Predicting HIV-1 Protease and Reverse Transcriptase Drug Resistance Using Fuzzy Cognitive Maps

Isel Grau, Gonzalo Nápoles, and María M. García

Universidad Central “Marta Abreu” de Las Villas, Santa Clara, Cuba
{igrau,gnapoles,mmgarcia}@uclv.edu.cu

Abstract. Several antiviral drugs have been approved for treating HIV infected patients. These drugs inhibit the function of proteins which are essential in the virus life cycle, thus preventing the virus reproduction. However, due to its high mutation rate the HIV is capable to develop resistance to administered therapy. For this reason, it is important to study the resistance mechanisms of the HIV proteins in order to make a better use of existing drugs and design new ones. In the last ten years, numerous statistical and machine learning approaches were applied for predicting drug resistance from protein genome information. In this paper we first review the most relevant techniques reported for addressing this problem. Afterward, we describe a Fuzzy Cognitive Map based modeling which allows representing the causal interactions among the protein positions and their influence on the resistance. Finally, an extended comparison experimentation is carried out, which reveals that this model is competitive with well-known approaches and notably outperforms other techniques from literature.

Keywords: HIV proteins, Drug resistance, Prediction, Fuzzy Cognitive Maps.

1 Introduction

In the last two decades several antiretroviral (ARV) drugs have been designed for treating Human Immunodeficiency Virus (HIV). The main goal of ARV therapies is to inhibit the function of essential proteins for the virus life cycle such as *protease*, *reverse transcriptase* or *integrase*. For instance, the *reverse transcriptase* protein catalyzes the reverse transcription process, which transforms RNA into DNA and incorporates the resulting DNA into the host cell. As a result, the infected cell produces viral particles which are matured by *protease* protein, cleaving precursor proteins and therefore completing the virus replication process. Evidently, inhibiting these processes could help on preventing the virus reproduction. Nevertheless, due to its high mutation rate, the HIV is capable to develop resistance to administered drugs, evading the immune system and causing the therapy failure.

Consequently, understanding resistance mechanisms is critical for designing more effective treatment strategies or even developing new ARV drugs [1]. In general, the resistance testing of an observed mutation can be performed by two main approaches: the genotypic and the phenotypic tests. The genotypic testing is based on identifying drug-resistance mutations (and the combination of them) which have been associated to decreased susceptibility of a target drug; while phenotypic testing measures the

viral replication in presence of different ARV concentrations. In clinical practice, genotype assays are more frequently used than phenotype ones since they are less expensive in time and effort; however, they only provide indirect evidence of resistance. On the other hand, phenotypic testing is more useful for determining the susceptibility of new approved ARV drugs, where patterns on resistance have not yet been well described. Phenotype assays are clinically suitable for viruses with complex mutational patterns where genotype interpretation becomes really difficult [1].

The paired results of such tests, performed for several protein mutations, constitute a valuable historical data in order to understand the HIV behavior. Based on this knowledge, in the last ten years several statistical and machine learning methods have been proposed for predicting the phenotype resistance to a target drug from the genotype information (known as virtual phenotype), that is, the resistance degree of a mutation (target attribute) given its amino acid sequence (predictive features). In some cases these data-driven approaches lead to parsimonious models, but in general they are harder to interpret [2].

In a recent attempt to use more interpretable techniques, in previous works [3, 4] the authors proposed a model based on Fuzzy Cognitive Maps (FCM) [5] with the goal of discovering knowledge on the causal patterns among the sequence positions and the phenotype resistance. Although this research was mainly focused on the causality interpretation of learned maps, we observed that the prediction accuracies notably outperformed several well-known classifiers. Inspired on this result, we propose an extended comparison experiment for measuring the accuracy of this model using historical data from several *protease* and *reverse transcriptase* inhibitors. Before, we first review the most relevant techniques reported in the last ten years for addressing this classification problem. In addition, the FCM model is described in Section 3, but now it is investigated from the prediction point of view.

2 Computational Approaches for Drug Resistance Analysis

Since not only the number of approved ARV drugs is increasing, but also the resistant mutations to these treatments, the use of intelligent systems have progressively become more important for understanding the resistance phenomena in HIV proteins. Actually, in the last years the use of computational methods for prediction or interpretation of HIV drug resistance has been growing. In general terms, such models constitute a very useful tool for guiding physicians in designing complex individual therapies and drug experts in the development of new ARV [1].

At beginning several rule-based systems were introduced, using the knowledge of physicians and also data about mutations previously associated with resistance from clinical trials. This is the case of Rega [6], ANRS [7] and VGI systems [8]. In fact, at the same time the Stanford HIV Drug Resistance Database project enabled the public access to an algorithm known as HIVdb [9]. This approach uses, in addition to rules, a drug penalty score for inferring five levels of resistance. Moreover, this project has a platform (HIValg) for comparing the output of several drug interpretation algorithms; which was used in [10] for determining those relevant mutation patterns responsible of observed discordance among the investigated rule-based approaches.

Subsequently, more accurate computational techniques such as support vector regression were employed in Geno2Pheno [11], which is another web service for drug resistance prediction. Also, different types of neural networks were explored [12-14] where bidirectional recurrent neural networks [15] reported competitive performance in terms of accuracy. Regression models also had been studied; for instance, a standard stepwise linear regression [16] outperformed other genotypic interpretation algorithms publicly available so far, including decision trees, support vector machine and four rule-based algorithms (HIVdb, VGI, ANRS and Rega). Later, in [17] Rabinowitz et al. introduced two regression techniques using convex optimization and perform a comparison against the most relevant approaches at the moment.

On the other hand, in the same year Rhee et al. [18] published the results of five previously proposed predictors including decision trees, linear regression, linear discriminant analysis, neural networks, and support vector regression, using high quality filtered knowledge bases. These historical data is publicly available for experimental comparisons of new algorithms. More recently, in reference [19] was described a linear regression called itemset boosting that works particularly well for predicting the resistance of nucleotide *reverse transcriptase* inhibitors. As well, in [20] least-angle regression was performed to identify *protease* mutations associated with reduced susceptibility to at least one *protease* inhibitor, and least-squares regression was employed in order to quantify the contribution of *protease* mutations to reduced susceptibility. Finally, in [21] the author implements a procedure based on n-grams to generate sequence attributes; where results are complementary to other sequence-based approaches, reporting competitive features in performance.

3 A Model Based on Fuzzy Cognitive Maps Theory

In this section we briefly describe the FCM model proposed in [3, 4] which was conceived for studying the causal influence of the *protease* protein positions on the resistance when a mutation occurs. Next, we explain the generalization of this model for any other HIV protein as a tool for describing the drug resistance activity.

As a first step each protein position is represented as a map concept, while another node for denoting the resistance degree to a specific drug is also defined. Afterwards, causal connections among all input concepts are created, representing the interaction (causal influence) among all protein positions. Also, connections between each map concept and the final resistance concept are established. This topology is supported by the fact that there exist relations among not necessarily adjacent positions of the sequence due to the three-dimensional structure of the protein; where a change in the amino acid of a specific position (i.e. mutation) could be relevant for the drug resistance [4]. For better understanding of this scheme, following figure 1 illustrates the general conception of the FCM that results from this stage.

Then, in order to determine the causality among positions and the resistance variable, a learning process based on Swarm Intelligence is carried out. This learning algorithm uses historical data publicly available for finding a causal matrix that minimizes the difference between the reported resistance and the value of the resistance concept (map inference), for all mutations reported [3, 4].

Fig. 1. Topology for describing a HIV protein through the FCM theory. Concepts denote positions of the sequence, and links stand for causal connection among amino acids.

It is fair to mention that, with the purpose of reducing the model dimensionality, only sequence positions previously associated with resistance are considered. These positions are selected from both numerical and biological perspective, as a result of available research in this field. As a result, predicting the resistance target from descriptors nodes means to solve the related classification/regression problem as follow. First, the activation value of each concept is taken as the contact energy [22] of corresponding amino acid normalized in the range $[0,1]$. As a second step, the map inference process is triggered and the value of the resistance concept is examined. For example, for a regression perspective this value will denote the normalized degree of resistance; whereas for a classification perspective, a drug-specific cut-off for determining the resistance class (0-susceptible and 1-resistant) should be used.

It is also remarkable that typical FCM can't solve classification problems [23]; however from empirical simulations we noticed that this model was able to outperform other well-known classifiers. In fact, in the following section we carry out an extensive set of experiments for fully exploring the prediction ability of this model against other classifiers for both *protease* and *reverse transcriptase* proteins.

4 Simulations and Results

In the present section we study the inference ability of the FCM model for solving the bioinformatics problem enunciated before. To do that, we use historical data associated with 7 *protease* inhibitors and 11 *reverse transcriptase* inhibitors taken from [9]. The *protease* inhibitors used in this work are: Amprenavir (APV), Atazanavir (ATV), Indinavir (IDV), Lopinavir (LPV), Nelfinavir (NFV), Ritonavir (RTV) and Saquinavir (SQV). While two kinds of *reverse transcriptase* inhibitors are used: nucleoside/nucleotide and nonnucleoside inhibitors. The nucleoside/nucleotide ones are: Lamivudine (3TC), Abacavir (ABC), Zidovudine (AZT), Stavudine (D4T), Zalcitabine (DDC), Didanosine (DDI), Emtricitabine (FTC) and Tenofovir (TDF); whereas nonnucleoside are: Delavirdine (DLV), Efavirenz (EFV) and Nevirapine (NVP). Then, we evaluate the FCM model against some approaches from literature, for solving the related classification problem having two and three classes.

4.1 Solving the Sequence Classification Problem for Two Classes

The idea here is to compare the accuracy of the FCM model against other classifiers for solving the binary classification problem (susceptible and resistant). In all cases we use the cut-off values reported in [9] as thresholds for determining the class of each instance. In addition we use the following parameters settings for the learning algorithm [3,4]: 80 particles, five variable neighborhoods, 200 generations, and the allowed number of generations without progress is set to 40. For comparison we used next methods: Support Vector Machine with linear kernel (SVML), polynomial of degree 1 (SVM1), degree 2 (SVM2), degree 3 (SVM3), and radial basis (SVMR); in addition we used a Multilayer Perceptron (MLP) and a Bidirectional Recurrent Neural Network (BRNN), all taken from [15]. To conclude, we consider an Artificial Neural Network (ANN) from [14] and a novel ensemble classifier from [24] called Multi-Expert by Hard Instances (MEHI) which has reported promising accuracies.

Table 1 shows the accuracy from a 10-fold cross-validation process using data of six *protease* inhibitors, corresponding to the complete unfiltered datasets of the Phenosense assay. From these results the following conclusions are drawn: for ATV, SQV, LPV and IDV the investigated model notably outperforms other algorithms, while for RTV and NFV it reports quite competitive results (where MEHI computes the best accuracies). However, it is remarkable that, for ATV the FCM model is able to outperform MEHI in 12 percent points, whereas for drugs RTV and NFV the FCM model and the MEHI algorithm only differs at most in two percent points.

In literature, there are few reports concerning to algorithms for solving this binary classification problem for *reverse transcriptase* inhibitors using datasets with complete sequence. Despite this inconvenient, in [25] Grau et al. proposed a Recurrent Neural Network (RNN) that uses a modified backpropagation through time algorithm for dealing with instances of variable length, allowing handling the complete sequence. Following Table 2 summarizes the comparison accuracies between the FCM model and the RNN approach. Here the FCM model largely outperforms the RNN at most inhibitors, which confirm the suitability of FCMs to deal with this problem.

Table 1. Classification accuracy obtained for *protease* inhibitors (two classes)

Drug	SVML	SVM1	SVM2	SVM3	SVMR	ANN	MLP	BRNN	MEHI	FCM
ATV	0.78	0.71	0.68	0.72	0.70	-	0.80	0.81	0.84	0.96
SQV	0.87	0.80	0.69	0.85	0.82	0.91	0.85	0.91	0.85	0.95
LPV	0.88	0.85	0.85	0.88	0.85	0.92	0.92	0.94	0.93	0.98
RTV	0.91	0.84	0.79	0.92	0.86	0.96	0.90	0.94	0.99	0.97
IDV	0.91	0.86	0.83	0.92	0.88	0.95	0.86	0.92	0.97	0.99
NFV	0.84	0.75	0.70	0.84	0.80	0.95	0.86	0.93	0.96	0.95

Table 2. Classification accuracy obtained for *reverse transcriptase* inhibitors (two classes)

Model	3TC	ABC	AZT	D4T	DDC	DDI	DLV	EFV	FTC	NVP
RNN	0.70	0.61	0.56	0.89	0.78	0.81	0.58	0.56	0.86	0.65
FCM	0.96	0.94	0.98	0.93	0.80	0.79	0.92	0.95	0.98	0.95

4.2 Solving the Sequence Classification Problem for Three Classes

In this subsection we extend the experimentation by comparing the FCM approach against other classifiers for solving the related classification problem, now using three classes (susceptible, intermediate and resistant). Besides, the same parameter setting of the FCM learning algorithm used in the above section is adopted. At this moment two kind of datasets are studied; the first ones corresponds to the complete unfiltered dataset, while the second are high quality filtered datasets both available in [9]. For comparison we used following approaches: Random Forest classifier using relative frequency approach (RF1) and Random Forest classifier using a counts method (RF2), Reduced-Error Pruned Tree with relative frequency procedure (REPT1) and Reduced-Error Pruned Tree with counts approach (REPT2), all taken from [21]. Also, we consider: Decision Trees (DT), Neural Networks (NN), Least-Squares Regression (LSR), Support Vector Regression (SVR) and also Least-Angle Regression (LARS), all taken from [18]. Table 3 and 4 show the computed accuracy from a 5-fold cross-validation process for *protease* and *reverse transcriptase* inhibitors.

Table 3. Classification accuracy obtained for *protease* inhibitors (three classes)

Drug	High quality filtered datasets						Complete unfiltered datasets				
	SVR	LSR	LARS	DT	NN	FCM	REPT1	RF1	REPT2	RF2	FCM
APV	0.82	0.81	0.81	0.77	0.74	0.61	-	-	-	-	-
ATV	0.69	0.68	0.76	0.71	0.64	0.70	0.74	0.75	0.76	0.76	0.78
IDV	0.77	0.78	0.77	0.75	0.73	0.78	0.78	0.80	0.75	0.80	0.72
LPV	0.80	0.79	0.83	0.77	0.76	0.85	0.80	0.82	0.80	0.81	0.78
NFV	0.79	0.79	0.80	0.76	0.73	0.70	0.80	0.80	0.79	0.82	0.75
RTV	0.86	0.86	0.88	0.84	0.81	0.80	0.87	0.86	0.87	0.84	0.79
SQV	0.81	0.81	0.82	0.75	0.76	0.73	0.80	0.79	0.80	0.80	0.74

Table 4. Classification accuracy obtained for *reverse transcriptase* inhibitors (three classes)

Drug	High quality filtered datasets						Complete unfiltered datasets				
	SVR	LSR	LARS	DT	NN	FCM	REPT1	RF1	REPT2	RF2	FCM
3TC	0.84	0.83	0.88	0.90	0.90	0.86	0.89	0.87	0.87	0.90	0.83
ABC	0.65	0.63	0.77	0.69	0.66	0.68	0.68	0.68	0.66	0.67	0.62
AZT	0.7	0.64	0.76	0.70	0.71	0.83	0.75	0.75	0.73	0.70	0.86
D4T	0.68	0.66	0.78	0.75	0.72	0.78	0.74	0.79	0.76	0.78	0.72
DDC	-	-	-	-	-	-	0.80	0.75	0.80	0.76	0.74
DDI	0.67	0.61	0.75	0.74	0.71	0.85	0.69	0.73	0.69	0.71	0.74
DLV	0.78	0.73	0.84	0.84	0.78	0.67	0.76	0.70	0.76	0.71	0.78
EFV	0.82	0.78	0.87	0.84	0.77	0.63	0.78	0.74	0.76	0.73	0.73
FTC	-	-	-	-	-	-	0.96	0.83	0.94	0.89	0.98
NVP	0.78	0.74	0.87	0.91	0.81	0.91	0.84	0.79	0.82	0.77	0.88
TDF	0.69	0.46	0.70	0.68	0.73	0.66	0.75	0.75	0.68	0.74	0.70

From Table 3 and 4 it is observed that the overall performance of all classifiers is reduced regarding to the classification problem using two classes. For example, in the *protease* filtered dataset, the studied algorithm reports better accuracies for IDV and LPV, whereas for the other drugs it computes competitive results. On the other hand, for the *reverse transcriptase* filtered dataset the FCM model performs better for AZT, DDI and NVP; reporting reasonable accuracies for remaining inhibitors, except for non-nucleoside drugs DLV and EFV where the percent of correct classified instances notably decreases. Though, using the complete *reverse transcriptase* datasets, FCM is able to outperform other classifiers for AZT, DDI, DLV, FTC, and also NVP; showing competitive results for remaining inhibitors. The reduction in the performance could be due to inconsistency or imbalanced knowledge bases. Future work will be focused on improving these results by introducing a new classification strategy for FCM which takes into account these issues and uses an alternative topology and stability criteria in the inference process. In addition this study will be extended for solving the related regression problem.

5 Conclusions

Understanding the complex behavior of HIV includes the prediction of resistance features to existing drugs. However, predicting phenotype from genotype information involves a challenging sequence classification problem, which has been addressed in literature by using well-known classifiers, but the prediction accuracies are still unsatisfactory. Recently was proposed a model based on Fuzzy Cognitive Maps for analyzing causal patterns among positions on *protease* sequences. While this study was oriented to the knowledge discovering, we noticed that reported prediction accuracies were promising. In this paper we explored this feature, and next aspects are concluded: (i) The FCM model using two prediction classes (susceptible and resistant) significantly outperformed other evaluated classifiers for both *protease* and *reverse transcriptase* datasets, (ii) The FCM model using three prediction classes (susceptible, intermediate and resistant) decreased its performance, although it is competitive in most cases with respect to other classifiers, therefore complementing reported approaches from literature.

References

1. Tang, M.W., Shafer, R.W.: HIV-1 Antiretroviral Resistance Scientific Principles and Clinical Applications. *Drugs* 72(9), 1–25 (2012)
2. Beerenwinkel, N., et al.: Computational methods for the design of effective therapies against drug resistant HIV strains. *Bioinformatics* 21, 3943–3950 (2005)
3. Grau, I., Nápoles, G., León, M., Grau, R.: Fuzzy Cognitive Maps for Modelling, Predicting and Interpreting HIV Drug Resistance. In: Pavón, J., Duque-Méndez, N.D., Fuentes-Fernández, R. (eds.) *IBERAMIA 2012*. LNCS, vol. 7637, pp. 31–40. Springer, Heidelberg (2012)
4. Nápoles, G., Grau, I., León, M., Grau, R.: Modelling, aggregation and simulation of a dynamic biological system through Fuzzy Cognitive Maps. In: Batyrshin, I., Mendoza, M.G. (eds.) *MICAI 2012, Part II*. LNCS, vol. 7630, pp. 188–199. Springer, Heidelberg (2013)
5. Kosko, B.: Fuzzy Cognitive Maps. *Int. Journal of Man-Machine Studies* 24, 65–75 (1986)

6. Laethem, K., et al.: A genotypic drug resistance interpretation algorithm that significantly predicts therapy response in HIV-1-infected patients. *Antiviral Therapy* 7, 123–129 (2002)
7. Rousseau, M.N., et al.: Patterns of resistance mutations to antiretroviral drugs in extensively treated HIV-1-infected patients with failure of highly active antiretroviral therapy. *Journal of Acquired Immune Deficiency Syndromes* 26, 36–43 (2001)
8. Reid, C., et al.: A dynamic rules-based interpretation system derived by an expert panel is predictive of virological failure. *Antiviral Therapy* 7, S91 (2002)
9. Rhee, S.Y., et al.: Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Research* 31, 298–303 (2003)
10. Ravela, J., et al.: HIV-1 Protease and Reverse Transcriptase Mutation Patterns Responsible for Discordances Between Genotypic Drug Resistance Interpretation Algorithms. *Journal of Acquired Immune Deficiency Syndromes* 33, 8–14 (2003)
11. Beerenwinkel, N., et al.: Geno2pheno: estimating phenotypic drug resistance from HIV-1 genotypes. *Nucleic Acids Research* 31, 3850–38505 (2003)
12. Draghici, S., Potter, R.B.: Predicting HIV drug resistance with neural networks. *Bioinformatics* 19, 98–107 (2003)
13. Woods, M., Carpenter, G.A.: Neural Network and Bioinformatic Methods for Predicting HIV-1 Protease Inhibitor Resistance. Technical Report 02215 (2007)
14. Pasomsub, E., et al.: The Application of Artificial Neural Networks for Phenotypic Drug Resistance Prediction: Evaluation and Comparison with Other Interpretation Systems. *Jpn. Journal of Infectious Diseases* 63, 87–94 (2010)
15. Bonet, I., García, M.M., Saeys, Y., Van de Peer, Y., Grau, R.: Predicting Human Immunodeficiency Virus (HIV) Drug Resistance Using Recurrent Neural Networks. In: Mira, J., Álvarez, J.R. (eds.) *IWINAC 2007*. LNCS, vol. 4527, pp. 234–243. Springer, Heidelberg (2007)
16. Wang, K., et al.: Simple linear model provides highly accurate genotypic predictions of HIV-1 drug resistance. *Antiviral Therapy* 9, 343–352 (2004)
17. Rabinowitz, M., et al.: Accurate prediction of HIV-1 drug response from the reverse transcriptase and protease amino acid sequences using sparse models created by convex optimization. *Bioinformatics* 22, 541–549 (2005)
18. Rhee, S.Y., et al.: Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *PNAS* 103, 17355–17360 (2006)
19. Saigo, H., Uno, T., Tsuda, K.: Mining complex genotypic features for predicting HIV-1 drug resistance. *Bioinformatics* 23, 2455–2462 (2007)
20. Rhee, S.Y., et al.: HIV-1 Protease Mutations and Protease Inhibitor Cross-Resistance. *Antimicrobial Agents and Chemotherapy* 54, 4253–4261 (2010)
21. Masso, M.: HIV-1 Prediction of Human Immunodeficiency Virus Type 1 Drug Resistance: Representation of Target Sequence Mutational Patterns via an n-Grams Approach. In: *IEEE International Conference on Bioinformatics and Biomedicine*, pp. 1–6 (2010)
22. Song, H.J., et al.: An Extension to Fuzzy Cognitive Maps for Classification and Prediction. *IEEE Transactions on Fuzzy Systems* 19, 116–135 (2011)
23. Miyazawa, S., Jernigan, R.L.: Contacts energies Self-Consistent Estimation of Inter-Residue Protein Contact Energies Based on an Equilibrium Mixture Approximation of Residues. *PROTEINS: Structure, Function, and Genetics* 34, 49–68 (1999)
24. Bonet, I., et al.: Multi-Classifer Based on Hard Instances- New Method for Prediction of Human Immunodeficiency Virus Drug Resistance. *Current Topics in Medicinal Chemistry* 13, 685–695 (2013)
25. Grau, I., Nápoles, G., Bonet, I., Garcia, M.M.: Backpropagation Through Time Algorithm for Training Recurrent Neural Networks using Variable Length Instances. *Computación y Sistemas* 17, 15–24 (2013)