in the second sample there were only seven variants, where one of them dominated the others.

In the case of the subsequence RS the situation is quite different. At the beginning of the therapy, 12 and 20 different variants of the sequence were found in samples from patient N and P, respectively (cf. Fig. 7). However, after two weeks in the sample coming from patient N the number of variants was equal to 14, while in the case of patient P there were 13 variants. In both cases there was one variant slightly dominating the others but in the sample taken from patient P the domination was a bit stronger.

So, it may be concluded that in the case of the analyzed virus populations the changes in HVR1 most clearly distinguish the two patients.

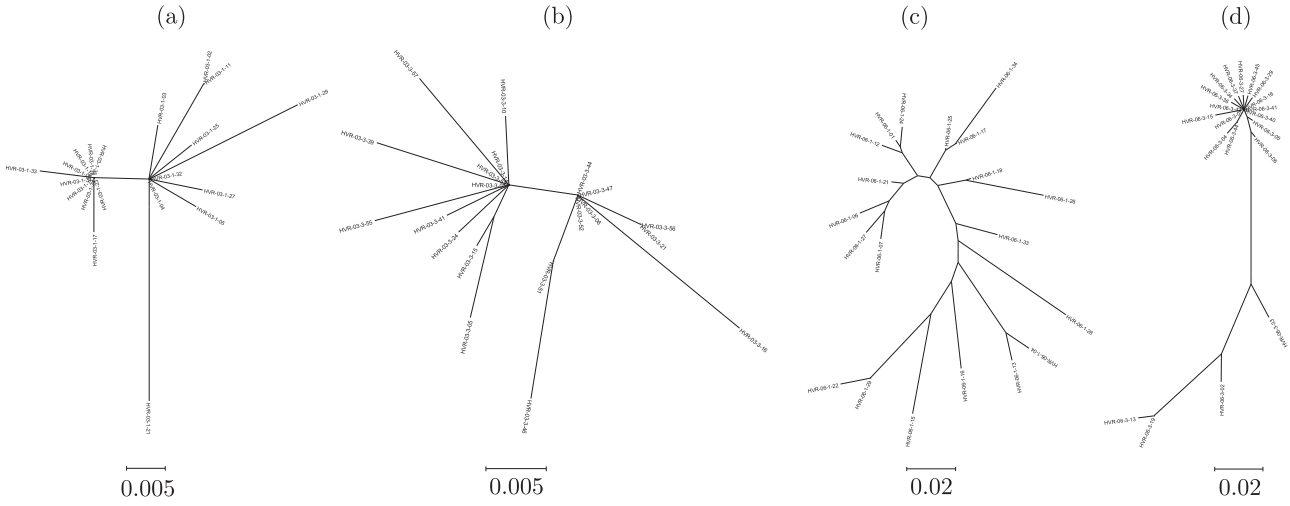

Fig. 8. The phylogenetic trees of HCV constructed based on the sequences taken from the sample of patient N – prepared for the entire considered region (AF): 8a – before starting the treatment; 8b – after two weeks of the therapy. The phylogenetic trees of HCV constructed based on the sequences taken from the sample of patient P – prepared for the entire considered region (AF): 8c – before starting the treatment; 8d – after two weeks of the therapy

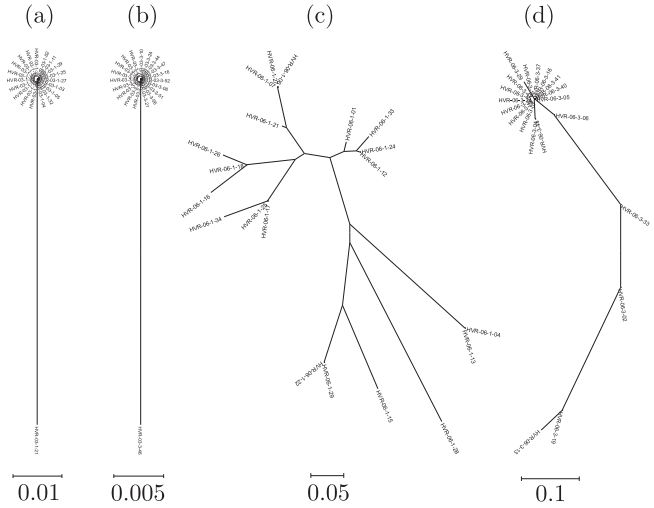

Fig. 9. The phylogenetic trees of HCV constructed based on the sequences taken from the sample of patient N prepared for HVR1 region: 9a – before starting the treatment; 9b – after two weeks of the therapy. The phylogenetic trees of HCV constructed based on the sequences taken from the sample of patient P – prepared for HVR1 region: 9c – before starting the treatment; 9d – after two weeks of the therapy

3.4. Evolutionary trees analysis. Although the knowledge about the number of different sequences in the samples is very important since it tells a lot about the genetic diversity of the virus, it tells nothing how closely the different genomes are related to each other. For example, there may be two quite different populations in the two patients' organisms, each of them consisting of the same number of genome types but in one of them the sequences may be very similar to each other while in the second population the genomes may differ a lot. Information of this kind may be very important since it may suggest that some variants of the genomes are better suited to survive in the host organism than the others. Clearly, such information can be discovered by constructing an evolutionary tree for each sample. Similarly like previously the trees have been constructed for the whole subsequence selected for the analysis, i.e. AF, as well as for HVR1 and the RS region. Again, Neighbor Joining method and MEGA package have been used. Figs. 8, 9, and 10, respectively, show the obtained results.

As one can notice, the most obvious difference between the trees constructed for the two patients may be observed in the case where HVR1 was the base for the analysis (cf. Fig. 9). In fact, in both samples taken from patient N there are only two virus variants, while in the case of patient P the evolutionary distances are quite large at the beginning of the treatment and then they are reduced to zero for a large subset of the population after two weeks. However, in this case there are still some variants of the virus which differ considerably from the dominating variant of the sequence.

The above result may be interpreted in such a way that the virus is "searching" for the best mutation which will allow it to resist against the immune system and the drugs applied. The virus present in patient N has "found" such an optimal mutation. However, it seems that the optimal variant of the virus has been present in the body of the patient N before the treatment and the therapy cannot force the virus to "search" for other variants since it is not necessary from the "viewpoint" of the virus. In the case of patient P such a good mutation was not present at the beginning of the therapy neither two weeks later, which makes probable the hypothesis that it will not be found in the next stages of the treatment (and in fact it was not found).

When the RS region is considered the difference between the ways of evolution of the two populations is not so obvious. However, also in this case it may be observed that in the population coming from patient N there are two subsets of the

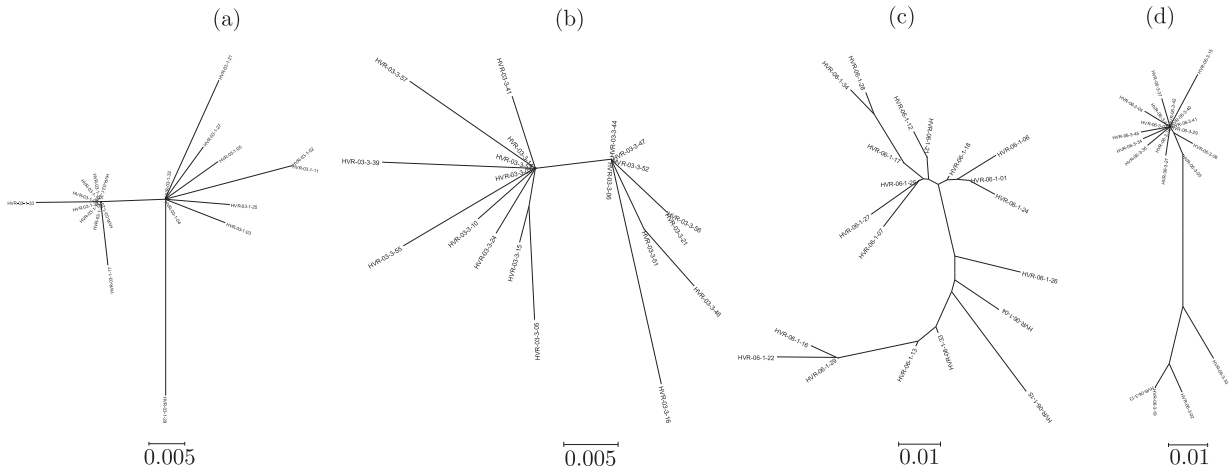

Fig. 10. The phylogenetic trees of HCV constructed based on the sequences taken from the sample of patient N – prepared for the considered region without HVR1 (RS): 10a – before starting the treatment; 10b – after two weeks of the therapy. The phylogenetic trees of HCV constructed based on the sequences taken from the sample of patient P – prepared for the considered region without HVR1 (RS): 10c – before starting the treatment; 10d – after two weeks of the therapy

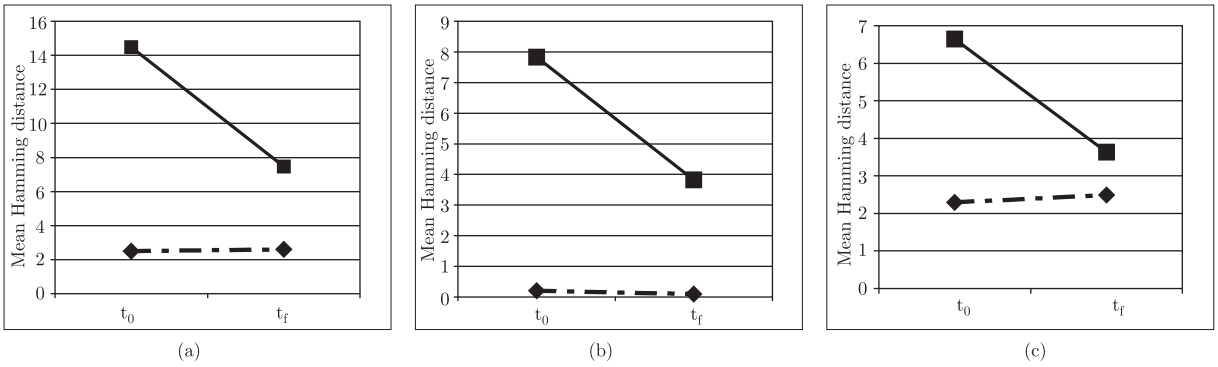

Fig. 11. The genetic diversity of HCV computed for patient P (solid line) and for patient N (dash dot line) before starting the treatment and after two weeks of the therapy. 11a – genetic diversity computed on entire considered region (AF), 11b – genetic diversity computed on HVR1, 11c – genetic diversity computed on considered region without HVR1 (RS)

sequences relatively closely related to each other, while the evolution of the population present in the organism of patient P is very similar to the evolution observed on the base of HVR1, but here there is no so obviously dominating variant of the sequence like in the case of HVR1.

3.5. Genetic diversity. Similar results to those obtained by the evolutionary trees analysis may be obtained by the calculation of the genetic diversity. In this case an average diversity in the analysed population is determined by calculating the Hamming distance between any pair of sequences, i.e. the number of amino acid differences between the two sequences. The mean Hamming distance being the average of the distance taken for all pairs of the analyzed sequences reflects the genetic diversity of the population.

The Hamming distance for sequences s_i and s_j is defined as

$$h(s_i, s_j) = \sum_{k=1}^{n} x(k)$$

where

$$x(k) = \begin{cases} 0 \text{ if } s_i(k) = s_j(k) \\ 1 \text{ otherwise} \end{cases}$$

$s_i(k)$ denotes character at position k in sequence s_i and n is the length of the sequences. The mean Hamming distance for set S of sequences is calculated according to the following formula:

$$H(S) = \frac{\sum\limits_{s_i, s_j \in S, s_i \neq s_j} h(s_i, s_j)}{\binom{|S|}{2}}.$$

Like previously, it has been calculated for all three segments, i.e. the AF subsequence, HVR1, and the RS region. The results are shown in Fig. 11.

Let us observe that the genetic diversity calculated in the way described above is an average of the evolutionary distances observed on the basis of the trees presented in the previous subsection. However, the most interesting fact observed in Fig. 11 is the change of the diversity after the two weeks of

the therapy. Clear differences between the two patients can be seen. Indeed, in the case of patient N, the diversity almost does not change not depending on the analyzed fragment of the subsequence, while the diversity decreases in all three cases, when patient P is considered.

The other difference is that in the population present in the organism of patient N the diversity is smaller than in the second population, especially at the beginning of the therapy.

4. Conclusions

Earlier Farci and coworkers applied computational analysis to test how IFN therapy affects the HCV population in patients with CHC [26]. They postulated that changes in the genetic diversity of HCV predict the therapeutic outcome. According to these authors positive response to the applied antiviral drug can be expected if the number of viral variants circulating in patient's blood is significantly reduced during the first two week of treatment.

The major aim of our studies was to determine whether some parameters characterizing the structure of the HCV population can be used to predict the final outcome of IFN-ribavirin therapy in patients suffering from CHC. To this end, subsequence AF representative for the entire HCV genome was identified. Samples of this subsequence were isolated from the blood of patients P and N just before therapy (20 samples for each patient) and then after two weeks of treatment (20 samples for each patient). Finally, all collected variants of subsequence AF were subjected to a detailed analysis. It was carried out for the entire subsequence AF as well as for its fragments HVR1 and RS. For every examined HCV population its genetic diversity was determined, a phylogenetic tree was constructed and Hamming distance was calculated. As a result, we found that HVR1 is the optimal segment to analyze, since parameters describing this region, best reflect the differences between viral populations present in patient P with sustained response and in patient N nonresponding to IFN-ribavirin therapy. The above presented data also indicate that it should be easier to predict a final outcome of IFN-ribavirin therapy based on phylogenetic trees and Hamming distances than on the number of virus variants in the analyzed populations. What is more, it seems that prediction can be done already before the treatment, not after two weeks – as it was postulated by Farci and coworkers [26]. Obviously, these preliminary results require verification in a broader experiment, which is a subject of our further research. Taking into account growing complexity of the problems analyzed, parallel approaches to phylogenetic tree construction could be used which may allow for an application of exact optimization algorithms like branch and bound method (cf. [27]).

Acknowledgements. This research was partially supported by the Polish Government through a grant (6 PO5A 054 21) from the State Committee for Scientific Research (KBN). P. J. is additionally supported by the Scholarship from the President of the Polish Academy of Sciences.

REFERENCES

[1] J. Holland, K. Spindler, F. Horodyski, E. Grabau, S. Nicol, and S. Van de Pol, "Rapid evolution of RNA genomes", *Science* 215, 1577–1585 (1982).

[2] J. Holland, J.C. DelaTorre, and D.A. Steinhauer, "RNA virus populations as quasispecies", *Genetic Diversity of RNA Viruses*, Springer Verlag, 1–20 (1992).

[3] D.B. Smith, J. McAllister, C. Casino, and P. Simmonds, "Virus 'quasispecies': making a mountain out of a molehill?", *J. Gen. Virol.* 78, 1511–1519 (1997).

[4] M. Figlerowicz, M. Alejska, A. Kurzyńska-Kokorniak, and M. Figlerowicz, "Genetic variability: the key problem in the prevention and therapy of RNA-based virus infections", *Med. Res. Rev.* 23(4), 488–518 (2003).

[5] World Health Organization, "Global surveillance and control of hepatitis C", *J. Med Virol.* 6, 35–47 (1999).

[6] M. Aniszewska, B. Kowalik-Mikolajewska, and S. Dobosz, "Hepatitis C – epidemiology, diagnosis and treatment in children", *Epidemic Review* 55(4), 503–510 (2001), (in Polish).

[7] M. Bendelini, M.L. Vatteroni, F. Maggi, and M. Pistello, "Hepatitis C virus", in *Viral Hepatitis*, 65–127 ed. S. Specter, Humana Press Inc., 1999.

[8] J B. Schwimmer and W.F. Balistreri, "Transmission, natural history, and treatment of hepatitis C virus infection in the pediatric population", *Semin. Liver Dis.* 20(1), 37–46 (2000).

[9] G. Fattovich and S.W. Schalm, "Hepatitis C and cirrhosis", in *Hepatitis C*, 241–264 eds. T.J. Liang, J.H. Hoofnagle, San Diego, Academic Press, 2000.

[10] A.M. Di Bisceglie, "Hepatitis C and hepatocellular carcinoma", in *Hepatitis C*, 265–276 eds. T. J. Liang , J. H. Hoofnagle, San Diego, Academic Press, 2000.

[11] S. Crotty, D. Maag, J.J. Arnold, W. Zhong, J.Y.M. Lau, Z. Hong, R. Andino, and C.E. Cameron, "The broad – spectrum antiviral ribonucleoside ribavirin is an RNA virus mutagen", *Nature Medicine* 6, 1375–1379 (2000).

[12] S. Crotty, C. Cameron, and R. Andino, "RNA virus error catastrophe: direct molecular test by using ribavirin", *PNAS 98*, 6895–6900 (2001).

[13] D. Landau, J.P. Batisse, C. Van Huyen, F. Piketty, G. Bloch, L. Pialoux, J.P. Belec, P. Petite, and L. Weiss, "Efficacy and safety of combination therapy with interferon – alpha 2b and ribavirin for chronic hepatitis C", *HIV – infected patients. AIDS* 14, 839–844 (2000).

[14] S.P. Lawrence, "Advances in the treatment of hepatitis C", *Adv. Intern. Med.* 45, 65–105 (2000).

[15] J.G. McHutchison and T. Poynard, "Combination therapy with interferon plus ribavirin for the initial treatment of chronic hepatitis C", *Semin. Liver Dis.* 19 (Suppl 1), 57–5 (1999).

[16] P. Simmonds, "Hepatitis C genotypes", in *Hepatitis C*, 25–51 eds. T. J. Liang , J.H. Hoofnagle, San Diego, Academic Press, 2000.

[17] K. Moriishi and Y. Matsuura, "Mechanisms of hepatitis C virus infection", *Antivir. Chem. Chemother* 14, 285–297 (2003).

[18] R. Bartenschlager and V. Lohmann, "Replication of hepatitis C virus", *J. Gen Virol.* 81, 1631–1648 (2000).

[19] D.B. Smith and S.C. Inglis, "The mutation rate and variability of eucaryotic viruses: an analytical review", *J. Gen. Virol.* 68, 2729–2740 (1987).

[20] D.A. Steinhauer and J.J. Holland, "Direct method for quantitation of extreme polymerase error frequencies at selected single base sites in viral RNA", *J. Virol.* 57, 219–228 (1986).

[21] P. Friebe and R. Bartenschlager, "Genetic analysis of sequences in the 3' nontranslated region of hepatitis C virus that are important for RNA replication", *J. Virol.* 76 (11), 5326–5338 (2002).

[22] P. Friebe, V. Lohmann, N. Krieger and R. Bartenschlager, "Sequences in the 5' nontranslated region of hepatitis C virus required for RNA replication", *J. Virol.* 75 (24), 12047–12057 (2001).

[23] P. Sarnow, "Viral internal ribosome entry site elements: novel ribosome-RNA complexes and roles in viral pathogenesis", *J. Virol.* 77 (5), 2801–2806 (2003).

[24] A.J. Weiner, H.M. Geysen, C. Christopherson, J.E. Hall, T.J. Mason, G. Saracco, F. Bonino, K. Crawford, C.D. Marion, and K.A. Crawford et al., "Evidence for immune selection of hepatitis C virus (HCV) putative envelope glycoprotein variants: potential role in chronic HCV infections", *PNAS* 89, 3468–3472 (1992).

[25] S. Kumar, K. Tamura, and M. Nei, "MEGA3: integrated software for molecular evolutionary genetic analysis and sequence alignment", *Briefings in Bioinformatics* 5 (2004), (to be published).

[26] P. Farci, R Strazzera, H.J. Alter, S. Farci, D. Degioannis, A. Coiana, G. Peddis, F. Usai, G. Serra, L. Chessa, G. Diaz, A. Balestrieri, and R.H. Purcell, "Early changes in hepatitis C viral quasispecies during interferon therapy predict the therapeutic outcome", *PNAS* 99, 3081–3086 (2002).

[27] J. Błażewicz, P. Formanowicz, P. Kędziora, and P. Wojciechowski, "Parallel algorithms for evolutionary history reconstruction", *Lecture Notes in Computer Science* 3019, 1138–1145 (2004).