Computational and Mathematical Methods in Medicine
Vol. 11, No. 2, June 2010, 185–199

Taylor & Francis
Taylor & Francis Group

Towards prediction of HCV therapy efficiency

Szymon Wasik[a]*, Paulina Jackowiak[b1], Jacek B. Krawczyk[c2], Paweł Kedziora[a3], Piotr Formanowicz[ab4], Marek Figlerowicz[b5] and Jacek Błażewicz[ab6]

[a]*Institute of Computing Science, Poznan University of Technology, Piotrowo 2, 60-965 Poznan, Poland;* [b]*Institute of Bioorganic Chemistry, Polish Academy of Sciences, Z. Noskowskiego 12/14, 61-704 Poznan, Poland;* [c]*Faculty of Commerce and Administration, Victoria University of Wellington, PO Box 600, Wellington, New Zealand*

(*Received 6 February 2009; final version received 30 June 2009*)

We investigate a correlation between genetic diversity of hepatitis C virus population and the level of viral RNA accumulation in patient blood. Genetic diversity is defined as the mean Hamming distance between all pairs of virus RNA sequences representing the population. We have found that a low Hamming distance (i.e. low genetic diversity) correlates with a high RNA level; symmetrically, high diversity corresponds to a low RNA level. We contend that the obtained correlation strength justifies the use of the viral RNA level as a measure enabling prediction of efficiency of an established therapy. We also propose that patient qualification for therapy, based on viral RNA level, improves its efficiency.

Keywords: HCV; genetic diversity; therapeutic efficiency; RNA levels; transition probabilities

AMS Subject Classification: 15A51; 62P10

1. Introduction

The aim of this paper is to propose a method for early-stage hepatitis C virus (HCV) patients' assessment, under which predictions can be made about efficiency of a treatment.

HCV is one of the most prevalent human pathogens, infecting more than 170 million people worldwide [12]. Similar to other RNA-based viruses, it exists as a quasispecies in an individual organism [3,11], i.e. as a pool of phylogenetically related but genetically slightly distinct variants. With its capacity for long-lasting persistence in the host, HCV causes chronic infections that can lead to liver fibrosis, cirrhosis and hepatocellular carcinoma [1,10].

Interferon alpha in combination with ribavirin is currently used as a standard therapy for chronic hepatitis C (CHC). Unfortunately, only about 40% of patients infected with HCV (1a or 1b, which belongs to the most common genotypes in Europe and in the USA) develop sustained response (SR) to the treatment. The remainder does not respond or display a transient response (TR) only [8]. Interestingly, no specific correlation between clinical parameters describing patients (e.g. age, sex, aminotransferases level, etc.) and the therapeutic outcome has been found. However, there is evidence to suggest that there

*Corresponding author. Email: szymon.wasik@cs.put.poznan.pl

ISSN 1748-670X print/ISSN 1748-6718 online

DOI: 10.1080/17486700903170712
http://www.informaworld.com

exists some correlation between patients' response to therapy and composition of the HCV quasispecies [2,9].

Earlier we observed in [4,5] that a highly diversified viral population can be linked to a positive therapy outcome in children (SR). On the other hand, a homogeneous population forecasts a treatment failure (TR or no response (NR)).

Briefly, those findings came from an analysis of the quasispecies' structure, based on observations of a fragment of E1/E2 protein coding sequence, before subjecting the patients to interferon–ribavirin therapy. Three parameters were used to describe an HCV population:

(1) quasispecies complexity (the number of different viral variants);
(2) phylogenetic trees; and
(3) genetic diversity represented by mean Hamming distance [4,5].

Among them, the last two parameters were shown to best reflect the differences between HCV populations isolated from patients with various treatment outcomes. Our results indicated that the therapy was successful when the tested region exhibited substantial polymorphism at $T0$ reflected in a high Hamming distance (Figure 1(a) – green bars) and a widely branched tree (Figure 1(b) – green trees). On the other hand, heterogeneity decreased in patients whose treatment was ineffective, reaching the lowest level in non-responders (Figure 1(a) – blue and red bars; (b) – blue and red trees).

In this paper, we propose a method of HCV population analysis useful for the prediction of CHC treatment outcomes. At present, all patients with CHC who meet certain standard inclusion criteria are subjected to the interferon–ribavirin treatment. Since the therapy is effective for about 40% of them only, it would be advantageous to modify and restrict the criteria, in order to improve the treatment efficiency, i.e. to limit it to patients in whom the therapy benefits outbalance the risks. This appears to be of special importance in view of the fact that the administration of currently used medicines if often connected with significant side effects, including flulike, gastrointestinal and psychiatric symptoms [7]. We intend to use the above depicted results to propose a mathematical approach that could support a process of patients' qualification for the treatment.

The direct measurements of HCV genetic diversity using the Hamming distance and/or phylogenetic trees' display are time consuming and costly. This is the main hindrance for their use as therapy predictors. The contribution of this paper is in observing a correlation

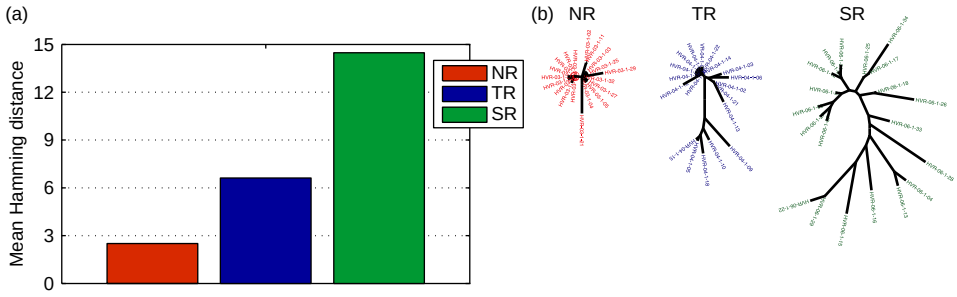

Figure 1. Selected parameters characterizing HCV populations present in patients with diverse responses to interferon–ribavirin therapy. (a) Genetic diversity reflected by mean Hamming distance. The colours of the bars correspond to the three types of the patients' response to treatment: red, NR; blue, TR; green, SR. (b) Phylogenetic trees. The trees' branches are represented by black solid lines, while the names of the analysed sequences are coloured in accordance with the patients' response to the treatment: red, NR; blue, TR; green, SR.

between HCV genetic diversity and some routinely determinable parameters. The parameters are viral RNA accumulation and alanine aminotransferase (ALT) level[7]. We will show that there is a virus RNA threshold level that can separate the patients who will most likely respond positively to the treatment, from the remaining ones. We contend that measuring the ALT levels cannot be used as a success predictor.

The reminder of this paper is organized as follows. In Section 2, we describe and analyse available case-study data on HCV infected patients. We explain in Section 2.5 how the state transitions can be inferred from the case study. The state transition analysis is carried on in Section 3. Section 4 summarizes our findings.

2. RNA levels model

2.1 Current treatment scheme

The standard therapy (including a follow-up) for CHC, caused by HCV that belongs to the most common genotypes, lasts 72 weeks. The level of HCV RNA accumulation is determined for each patient just before the administration of interferon and ribavirin ($T0$) and then after 24, 48 and 72 weeks ($T24$, $T48$ and $T72$, respectively). Patients with a detectable amount of HCV RNA in $T24$ are excluded from further therapy. The others are treated till $T48$. At $T72$, patients' response to the treatment is assessed and qualified as: (i) sustained (with an undetectable amount of HCV RNA at $T24$, $T48$ and $T72$), (ii) TR (with an undetectable amount of HCV RNA at $T24$, but measurable again at $T48$ and $T72$, or undetectable amount of HCV RNA at $T24$ and $T48$, but measurable again at $T72$), or (iii) NR (with a detectable amount of HCV RNA at $T24$, $T48$ and $T72$).

The numerical data used as a basis for the analysis presented in this paper have been collected from patients treated according to the described scheme (Appendix A, Tables A1 and A2).

2.2 Genetic diversity proxies

For display reasons, we code patients on the basis of the mean Hamming distance as follows (Figure 1(a)):

(1) *red* with low genetic diversity – mean Hamming distance below 4.5;
(2) *blue* with medium genetic diversity – mean Hamming distance between 4.5 and 6.61; and
(3) *green* with high genetic diversity – mean Hamming distance above 6.61.

Unfortunately, obtaining data necessary for construction of a phylogenetic tree is a long and expensive process, so only 15 cases were studied. On the other hand, the amount of virus RNA in blood is relatively easy to be measured. We have the data about RNA levels for 109 patients, including the 15 patients, for whom the phylogenetic trees are known. We used this data set to examine a relationship between the levels of virus RNA and the genetic diversity. We contend that if there is a correlation between these values, then we can use the RNA levels in patients to proxy the virus population genetic diversity and forecast a response to the treatment.

So, we examine the correlation between patients' genetic diversity measured by the mean Hamming distance and the RNA levels. We use data collected on patients in Table A1 (Appendix A, p. 13). The analysis is presented as a linear regression in Figure 2.

The R^2 coefficient for this regression is 0.39. This means that about 40% of variability in the mean Hamming distance can be explained by the changes in the RNA level.

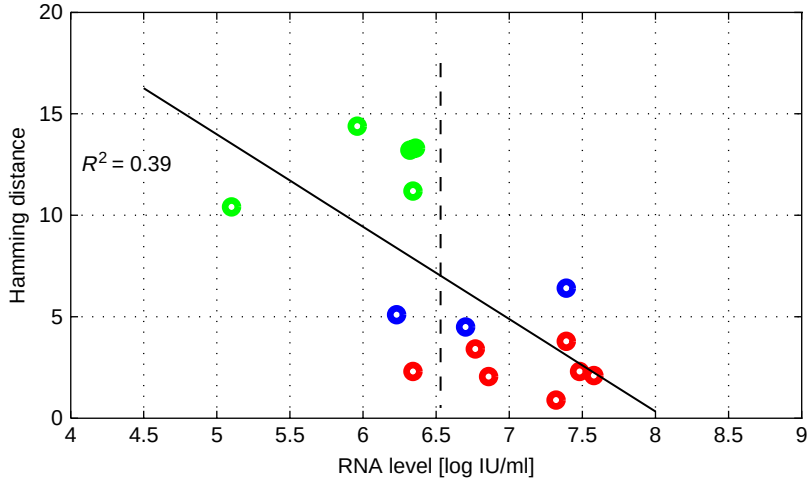

Figure 2. Linear regression between virus RNA level and Hamming distance of sequences at $T0$. Colours represent a group, to which patients belong, see items 2-2 above. The vertical line represents a division of states which will be explained later in Section 2.4. Data presented in this chart are taken from Table A1.

The value $R^2 = 0.39$ can be considered as rather large. However, the available sample size is small, which raises the question of significance of the obtained results. To decide upon it, we have applied the F-test for the model and the t-tests for the coefficients. The tests have confirmed significance of the model. The computed value of $F = 9.72$ is greater than $F_{\text{crit}} = 4.67$; also $(t = -3.11) < (-t_{\text{crit}} = -2.16)$ at significance level 0.05. Bearing in mind that the critical values of F_{crit} and t_{crit} allow for the small number of observations we remain satisfied that the results presented in this paper are statistically significant. In the end, we are confident that we can use patients' RNA levels as a proxy for HCV genetic diversity and hence forecast patient curability on the basis of the former.

Nevertheless, we are aware of the small sample, which is the base for our results and admit the possibility that they could change should the sample increase. The results that follow in this paper are justified by the above tests. We endeavour to replicate our study for a larger set of observations, once they become available. This would strengthen the practicality of our results.

As an alternative for the RNA levels we have also examined HCV genetic diversity dependence on the ALT levels. The data about ALT are also presented in Table A1 and the linear regression chart is presented in Figure 3. It can be noticed that the ALT level is extremely large for one point (the rightmost point in Figure 3). Accordingly, we analysed two sets of data – with and without this point. For these data, the R^2 coefficient is equal to 0.11 or 0.19, respectively. This is, in both cases, much smaller than R^2 for the RNA levels. This suggests that the ALT level is not a good choice for a proxy for HCV genetic diversity and will be excluded from further analysis.

2.3 Data distribution

A histogram of RNA level data divided into eight groups in $T0$ is presented in Figure 4. Each group has a width equal to $0.62(\log \text{IU}/\text{ml})$. A histogram of RNA level data divided

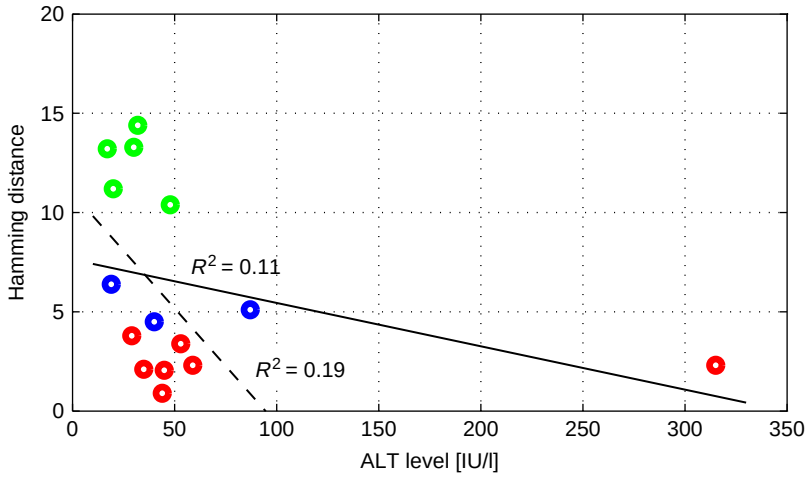

Figure 3. Linear regression between ALT level and Hamming distance of sequences at $T0$. The solid line represents regression between all data points, the dashed line between all points except the rightmost one. Data presented in this chart are listed in Table A1.

into 10 groups in $T24$, $T48$ and $T72$ is presented in Figure 5. In $T24$ and later patients with no RNA particles appear so the number of groups had to be increased. Each group has a width equal to $1.0(\log \mathrm{IU}/\mathrm{ml})$. The source data are presented in Table A2 (Appendix A, p. 14).

Should the group of healthy patients be removed (i.e. the group with RNA level equal to 0) the distributions presented in Figures 4 and 5 could be approximated by normal distributions[8].

We have measured mean and standard deviation of the virus RNA distribution at $T0$ as $m = 5.84$ and $\sigma = 0.88$, respectively. We will use these parameters to arrange patients into groups whose virus evolution can be considered similar.

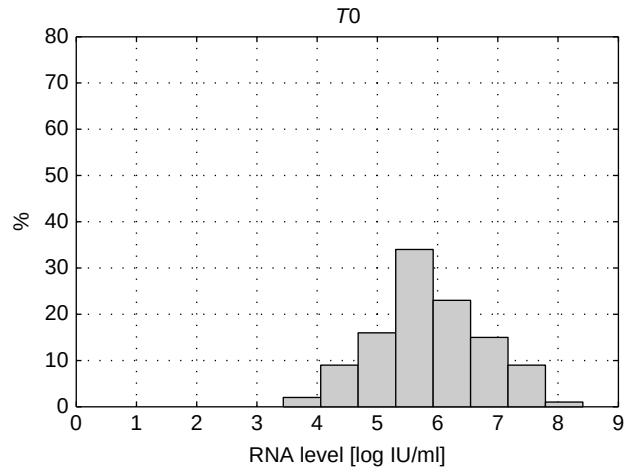

Figure 4. Distribution of RNA level data in $T0$.

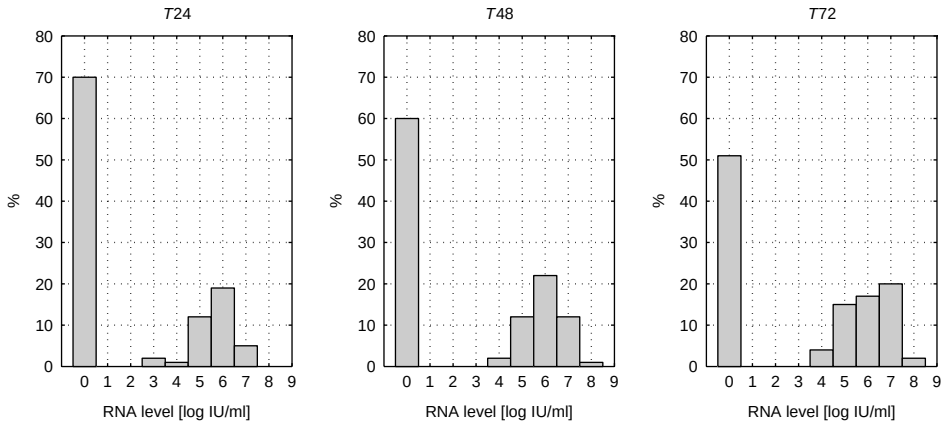

Figure 5. Distribution of RNA level data after 24, 48 and 72 weeks.

We notice that for some patients in Table A2, we have only information that the viral RNA was detected without an exact numerical value of its level. We have used the above distributions (Figures 4 and 5) to generate the 'missing' values. In particular, we have built a random number generator that produced numbers whose distribution was exactly the same as observed in the patients with full records. In effect, the resulting data spreads are the same as in the full record patients. In this fashion, we have increased the data set and utilized information about all patients that participated in the experiment whose levels of the viral RNA was significant.

2.4 Transition states

An analysis of Figures 4 and 5 suggests that a patient will belong to one of the following three groups over the course of 72 weeks:

- Group N or *no RNA* – where the RNA level is below or equal to $N_{\max} = 2.43(\log \mathrm{IU/ml})$. This is the level of RNA that cannot be detected with currently used methods. Patients with RNA levels in this range are considered uninfected.
- Group M or *medium RNA level* – where the RNA level is between $2.43(\log \mathrm{IU/ml})$ and $M_{\max} = 6.53(\log \mathrm{IU/ml})$. The latter value separates the *green* group of patients from the 'first' patient of the *blue* group (Figure 2) and is close to $\sigma + m$ reported for the distribution at $T0$. We propose this value as the upper limit of this group.
- Group H or *high RNA level* – where the RNA level is above $6.53(\log \mathrm{IU/ml})$, that is the rest of the patients.

The number of patients in each group at the beginning of each week is presented in Figure 6.

From this chart, we conclude that after 24 weeks of treatment about 60% of patients appeared to be uninfected (group N). Together with an increase of the number of patients in this group there is a decrease of the number of patients in the group with the high RNA level (group H). We observe that during the following weeks the number of people without the virus starts to decrease and the number of people with a high level of virus starts to increase which signifies the recurrence of the infection.

Analytically, we can say that the virus RNA levels 'evolve' between the above groups.

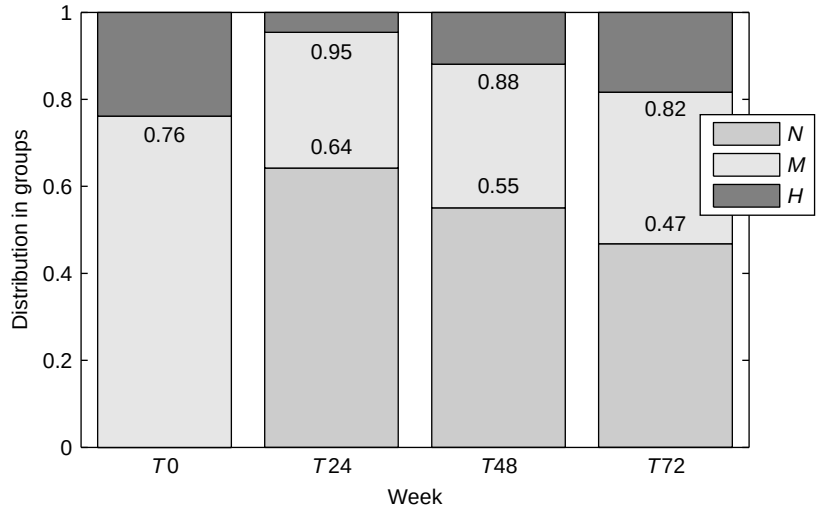

Figure 6. Distribution of patients between groups in different weeks when all patients are treated.

2.5 Transition matrices

The chart presented in Figure 6 provides a rather limited possibility for an analysis. However, it shows how the number of patients belonging to each group varies between the weeks. We contend that the above transitions are typical and postulate that they are governed by transition probabilities of patients between the groups.

We define a transition probability matrix as follows:

$$\mathbf{T}_{i,j} = \begin{bmatrix} p_{N,N}^{(i,j)} & p_{N,M}^{(i,j)} & p_{N,H}^{(i,j)} \\ p_{M,N}^{(i,j)} & p_{M,M}^{(i,j)} & p_{M,H}^{(i,j)} \\ p_{H,N}^{(i,j)} & p_{H,M}^{(i,j)} & p_{H,H}^{(i,j)} \end{bmatrix}. \tag{1}$$

Here, i and j are indices of weeks between which a transition occurs ($i, j \in \{T0, T24, T48, T72\}$) and $p_{g,h}^{(i,j)}$ is a transition probability between two groups $g, h \in \{N, M, H\}$ and weeks i and j.

We have calculated three matrices $\mathbf{T}_{T0,T24}$, $\mathbf{T}_{T24,T48}$ and $\mathbf{T}_{T48,T72}$ that represent transitions $T0 \rightarrow T24$, $T24 \rightarrow T48$ and $T48 \rightarrow T72$. We have also calculated two additional matrices for the transitions between the beginning and the end of treatment ($T0 \rightarrow T72$) and between the end of the first phase and the end of treatment ($T24 \rightarrow T72$).

We can also define a vector containing the number of patients at the beginning of week i as:

$$P_i = [P_{i,N}\, P_{i,M}\, P_{i,H}], \tag{2}$$

where $P_{i,g}$ denotes the number of patients in group g at week i. Using Equations (1) and (2), we can calculate the number of patients in week j as:

$$P_j = P_i \cdot \mathbf{T}_{i,j}. \tag{3}$$

In our case study, we have

$$P_0 = [0 \quad 83 \quad 26]. \tag{4}$$

The transition matrices calculated using the data from Table A2 are presented as Equations (5)–(9). We notice that the first row coefficients in matrix $\mathbf{T}_{0,24}$ are irrelevant for our study because there are no patients in group N at the beginning of the treatment.

$$\mathbf{T}_{0,24} = \begin{bmatrix} N/A & N/A & N/A \\ 0.711 & 0.277 & 0.012 \\ 0.423 & 0.423 & 0.154 \end{bmatrix}, \tag{5}$$

$$\mathbf{T}_{24,48} = \begin{bmatrix} 0.843 & 0.143 & 0.014 \\ 0.029 & 0.765 & 0.206 \\ 0.0 & 0.0 & 1.0 \end{bmatrix}, \tag{6}$$

$$\mathbf{T}_{48,72} = \begin{bmatrix} 0.85 & 0.1 & 0.05 \\ 0.0 & 0.722 & 0.278 \\ 0.0 & 0.462 & 0.538 \end{bmatrix}, \tag{7}$$

$$\mathbf{T}_{0,72} = \begin{bmatrix} N/A & N/A & N/A \\ 0.566 & 0.337 & 0.096 \\ 0.154 & 0.385 & 0.461 \end{bmatrix}, \tag{8}$$

$$\mathbf{T}_{24,72} = \begin{bmatrix} 0.714 & 0.214 & 0.071 \\ 0.029 & 0.647 & 0.324 \\ 0.0 & 0.2 & 0.8 \end{bmatrix}. \tag{9}$$

We will use the above matrices to estimate patient numbers in groups N, M, H when the initial patient numbers (and distribution) are different from (4).

3. Therapeutic efficiency

3.1 A measure

We are aiming to assess how effective the current therapy is and suggest a method that could be used to improve its efficiency. Owing to its significant side effects of the present therapy it would be beneficial for the patients' well-being, if only those whose treatment has a high probability of developing an SR were subjected to the therapy. This would improve the treatment efficiency because non-curable patients would be spared suffering.

We define the therapeutic efficiency ratio as follows

$$\varepsilon = \frac{\text{total cured}}{\text{total treated}}, \quad \varepsilon \in [0, 1]. \tag{10}$$

We will assume that 'total treated' will be a rather large number so that degenerated cases (0 and 1) are avoided. We understand 'total cured' as the number of patients who have undetectable virus RNA level in $T72$.

If we allow 'total treated' to be a political decision variable that depends on the initial RNA level $\mu, \mu \in [3.0, 8.5]$[9] then therapeutic efficiency will depend on this variable as follows

$$\varepsilon(\mu) = \frac{\text{total cured}}{\text{total treated}\,(\mu)}, \quad \varepsilon \in [0, 1]. \tag{11}$$

It is easy to read from Table A2 (also, see the last bar in Figure 6) that

$$\varepsilon(8.5) = \frac{51}{109} = 47\%.$$

This value characterizes the current therapeutic policy, which consists of treating any patient that meets inclusion criteria irrespective of their virus RNA contents.

We notice another interesting number that may characterize therapeutic efficiency that is the composition of 'total cured'. From the data in Table A2, we can calculate that 92% of 'total cured' have come from the group whose RNA level is below 6.53(log IU/ml) (Section 2.4). This suggests that treating patients with the RNA levels below that number should increase the value of ε[10]. In the next section, we will assess ε for situations when submitting a patient to the therapy will depend on their initial RNA level.

3.2 Sensitivity of therapeutic efficiency to the initial RNA level

Assume that the virus RNA level in a patient in period t depends on its contents in period $t-1$ only. Under the current treatment method, no patient is treated between weeks 48 and 72. Hence, given the above assumption, we postulate that the transition probabilities in matrix $\mathbf{T}_{48,72}$ (Equation (7)) could be used to describe the change in the virus levels in the untreated patients at any time.

According to the current treatment method all patients are treated between weeks 0 and 24. We propose the following simulation exercise. Suppose that only patients with the RNA levels below a threshold level are subjected to the therapy. If we replace the last row in $\mathbf{T}_{0,24}$ (i.e. the row that describes the virus transitions in untreated patients) with the last row of matrix $\mathbf{T}_{48,72}$ we will receive matrix $\mathbf{T}'_{0,24}$. This matrix will represent probabilities of transitions when only patients with low virus RNA levels are treated between weeks 0 and 24. Using this matrix and Equation (3), we will calculate the number of patients in each group in week 72. The result is presented in Equation (12) which produces the expected number of patients (it was rounded to the integer number).

$$P'_{72} = P_0 \cdot \mathbf{T}'_{0,24} \cdot \mathbf{T}_{24,48} \cdot \mathbf{T}_{48,72} = \begin{bmatrix} 43 & 41 & 25 \end{bmatrix}. \tag{12}$$

The patient distributions among the groups in this case are shown in Figure 7. We can see that the comparison of 'total cured' (the fourth bar's lowest block) to 'total treated' (the first bar's lowest block) results in $\varepsilon = 0.52$, which is higher than before.

Presumably, the fewer the patients we treat, the higher the therapeutic efficiency. However, fewer patients treated means that some of them would remain sick but could have been cured under higher $M_{\max}$, which is the RNA level that separates the treated from untreated patients (Section 2.4). We will now vary $M_{\max}$ and examine sensitivity of ε to those variations.

So, for a few values of $M_{\max}$ we use the above algorithm, i.e. we

(1) calculate transition matrices $\mathbf{T}$ using a new value of $M_{\max}$ (Section 2.5);
(2) replace the last row of $\mathbf{T}_{0,24}$ with the last row of $\mathbf{T}_{48,72}$;

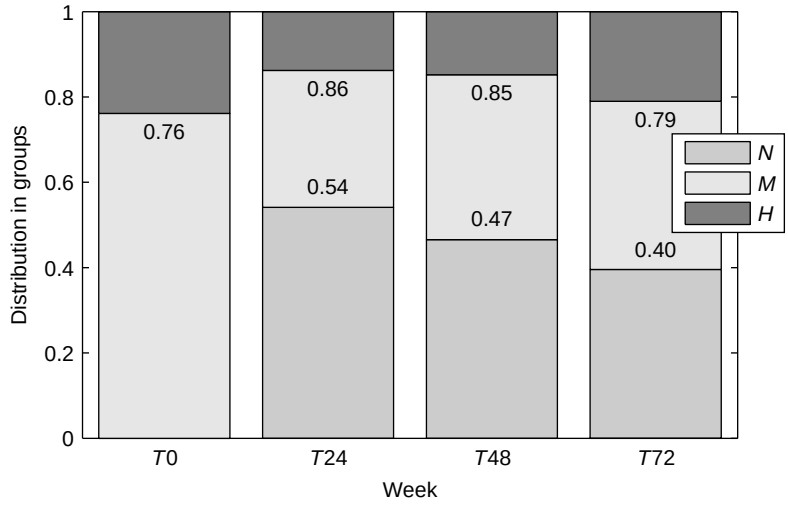

Figure 7. Distribution of patients between groups in different weeks if only group M is treated.

(3) calculate the number of patients P_{72} at $T72$ using Equation (3); and
(4) calculate the value of $\varepsilon(\mu)$, here $\mu = M_{\max}$.

The above algorithm uses transition matrices to estimate the value of ε. We compared it to the algorithm that does not use matrices $\mathbf{T}$ and calculates the value of ε using the raw data. This algorithm is following:

(1) Select only these patients who have the level of RNA at $T0$ lower or equal to $M_{\max}$. Let P_{all} denote the number of these patients.
(2) Calculate the number of patients who had SR in the selected group of patients. Let P_{SR} denote the number of these patients.
(3) Calculate the value of ε using the following formula:

$$\varepsilon = \frac{P_{\text{SR}}}{P_{\text{all}}}. \tag{13}$$

The latter algorithm does not assume that the RNA level depends only on its content in the previous period but its disadvantage is that it can model only those cases that are explicitly provided in data. For example, it cannot model the case in which some patient in group H is healed although we know from matrix $\mathbf{T}_{24,72}$ that it is possible (it applies to 3% of patients in group H).

The results are presented in Figure 8. We can observe that the algorithm which does not use transition matrices has a significant local minimum at $4.25(\log \text{IU}/\text{ml})$. This minimum occurs because for virus RNA level lower than $4.5(\log \text{IU}/\text{ml})$ there is a very limited number of patients and most of them did not have SR. This is the instability that often occurs in real data, especially when the number of cases is small. This minimal ε value is calculated based on only six patients' data and when the number of patients becomes more statistically significant when the increase of $M_{\max}$ ε raises to the higher value. This problem does not occur for the algorithm that uses transition matrices. To estimate the value of ε it uses results of statistical analysis instead of non-processed data and that is why it is more noise-resistant. On the other hand, this shape of the curve

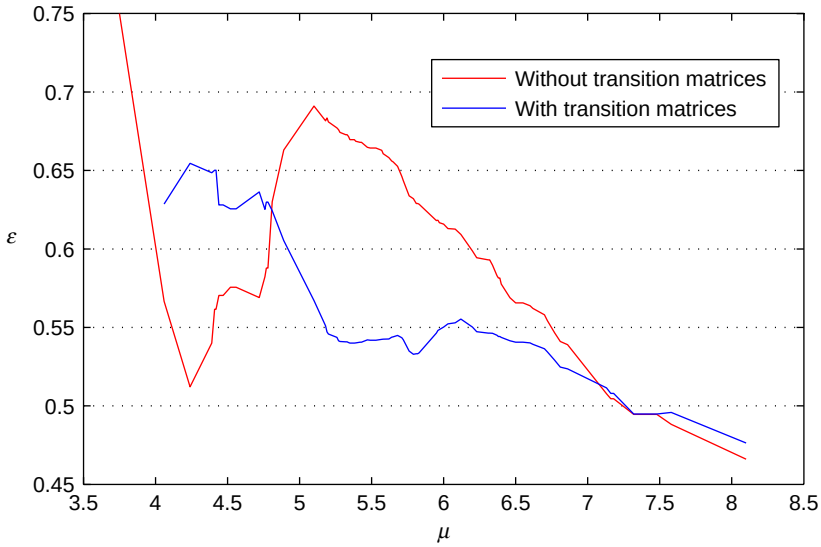

Figure 8. The value of ε for different values of μ. The central moving average algorithm was used to smooth data. (For each point the average value of points from the window with a width equal to 0.75(log IU/ml) was calculated.)

reflects also the complexity of this biological system where the outcome of virus infection and therapy depends on many factors, not only on the viral RNA level.

We can also observe that increasing $M_{\max}$ decreases the value of ε which is consistent with other reports (for example [6]). It can be noticed that after exceeding the RNA level of $M_{\max} = 5.25(\log \text{IU/ml})$ the efficiency of the therapy modelled with transition matrices decreases significantly and the efficiency of the therapy modelled without transition matrices stays at the constant level. Using these observations, it can be suggested that the optimal $M_{\max}$ is somewhere at $5.25(\log \text{IU/ml})$.

4. Concluding remarks

It was proposed in [4,5] that phylogenetic trees and the mean Hamming distance of HCV populations can be useful in predicting a CHC treatment outcome. In this article we employed a linear regression analysis to test the dependency between the genetic variability of the virus and its accumulation reflected in the RNA level in blood. R^2 coefficient for the regression, confirmed by the F- and t-tests, turned out to be reasonably high (0.39). This enabled us to use the level of viral RNA accumulation as a proxy for genetic variability. We then associated the types of responses to treatment with the three patient groups (N, M, H), separated by the different virus RNA levels. Next, we constructed matrices describing transitions between the groups. Finally, we analysed the therapeutic efficiency using unprocessed data and results of the algorithm that uses transition matrices. Our results indicated that RNA accumulation below $5.25(\log \text{IU/ml})$ can be regarded as a threshold separating patients with high probability to develop an SR from those whose response was null or temporary. Gathering more data from a larger group of patients could contribute to improving (i.e. 'tuning' the threshold) the current criteria of patient qualification for CHC therapy.

Notwithstanding the encouraging result, we endeavour to confirm our findings using a larger dataset than currently available.

Acknowledgements

This study was partially supported by the Polish Government through Grants N301 019 31/0483 and N519 314635 from the Ministry of Science and Higher Education. In addition, PJ was supported by the scholarship provided by the President of the Polish Academy of Sciences.

Notes

1. Email: paulinaj@ibch.poznan.pl
2. Email: j.krawczyk@vuw.ac.nz
3. Email: pawel.kedziora@cs.put.poznan.pl
4. Email: piotr.formanowicz@cs.put.poznan.pl
5. Email: marekf@ibch.poznan.pl
6. Email: jacek.blazewicz@cs.put.poznan.pl
7. ALT is an enzyme found predominantly in liver and a highly sensitive indicator of hepatocellular abnormalities. When liver damage occurs, serum ALT level rises.
8. It was verified with the Lillifors test at the 0.05 significance level that the data distribution, with the patients with undetectable viral RNA level removed, is normal.
9. The threshold limits have been chosen to include the RNA extreme values registered in the Poznan case study (Table A2).
10. We are conscious that the percentage of the cured patients whose initial RNA level was above 6.53(log IU/ml) is not zero; however, it is relatively low. Consequently, it might be socially acceptable to treat only those patients whose RNA levels are sufficiently low to predict SR and at the same time they are not too low, to not exclude too many patients from treatment.

References

[1] A. Alberti, L. Chemello, and L. Benvegnu, *Natural history of hepatitis C*, J. Hepatol. 31(Supplement 1) (1999), pp. 17–24.
[2] 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*, Proc. Natl Acad. Sci. USA 99 (2002), pp. 3081–3086.
[3] M. Figlerowicz, M. Alejska, A. Kurzynska-Kokorniak, and M. Figlerowicz, *Genetic variability: The key problem in the prevention and therapy of RNA-based virus infections*, Med. Res. Rev. 23 (2003), pp. 488–518.
[4] M. Figlerowicz, P. Jackowiak, M. Alejska, N. Malinowska, A. Kowala-Piaskowska, P. Kedziora, P. Formanowicz, J. Blazewicz, and M. Figlerowicz, *Two types of viral quasispecies identified in children suffering from chronic hepatitis C*, J. Hepatol. 50 (2009), p. S127.
[5] P. Kedziora, M. Figlerowicz, P. Formanowicz, M. Alejska, P. Jackowiak, N. Malinowska, A. Fratczak, J. Blazewicz, and M. Figlerowicz, *Computational methods in diagnostics of chronic hepatitis C*, Bull. Pol. Acad. Sci. Tech. 53 (2005), pp. 273–281.
[6] J.Y.N. Lau, G.L. Davis, J. Kniffen, K.P. Qian, M.S. Urdea, M. Chan, C.S. abd Mizokami, P.D. Neuwald, and J.C. Wilber, *Significance of serum hepatitis C virus RNA levels in chronic hepatitis C*, Lancet 341 (1993), pp. 1501–1504.
[7] J.G. McHutchinson and J.H. Hoofnagle, *Therapy of Chronic Hepatitis C*, Academic Press, San Diego, CA, 2000, pp. 203–239.
[8] 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) (1999), pp. 57–65.
[9] J.M. Pawlotsky, G. Germanidis, A.U. Neumann, M. Pellerin, P.O. Frainais, and D. Dhumeaux, *Interferon resistance of hepatitis C virus genotype 1b: Relationship to nonstructural 5A gene quasispecies mutations*, J. Virol. 72 (1998), pp. 2795–2805.
[10] L.B. Seeff, *Natural history of chronic hepatitis C*, Hepatology 36 (2002), pp. S35–S46.
[11] P. Simmonds, *Genetic diversity and evolution of hepatitis C virus – 15 years on*, J. Gen. Virol. 85 (2004), pp. 3173–3188.
[12] World Health Organization, *Hepatitis C – global prevalence (update)*, Wkly Epidemiol. Rec. 74 (1999), pp. 425–427.

Appendix A: The Poznan case study

Table A1. ALT and HCV RNA level at the beginning of treatment (at $T0$) and mean Hamming distance.

Patient	MHD	RNA (log IU/ml)	ALT (IU/l)
P1-1	0.9	7.32	44
P1-2	2.1	7.58	35
P1-3	2.3	6.34	59
P1-5	3.8	7.39	29
P2-4	2.3	7.48	40
P2-5	3.4	6.77	53
P2-10	2.05	6.86	45
P1-4	6.4	7.39	19
P1-7	5.1	6.23	87
P2-8	4.5	6.7	40
P1-6	14.4	5.96	32
P1-8	13.3	6.36	30
P1-9	11.2	6.34	20
P1-10	10.4	5.1	48
P2-2	13.2	6.32	17

These data were used to calculate the correlation in Section 2.2. Patients presented in this table and data about Hamming distance are the subset of data presented in Table A2.

Table A2. HCV RNA level in 0, 24, 48 and 72 weeks after beginning of treatment in log IU/ml.

Patient	$T0$	$T24$	$T48$	$T72$	Response
6	3.75	0.00	0.00	0.00	SR
68	4.06	6.00	+	+	NR
38	4.24	0.00	0.00	0.00	SR
74	4.39	5.59	+	+	NR
47	4.41	0.00	0.00	4.29	TR
51	4.42	0.00	0.00	+	TR
22	4.44	0.00	0.00	0.00	SR
15	4.47	0.00	0.00	0.00	SR
45	4.52	3.04	0.00	0.00	SR
2	4.54	0.00	0.00	0.00	SR
43	4.56	0.00	0.00	0.00	SR
16	4.72	0.00	0.00	0.00	SR
41	4.76	0.00	0.00	0.00	SR
44	4.77	0.00	0.00	0.00	SR
70	4.78	4.54	+	+	NR
93	4.81	5.53	6.25	5.70	NR
23	4.89	0.00	0.00	0.00	SR
35	5.10	0.00	0.00	0.00	SR
P1-10	5.10	0.00	0.00	0.00	SR
5	5.18	0.00	0.00	0.00	SR
57	5.18	0.00	4.96	+	TR
46	5.19	0.00	0.00	0.00	SR
24	5.20	0.00	0.00	0.00	SR
8	5.26	0.00	0.00	0.00	SR
66	5.27	4.76	+	+	NR
7	5.28	0.00	0.00	0.00	SR
13	5.28	0.00	0.00	0.00	SR
3	5.32	0.00	0.00	0.00	SR

Table A2 – *continued*

Patient	$T0$	$T24$	$T48$	$T72$	Response
21	5.33	0.00	0.00	0.00	SR
28	5.33	0.00	0.00	0.00	SR
19	5.34	0.00	0.00	0.00	SR
10	5.35	0.00	0.00	0.00	SR
55	5.36	0.00	5.00	+	TR
87	5.37	5.51	+	+	NR
64	5.38	4.66	+	+	NR
17	5.39	0.00	0.00	0.00	SR
50	5.42	0.00	0.00	5.44	TR
1	5.43	0.00	0.00	0.00	SR
37	5.47	0.00	0.00	0.00	SR
48	5.47	0.00	0.00	5.39	TR
73	5.47	6.21	+	+	NR
86	5.47	5.02	+	+	NR
53	5.49	0.00	0.00	5.51	TR
26	5.50	0.00	0.00	0.00	SR
36	5.50	0.00	0.00	0.00	SR
72	5.53	4.68	+	+	NR
94	5.53	5.40	+	+	NR
76	5.57	5.71	+	+	NR
9	5.58	0.00	0.00	0.00	SR
18	5.62	0.00	0.00	0.00	SR
52	5.64	0.00	0.00	+	TR
29	5.65	0.00	0.00	0.00	SR
30	5.68	0.00	0.00	0.00	SR
32	5.71	0.00	0.00	0.00	SR
34	5.72	0.00	0.00	0.00	SR
60	5.72	0.00	+	+	TR
42	5.76	0.00	0.00	0.00	SR
40	5.79	0.00	0.00	0.00	SR
79	5.81	2.99	+	+	NR
67	5.82	4.65	5.83	+	NR
4	5.83	0.00	0.00	0.00	SR
78	5.95	5.14	+	+	NR
P1-6	5.96	0.00	0.00	0.00	SR
25	5.97	0.00	0.00	0.00	SR
39	5.97	0.00	0.00	0.00	SR
49	6.00	0.00	0.00	+	TR
27	6.03	0.00	0.00	0.00	SR
63	6.08	0.00	5.03	5.93	TR
82	6.12	5.80	+	+	NR
12	6.20	0.00	0.00	0.00	SR
P1-7	6.23	0.00	6.33	6.27	TR
65	6.31	4.42	+	+	NR
77	6.31	5.54	+	+	NR
P2-2	6.32	0.00	0.00	0.00	SR
P1-3	6.34	5.88	6.29	6.38	NR
P1-9	6.34	0.00	0.00	0.00	SR
P1-8	6.37	0.00	0.00	0.00	SR
61	6.38	0.00	+	+	TR
75	6.39	6.42	+	+	NR
20	6.40	0.00	0.00	0.00	SR
85	6.46	5.35	+	+	NR
81	6.50	5.70	+	+	NR

Table A2 – *continued*

Patient	$T0$	$T24$	$T48$	$T72$	Response
89	6.52	6.59	+	+	NR
62	6.53	0.00	5.39	+	TR
84	6.55	6.20	+	+	NR
31	6.60	0.00	0.00	0.00	SR
71	6.60	4.91	+	+	NR
14	6.62	0.00	0.00	0.00	SR
P2-8	6.70	0.00	0.00	6.90	TR
58	6.72	0.00	+	+	TR
88	6.72	5.63	+	+	NR
59	6.77	0.00	5.76	+	TR
P2-5	6.77	6.89	6.92	7.29	NR
33	6.81	0.00	0.00	0.00	SR
80	6.81	6.18	+	+	NR
83	6.81	5.63	+	+	NR
P2-10	6.86	6.69	6.79	6.81	NR
90	7.13	5.23	5.39	6.19	NR
11	7.16	0.00	0.00	0.00	SR
56	7.18	0.00	4.92	5.92	TR
P1-1	7.32	6.21	7.16	7.23	NR
91	7.33	5.69	5.77	7.15	NR
54	7.34	0.00	0.00	6.53	TR
69	7.35	4.55	+	+	NR
P1-5	7.39	7.45	7.41	7.48	NR
P1-4	7.39	0.00	6.26	7.06	TR
P2-4	7.48	7.42	7.52	7.91	NR
P1-2	7.58	6.49	7.29	7.36	NR
92	8.10	6.32	4.47	5.08	NR

'+' means that the virus was detected but exact data are unknown. Patients whose number starts with 'P' are those for whom data about mean Hamming distance are available. Data about mean Hamming distance are presented in Table A1. The meaning of abbreviations in response column is: SR, sustained response; TR, transient response; NR, no response.