ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

Recognition of HIV-1 subtypes and antiretroviral drug resistance using weightless neural networks

Caio R. Souza[1], Flavio F. Nobre[1], Priscila V.M. Lima[2], Robson M. Silva[2], Rodrigo M. Brindeiro[3], Felipe M. G. França[1]

1- COPPE, Universidade Federal do Rio de Janeiro - Brazil
2- DEMAT/ICE, Universidade Federal Rural do Rio de Janeiro – Brazil
3- Laboratory of Molecular Virology, Universidade Federal do Rio de Janeiro - Brazil

Abstract. This work presents an application of an improved version of the WiSARD weightless neural network in the recognition of different mutation types of HIV-1 and in the determination of antiretroviral drugs resistence. The data set used consists of 1205 gene sequence of the HIV-1 protease of subtypes B, C and F from patients under treatment failure. Experiments performed with the *bleaching* technique over the WiSARD model under different data representation strategies have shown promising results, both in terms of accuracy and standard deviation.

1 Introduction

AIDS is still one of the major public health problem. In Brazil, several measures have been taken by the government to detect infection as early as possible and ensure access to anti-HIV drugs [1]. Despite these efforts, AIDS treatment faces the emergence of antiretroviral drugs resistance.

Various studies have been done to predict which drug will result in a therapeutic failure. The main anti-HIV drugs inhibit important viral enzymes and so many of these studies used genomic analysis of the protease and the reverse transcriptase. The present work introduces an application of an improved version of the WiSARD weightless neural network model [3], in the categorization of subtypes of HIV-1 and in the determination of antiretroviral drugs resistance.

2 Materials and Methods

2.1 Database

The data set were provided by Molecular Virology Laboratory at the Federal University of Rio de Janeiro (UFRJ / Brazil). This database consists of 1205 gene sequence of the HIV-1 protease from patients with HIV-1. The samples are distributed in 6 classes according to their subtype and drug resistance: B resistance (62%), B naïve (8%), C resistance (12%), C naïve (4%), F resistance (11,5%), F naive (1,5). This is distribution is similar to the Brazilian reality [4].

2.2 WiSARD

The WiSARD (Wilkie, Stonham and Aleksander's Recognition Device) is a weightless neural network based on the set of answers from RAM's memories. It was

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

the earliest patented and commercially produced artificial neural network, and also the more representative weightless neural network (WNN) model [5].

The WiSARD input is divided into *M* equal tuple and each is related to a RAM, which must have 2^N bits, where *N* is the size of the tuples. These tuples are used to address the memories positions and access the stored knowledge. Therefore, we can see that the entry should be made up of *M* x *N* bits [6].

Initially all memories positions are set to "0". The learning phase consists on changing to "1" at the RAM neurons positions addressed by the binary training patterns. Thus, each neuron is extremely efficient at recognizing a pattern previously seen, but does not recognize similar patterns. Each tuple represents only part of the entry and the pattern is divided into *M* parts, the network stores each class in a separate discriminator, with *M* sets RAM's. These memories will be responsible for the network generalization and the noise tolerance [5].

In the recognition phase, a new pattern is presented to all discriminators. Each RAM reads the memory address pointed by the corresponding tuple and returns either 0 or 1. Each discriminator counts the number of 1´s to obtain a total score. The discriminator with the highest score is the response to that new pattern [5].

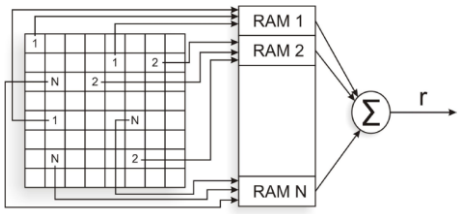

Figure 1. WiSARD discriminator [3].

The system also provides a *confidence* level *C* of the recognition process, which can be expressed by I = (*Rmax* - *R2MAX*) / *Rmax*; where *Rmax* is the highest score found and *R2MAX* is the second highest one.

The training and recognition process can be executed in constant time, because they are affected during training by the size of the input, and for recognition by the number of categories. Furthermore, the degree of generalization can be easily controlled, since it depends of the relationship between the size and the number of RAMs memories in the discriminator. The bigger each RAM memory is, the less the network has the ability to generalize [5]. Although this method has already shown good results in several studies, it has a limitation when two or more categories get tied scores. In this case one of them is chosen randomly.

2.3 Bleaching

In the *bleaching* technique, instead of recording Boolean values (0's and 1's) in a memory location, the number of times such memory location is written is stored. In the recognition phase, a bleaching value is used to verify if that memory location scores or not. Without this technique, the memories with value equal "1" punctuates; with bleaching, only location values that exceeds a current bleaching threshold will score [3].

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

The bleaching threshold starts with zero. If two or more categories are even, the bleaching value is increased by one, and the memories scores will be recalculated. This procedure is repeated until only one discriminator is elected, or until they all stop scoring. Only in this second case a choice must be made randomly.

3 Experiments

The WiSARD neural network stores information in RAM's addresses, and therefore, works primarily with binary data. Thus, it is essential to codify the HIV protease into a binary sequence. Silva codified the protease using the amino acids hydrophobicity scale and assigned to each of them a 20-bit binary number [6]. In the present work, other forms of encoding using the hydrophobicity scale, molecular weight and a combination of these two information were considered. These three scales were converted using the binary system, Gray coding, spreading its values from 1 to 2^{20}-1 in order to get bit Gray Code and also spreading the values of 1 to (2^{20}-1)*2/3, making the smallest and largest values of the scale composed solely of zeros and ones, respectively. Listed below are the encodings created:

- BIN20: 20-bit vector containing only one bit in 1.
- HIDROBIN: Hydrophobicity in binary scaled from 1 to 2^{20}-1.
- HIDROGRAY: Hydrophobicity in Gray code scaled from 1 to 2^{20}-1.
- HIDROGRAY23: Hydrophobicity in Gray code scaled from 1 to (2^{20}-1)*(2/3).
- MMBIN: Molecular mass in binary scaled from 1 to 2^{20}-1
- MMGRAY: Molecular mass in Gray code scale from 1 to 2^{20}-1.
- MMGRAY23: Molecular mass in Gray code scale from 1 to (2^{20}-1)*(2/3).
- HIDROMMBIN: Molecular mass followed by the value of hydrophobicity, both in binary scale from 1 to 2^{10}-1.
- HIDROMMGRAY: Molecular mass followed by the value of hydrophobicity both in Gray code and scale from 1 to 2^{10}-1.
- HIDROMMGRAY23: Molecular mass followed by the value of hydrophobicity, both in Gray code and scale from 1 to (2^{10}-1)*(2/3).

With these 10 encoding alternatives and using the WiSARD with 16-bits address memories we did three experiments: (1) Recognition of subtype and detecting the presence of any drug resistance; (2) Recognition of virus subtype and (3) Recognition of any drug resistance for individuals with subtype B HIV. This subtype has been chosen because it is prevalent not only in Brazil but also in Europe and North America, therefore, the most widely studied [7]. A total of 30 experiments were done, all using cross-validation leave-one-out (LOOCV).

4 Results

4.1 Accuracy Analysis

In the first experiment, the database was divided into six categories, and network was attempted to differentiate the subtypes and naive forms of the virus reaching accuracy from 65.5% to 69.1%, with an average of 67.5%. However, the resistance mutations do not have the same pattern for different subtypes. Therefore, in the second

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0. Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

experiment we first separate the samples according to the subtype of the virus, and then differentiate between resistant and naive forms.

This analysis can be done in two different ways: by grouping the samples by subtype before training or by joining the solutions according to the subtype, ignoring if the information is resistant or not. The accuracy varied from 89.5% to 94% when the samples are not grouped before training and when it was grouped the accuracy was between 86.5% and 93.5 %. It is important to note that in none of these cases the standard deviation was greater than 5%.

The good performance obtained, in some cases near 94%, showed that this approach could be useful in a future work, where the distinction of the drugs for which the patient will present resistance, would be made.

	ALL 6 CLASSES		WITHOUT GROUPING		GROUPING		SUBTYPE B	
ENCODING	AVG	STDV	AVG	STDV	AVG	STDV	AVG	STDV
BIN20	65,5%	3,4%	91,2%	2,5%	89,7%	1,6%	73,7%	4,4%
HIDROBIN	68,6%	5,1%	92,9%	2,4%	93,1%	2,6%	75,4%	5,3%
HIDROGRAY	68,5%	5,7%	92,3%	2,7%	90,7%	3,1%	76,4%	6,1%
HIDROGRAY23	68,7%	7,0%	93,3%	2,9%	93,5%	2,2%	76,1%	5,4%
MMBIN	69,1%	5,9%	94,0%	1,3%	93,5%	1,9%	74,2%	5,7%
MMGRAY	66,4%	8,0%	90,2%	5,0%	87,8%	4,7%	73,9%	3,6%
MMGRAY23	68,8%	7,1%	93,5%	2,2%	92,8%	2,4%	74,4%	5,9%
HIDROMMBIN	66,2%	5,3%	89,5%	2,6%	86,5%	4,0%	74,6%	7,0%
HIDROMMGRAY	66,4%	6,9%	91,1%	3,2%	88,4%	3,7%	73,8%	4,9%
HIDROMMGRAY23	67,1%	5,8%	90,8%	4,4%	88,6%	4,2%	75,9%	6,1%
AVERAGE	67,5%	6,0%	91,9%	2,9%	90,5%	3,1%	74,8%	5,4%

Table 1. Categorization results.

In the last experiment, we select only the subtype B samples and search for the presence or absence of resistance. The accuracy was between 73.7% and 76.4%, with and standard deviation ranging between 3.6% and 7.0%. These results were better than those obtained by differentiating the six categories. Considering these results with the excellent results obtained in the last experiments, it is our conclusion that categorization can be done in stages. Table 1 shows all the experiments results.

4.4 Bleaching Analysis

The main objective of this technique is to assist the WiSARD, giving a deterministic criterion for the cases of a tie. In this paper this technique was also used to increase the confidence of results. Thus, whenever there was a tie, the chosen categories were fixed and the increase of the difference between the both of them was observed.

Although the use of bleaching to improve confidence in the results increases processing time for the recognition phase, this showed very good results. In this study this technique ensures a deterministic recognition for all patters. Without the bleaching a random choice would have happened at almost 15% of all responses. For the experiment with samples of subtype B and using the encoding BIN20, less than half of the examples presented for recognition were chosen deterministically, indicating the poor performance obtained with this encoding.

Considering all experiments, this technique ensures an improved correct classification rate of 8.8%, and in the experiment with only subtype B and BIN20 encoding the increase was 34%. Furthermore, this technique allowed an increase in confidence in 83.6% of all experiments, showing that it not only improves accuracy, but it is also important to assess the credibility of the responses from the network.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

5 Discussion

Several machine learning algorithms have been used to study the changes in HIV-1 and to predict the virus resistance to a particular drug. Silva used Kernel Discriminant combined with genetic algorithm to identify new mutation positions [6]. After selecting the positions that characterize the resistance to a particular drug, prediction accuracies of drug resistance of 88% for Saquinavir, 81.25% for Nelfinavir and 84.93% for the Lopinavir were obtained.

In [8], neural networks were used to predict resistance to Indinavir and Saquinavir, obtaining an accuracy of 85%. Wang and Larder also use neural networks, and achieved an average accuracy of 88% for Lopinavir when using the protease sequence [9]. Beerenwinkel *et al.* worked with the protease and reverse transcriptase of the virus using decision trees and support vector machine (SVM). The results for SVM had accuracies from 54% to 85% for transcriptase inhibitors and 78% to 89% for protease inhibitors. For the decision trees, the accuracy varies between 58% and 97% for transcriptase inhibitors and 82% to 90% for the protease [10][11].

Wang *et al.* compared three methods: support vector machines, neural networks and "Random Forest", a method that consists in the use of several decision trees in parallel, and used both the protease and reverse transcriptase. The average absolute difference of change in viral load and the correlation between the predicted outcome and the change in viral load as metrics were used. The average absolute difference for these methods have respectively 0.600, 0.543 and 0.607 log10 copies/ml. Correlation metrics results are 62%, 68% and 70% respectively [12].

Although this work also aimed the prediction of the resistance of HIV-1, the adopted approach was different from other studies in the literature. Here, since we had samples from different HIV subtypes, initial target was the classification of HIV subtypes. Additionaly, the database used here did not have information regarding to which drug the patient showed resistant, and therefore, such categorization cannot be made, which led to the distinction of data in just two groups: naive and resistant. While this could impair the performance of the classifier, the results presented in Section 4, showed that this approach achieve in some cases much better results than those found through other methods. Furthermore, this technique provides high accuracy with very low standard deviation.

6 Conclusion

In this work we implement a WiSARD, a weightless neural network, that capable to recognize different mutation types of HIV-1. In these experiments it was also possible to evaluate how the technique of bleaching should be used to improve the accuracy.

The data used here were encoded in 20-bit binary sequences based on the chemical characteristics of the amino acids. In the experiments we used alone the molecular mass and hydrophobicity and also a combination of these two parameters. In addition, the matrix Bin20 was used for encoding, resulting in 10 different ways to represent the amino acids.

Experiments were performed with different focuses. Initially, tried to categorize the six groups present in the database, then tried to detect the subtype of the sample

only, without differentiating between resistant or naive and, finally sought to sort only between resistant and naive in subtype B samples.

For the experiments focusing on the recognition of the six groups, the highest degree of accuracy obtained was 69% with a standard deviation of 5% only. In the case of subtype categorization, the accuracy was above 90% and reached the best success rate of 93.9% with a standard deviation of only 1%. In the last experiment, which involved only subtype B samples, a success rate of 76.4% and standard deviation of 6% were obtained.

This paper also evaluated the bleaching technique, which works when a tie occurs. It was showed that without using this technique, only 72.6% of the experiments would reach a deterministic result. This means that, on average, approximately one in four recognitions would be random. Moreover, it was observed that the use of bleaching improved the degree of accuracy of the result by 18.5 percentage points and increased confidence in 84% of the experiments. Consequently, not only it can be concluded that this technique is necessary, but also that the technique here described can be of great help after fixing the groups obtained as soon as the tie occurs. Bleaching should be used so as to reveal the confidence that can be achieved. This method also showed a better ratio between accuracy and confidence.

References

[1] Ministério da Saúde, Brazil, DST/AIDS (http://www.aids.gov.br), accessed in 05/2011.

[2] Baxter, D., *et al.*, 2000. A randomized study of antiretroviral management based on plasma genotypic antiretroviral failing therapy, AIDS, v.14 (9), pp.83-92.

[3] Grieco, B. P. A.; Lima, P. M. V.; De Gregorio, M.; França, F. M. G., 2010. Producing pattern examples from "Mental" images. Neurocomputing (Amsterdam), V. 73, P. 1057-1064.

[4] Brindeiro, R., *et al.*, 2003. Brazilian network for HIV drug resistance surveillance (HIV-BResNet): a survey of chronically infected individuals, AIDS, v.17, pp.1063-1069.

[5] I. Aleksander, W. Thomas, and P. Bowden, "WISARD, a radical new step forward in image recognition", Sensor Rev., 4(3), 120-124, 1984.

[6] Silva, R. M., 2009. Genetic algorithm and kernel discriminant applied to identification of mutation in the hiv-1 resistance to the antiretroviral protease inhibitors. PhD Thesis, COPPE, Universidade Federal do Rio De Janeiro (UFRJ). Rio de janeiro.

[7] Spira, S., *et al.*, 2003. Impact of lade diversity on HIV-1 virulence, antiretroviral drug sensitivity and drug resistance. J. Antimicrob. Chemother. 51:229-240.

[8] Draghici, S., Potter, R., 2003. Predicting HIV drug resistance with neural networks, Bioinformatics, v.19, pp.98-107.

[9] Wang, D., Larder, B., 2003. Enhanced prediction of lopinavir resistance from genotype by use of artificial neural networks. J. Infect. Dis. 188, 653–660

[10] Beerenwinkel, N. *et al.*, 2002. Diversity and complexity of HIV-1 drug resistance: a bioinformatics approach to predicting phenotype from genotype. Proc. Natl. Acad. Sci. U. S. A. 99, 8271–8276

[11] Beerenwinkel, N. *et al.*, 2003. Geno2pheno: Estimating phenotypic drug resistance from HIV-1 genotypes. Nucleic Acids Res. 31, 3850–3855

[12] Wang, D. *et al.*, 2009. A comparison of three computational modeling methods for the prediction of virological response to combination HIV therapy. Artificial Intelligence in Medicine 47: 63—74.