Mutating HIV Protease Protein Using Ant Colony Optimization and Fuzzy Cognitive Maps: Drug Susceptibility Analysis

Isel Grau and Gonzalo Nápoles

Centro de Estudios de Informática, Universidad Central “Marta Abreu” de Las Villas, Cuba

{igrau, gnapoles}@uclv.edu.cu

Abstract. Understanding the dynamics of the resistance mechanisms in HIV proteins mutations is a key for optimizing the use of existing antiviral drugs and developing new ones. Several statistical and machine learning techniques have been proposed for predicting the resistance of a mutation to a certain drug using its genotype information. However, the knowledge publicly available for this kind of processing is majorly about resistant sequences, leading to highly imbalanced knowledge bases, which is a serious problem in classification tasks. In previous works, the authors proposed a methodology for modeling an HIV protein as a dynamic system through Fuzzy Cognitive Maps. The adjusted maps obtained not just allow discovering relevant knowledge in the causality among the protein positions and the resistant, but also achieved very competitive performance in terms classification accuracy. Based on these works, in this paper we propose an Ant Colony Optimization based method for generating possible susceptible mutations using the adjusted maps and biological heuristic knowledge. As a result, the mutations obtained allow drug experts to have more information of the behavior of the protease protein whenever a susceptible mutation takes place.

Keywords. HIV, drug resistance, mutations, fuzzy cognitive maps, modeling, ant colony optimization.

Mutación de la proteína proteasa del VIH utilizando optimización basada en colonia de hormigas y mapas cognitivos difusos: análisis de susceptibilidad a fármacos

Resumen. El conocimiento de los mecanismos de resistencia en las mutaciones de las proteínas del VIH es fundamental para optimizar el uso de los fármacos existentes, así como diseñar nuevos medicamentos. Varias técnicas de estadística y aprendizaje automatizado han sido propuestas en la literatura para intentar predecir la resistencia de una mutación a un fármaco determinado usando su información genotípica. Sin embargo el conocimiento disponible públicamente para este tipo de procesamientos está enfocado mayormente a las mutaciones resistentes, lo que provoca bases de conocimiento altamente desbalanceadas que constituyen un serio problema en las tareas de clasificación. En trabajos previos, los autores proponen una metodología para modelar una proteína del VIH como un sistema dinámico a través de Mapas Cognitivos Difusos. Los mapas ajustados obtenidos no solo permiten descubrir conocimiento en la causalidad entre las posiciones de la proteína y la resistencia, sino que alcanza un desempeño competitivo en términos de exactitud de la clasificación. Basado en estos trabajos, en este artículo proponemos un método basado en la técnica de Optimización de Colonias de Hormigas para generar nuevas mutaciones susceptibles utilizando los mapas ajustados y conocimiento biológico heurístico. Como resultado, las mutaciones obtenidas permitirían a los expertos en fármacos contar con mayor información sobre el comportamiento de la proteasa cuando aparece una mutación susceptible.

Palabras clave. VIH, resistencia a fármacos, mutaciones, mapas cognitivos difusos, modelación, optimización basada en colonia de hormigas.

1 Introduction

In the last few years, several antiretroviral drugs have been approved for treating the Human Immunodeficiency Virus (HIV). These drugs are designed for inhibiting the function of proteins that play an important role in the virus life cycle, such as protease, reverse transcriptase and integrase. However, due to its high mutation rate, this virus is capable to develop resistance to therapies designed by specialists. Therefore, the study of the resistance mechanisms in the proteins

ISSN 1405-5546
http://dx.doi.org/10.13053/CyS-18-1-2014-018

mutations is a key for optimizing the use of existing drugs and designing effective new ones [1].

Generally speaking, it is possible to determine the resistance of a mutation to a given drug by two different tests. The simplest one is the genotype testing, which consist on sequencing the patient and look for mutations previously associated with resistance. This method is relatively cheap but the interpretation could be too hard if multiple mutations take place. On the other hand, the phenotype test measures the quantity of drug concentration needed for inhibiting the protein function. This test is quite exact but is also costly in time and resources [2].

The information gathered from both experiments could be very useful in the study of the behavior of HIV proteins against different antivirals. In fact, in [3] is publicly available the paired results of these test. However, the historical data stored for this kind of processing is majorly about resistant sequences, leading to highly imbalanced knowledge bases. Imbalanced datasets are a very common problem in knowledge bases from real world problems. Frequently the minority class is usually the one that has the highest interest from the application point of view. In order to treat this issue, two major approaches have been proposed in literature: the data sampling which consist in modifying the dataset for obtaining a balanced distribution, and the algorithmic point of view which considers the imbalanced distribution in the learning process [4].

Despite imbalanced datasets, several machine learning and statistical techniques have used these historical data for training methods for virtual phenotyping, that is, to predict the phenotypic resistance using the genotypic information [1, 2, 5-9]. The numerous models proposed in literature offer a variety of tools for helping in designing therapies for patients without using the phenotype testing. Particularly, the authors in [10, 11] use the Fuzzy Cognitive Maps (FCM) theory for modeling the behavior of the HIV protease. In this proposal the causality patterns among all sequence positions and the resistance were learned using a Swarm Intelligence approach. As a result, the prediction accuracies obtained for five antiviral drugs were very promising and competitive, supporting the quality of the causal relations expressed in the adjusted map. In addition, the interpretation capabilities of the FCM allow the knowledge discovery of causality patterns of some punctual mutations and the resistance. As final contribution, these previous works, offer a simulation tool for studying the causality among all sequence positions when multiple mutations take place.

In this paper we extent these results by using the inference capabilities and the causal relations expressed in the obtained maps for generating protease sequences which report low resistance (susceptibility) to the studied drugs. To do so, an Ant Colony Optimization (ACO) approach is used to generate the susceptible sequences mutations, modeled as a discrete optimization problem. Also, biological knowledge about the frequency of each possible mutation in nature is used as heuristic knowledge. The generated mutations could expand the knowledge available about the resistance mechanisms and, to some extent, offer an alternative for treating the imbalanced distribution in the knowledge bases used for virtual phenotyping.

The rest of the paper is organized as follows. The next section makes an overview of previous works describing some theoretical aspects of FCM and explaining the protease modeling, learning process and their results. Section 3 proposes the ACO based method for generating susceptible mutation sequences. In Section 4 we discuss some experiments and their results. As a final point, conclusions and future work suggestions are given in Section 5.

2 Modeling Protease as a FCM

HIV protease protein can be seen as a dynamic system where all positions of the genomic sequence interact with each other to some degree, depending on the 3D structure of the protein. In addition, interactions taking place in the active site have strong influence in the drug ability to dock to the protein and thus inhibit its function. In fact, frequently, some punctual mutations in the sequence cause resistance to a given drug by preventing the docking of the antiviral to the active site of the protein.

The behavior of this complex biological system has been modeled and studied using different statistical and machine learning methods, mainly in order to predict the drug resistance from the protein genome sequence using historical data. Particularly, in previous works the authors proposed a methodology for modeling and simulate the behavior of resistance mechanism in HIV proteins through FCM.

2.1 Fuzzy Cognitive Maps

FCM are a soft computing technique which combines fuzzy logic and artificial neural network theories. They were proposed by Kosko in [12] as an extension to cognitive maps. Graphically, they are composed by nodes (concepts) representing descriptive variables of the system, and links (relations) expressing causality between two concepts.

From the fuzzy logic point of view, concepts are characterized by a fuzzy value in the range [0, 1] denoting the activation degree of the represented variable in the system. The causal links are weighted arcs representing the cause-effect relations between two concepts, and they can be described by a fuzzy value in the range [-1, +1]. The sign of the causal relations specify the direction of the change, for example, if there is a positive causality between two concepts then an increase in the source concept leads to an increase in the target variable, or if a negative causality exists between the concepts, then an augment in the source concept causes a reduction in the activation value of the target variable. Lastly, if the value is zero, there is no causal relation between the concepts. These fuzzy weights are often linguistically defined by experts in the application domain or could be automatically learned from historical data of the modeled problem.

On the other hand, from the connectionist point of view, FCM are a type of recurrent artificial neural network since they involve feedback in their connections. This aspect allows to express the dynamic of the system by describing the effects of a change in a variable on the other variables, which in turn can affect the node initiating the change, adding a temporal character to the modeling [13]. This is key feature for modeling bioinformatics sequence related problems. These structures are also an efficient inference engine, where the inference process is similar to neural networks.

$$A_i^{(t+1)} = S\left(\sum_{j=1}^{n} w_{ji} A_i^{(t)}\right), i \neq j \quad (1)$$

Once the causal weights are established and the initial values for all input concepts are given, the values of all concepts are computed through time according to the above expression, where A_i represents the activation value of the *i-th* concept, w_{ji} is the causal weight between the concepts and S is transformation function for normalizing the resulting activation value. The inference process is repeated a fixed number of times or until the map stability is reached, also known as hidden pattern. FCMs were a suitable choice for modeling the problem enunciated before since they can describe the biological system similarly to the mental representations of the experts.

2.2 Protease Modeling

Protease protein is defined by a sequence of 99 amino acids. As was mentioned before, there exist relations among not necessarily adjacent positions of the sequence, due to 3D structure of the protein. Here, a change in a specific position (to be considered as a mutation) could be relevant on the resistance. Following figure 1 illustrates the topology designed for representing the protease protein as a FCM.

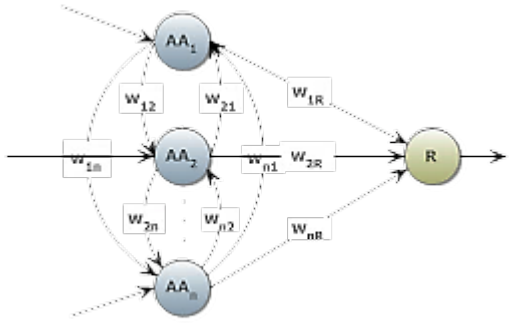

Fig. 1. FCM topology for representing HIV protease

Table 1. Average classification accuracy using a 5-fold cross-validation training evaluation (the best performing algorithm is emphasized in boldface)

Drug	DT	NN	LSR	SVR	LARS	RF	MLP	BRNN	FCM
APV	0.77	0.74	0.81	0.82	0.81	0.80	0.82	0.81	**0.87**
ATV	0.71	0.64	0.68	0.69	0.76	0.76	0.80	0.88	**0.92**
IDV	0.75	0.73	0.78	0.77	0.77	0.80	0.86	0.92	**0.97**
LPV	0.77	0.76	0.79	0.80	0.83	0.81	0.92	0.94	**0.95**
NFV	0.76	0.73	0.79	0.79	0.80	0.82	0.86	**0.93**	0.92
RTV	0.84	0.81	0.86	0.86	0.88	0.84	0.90	**0.94**	**0.94**
SQV	0.75	0.76	0.81	0.81	0.82	0.80	0.85	0.91	**0.94**

Each input concept stands for a sequence position described by one of the twenty possible amino acids. These nodes are fully connected with each other and also have causal influence on the resistance concept, which represents the output of the model. In the suggested configuration the authors describe the protease protein using the amino acids contact energies [14], which is a numerical descriptor statistically representing the proximity of an amino acid to the others, and to some extent describing the 3D structure of the protein. Also, it was proposed a feature selection based on sequence positions previously associated with resistance, in order to facilitate the interpretability of the final map.

In this case the causality of the map is automatically learned from historical data [3], using a variant of constricted PSO called PSO-RSVN [15, 16]. The supervised learning scheme applied is able to avoid stagnation and premature convergence states to local optima. Afterwards, the optimized maps characterize the resistance mechanisms of the protease protein for each antiviral drug taken into account. As a result, the prediction accuracies obtained in extended experiments using a 5-folds cross-validation were competitive with reported models in literature (see Table 1). Seven protease inhibitors were studied: Amprenavir (APV), Atazanavir (ATV), Indinavir (IDV), Lopinavir (LPV), Nelfinavir (NFV), Ritonavir (RTV) and Saquinavir (SQV). The algorithms used for comparison were: decision trees (DT), neural networks (NN), least-squares regression (LSR), support vector regression (SVR), and least angle regression (LARS) from [1]; a random forest (RF) using n-grams from [9], a multilayer perceptron (MLP) and a bidirectional recurrent neural network (BRNN) from [8].

Since the accuracy of the obtained maps constitutes a quality measure of the causality expressed in the relations, it was possible to develop a knowledge discovering process. The interpretability features of the FCMs helped to find causal patterns among the protein positions and the resistance. For each drug, positions with positive, negative or null causality over the resistance were identified and the effects of a punctual mutation in those positions were explained, offering useful information to drug specialists.

In the present paper, we extent these results starting from the aforementioned adjusted maps for generating susceptible possible mutations of the protease protein by using a discrete optimization approach.

3 Generating Susceptible Mutations of the HIV Protease Protein

As was discussed, the methodology proposed by Grau, Nápoles and coauthors [10, 11] allows analyzing the behavior of some HIV proteins, in order to understand the effect of mutations on the target drug resistance. As result, the existing causality between each protein position and the resistance concept was numerically established,

and also a novel scheme for simulating the effect of simple or multiple mutations was introduced.

However, from this work we notice that the FCM model not just is able to compute the biological causality, but also reported promising classification accuracies. More explicitly FCM model and the Swarm Intelligence based learning significantly outperformed other well-known classifiers such as Decision Trees, Support Vector Machine models, Multilayer Perceptron, Recurrent Neural Networks or Bayesian Networks. While those recurrent approaches computed best results.

As we known, HIV available knowledge bases are imbalanced, so there are many resistant cases, while susceptible mutations are quite limited. Clearly it could induce learning algorithms converge to local optima. Despite this inconvenient, FCM methodology detailed in [11] is quite robust to handle such situations. However, a serious drawback remains: the number of susceptible mutations reported in the scientific literature is still insufficient, limiting the comprehension of the HIV proteins behavior.

Then, it is possible to generate feasible mutations being susceptible to existing drugs? Next we introduce a novel scheme for mutating HIV *protease* protein using a learned FCM as suggested in [11] and a method based on ACO metaheuristic as generator of new mutations. Thus, the central idea of this scheme is to generate feasible mutations over the wild sequence as a typical combinatorial problem. To do that, we use the ACO metaheuristic where the information guiding the ant's movements is computed from biological knowledge extracted from historical data.

Perhaps the most relevant contribution from the machine learning point of view is that, in this scheme, the objective function value is computed through a learned FCM describing the protein behavior. In other words, the value of a candidate mutation is measured in terms of susceptibility to a specific drug, which is calculated over the FCM inference process. Of course, the FCM used in this method needs to be previously adjusted using the learning algorithm discussed in [11]. Following we justify this proposal more explicitly.

Mutations could be grouped into two groups: chromosomal mutations and gene mutations. The first ones are related with the chromosomes reordering, thus codifying changes in the molecular structure of the protein; while the second group is oriented to changing the nucleotides succession in the DNA sequence. Hereinafter, this work will be exclusively focused on gene mutations since most reported sequences in related literature are codified from this perspective. Nevertheless, the method introduced in this section could be easily adapted to chromosomal mutations as well.

As a further classification, gene mutations are grouped in four clusters as suggest [17]:

- **Silent mutations:** alter the current codon in a degenerated codon; it means that the amino acids codification does not suffer any modification. Such phenotypic mutations are unable to produce perceptible alterations, but instead they remain silent having a determinant role in the individual's evolution.
- **Frame shifted mutations:** these mutations induce the deletion or insertion of nucleotides over the protein sequence.
- **Missense mutations:** consist in the nucleotides replacement in codons which modify the interpretation of the codon, that is, the amino acids codification in the sequence.
- **Nonsense mutations:** transform a standard codon in a terminal codon (UAA, UAG, UGA). It is important to remark that, such mutations are particularly dangerous since they lead to the split of the proteomic chain.

As a remark, frame shifted mutations are quite frequent in reverse transcriptase mutations leading to instances having variable length [18]. But in protease mutations missense mutations are often reported in scientific literature. For this reason in this section we concentrate on generating artificial missense mutation for the protease protein, being susceptible to existing drugs. However, a suitable mechanism that allows generating these protease mutations is required. With this goal in mind, next subsection provides a brief background on ACO metaheuristic, and later the model design from a biological perspective will be introduced.

3.1 Ant Colony Optimization

The Ant Colony Optimization (ACO) metaheuristic is a well-known search method for solving combinatorial problems [19]. The biological idea behind this meta-heuristic is to simulate the behavior of a colony of individual agents (ants) when they are looking for food. Real ants in nature search for food in a random proximity to the nest. Once the ants found a source of food, they evaluate this source according to quality and quantity. Then, in the path back to the nest, they deposit a chemical pheromone trail on the ground, in order to guide the rest of the colony to the food source.

Inspired in this behavior, the ACO algorithm is a fully constructive model where each ant builds a solution of the problem by exploring a construction graph. The artificial ant moves from one state to another during the search process. Here states denote the components of the problem solution. In general terms, the preference of moving from one node to the other depends on two main values associated with each pathway:

– The artificial information η_{ij}, which is based in the pheromone trail deposited. It is iteratively updated by ants during the search process
– The heuristic information τ_{ij}, which is related with the application domain denoting the preference of moving from one state to another. It is important to notice that the heuristic information is known in advance and it is not updated during the search process.

In the search process, the probability of the k-th ant to move from state i to state j is computed by the expression (2); where $\mathcal{N}_i^k$ is the set of nodes that the ant has not yet visited, α and β are parameters specified by users for denoting the strength of the pheromone trail and the heuristic information on the decision, respectively.

$$P_{ij}^k(t+1) = \frac{\left[\tau_{ij}(t)\right]^{\alpha}\left[\eta_{ij}\right]^{\beta}}{\sum_{r \in \mathcal{N}_i^k}\left[\tau_{ir}(t)\right]^{\alpha}\left[\eta_{ir}\right]^{\beta}}, j \in \mathcal{N}_i^k \quad (2)$$

After the construction phase is complete, it is necessary to update de pheromone trails using the solutions found by ants. As a first stage, pheromone evaporation takes place uniformly reducing the pheromone trail in each path. Afterward, in a second moment, one or more solutions found are used to increase the value of such paths included in selected solutions. It is a sensible issue in the ACO metaheuristic. Actually, most of ACO variants primarily differ in the selected strategy for updating the pheromone trail at each cycle.

In this work we use a variant of ACO known as Max-Min Ant System (MMAS) [20]. The central features of this implementation are summarized as follows: (1) the allowed values for the pheromone trail are in the range $\tau_{min} < \tau_{ij} < \tau_{max,} \forall \tau_{ij}$ and (2) they are initialized with τ_{max} value, ensuring more exploration of the search space at the beginning of the search. Moreover, MMAS uses a strategy for updating pheromones very similar to Ant Systems [19], as describes following equation using the constant ρ, with $0 < \rho < 1$:

$$\tau_{ij}(t+1) = (1-\rho)\,\tau_{ij}(t) + \rho\tau_{ij}^{best}(t) \quad (3)$$

Consequently, a strong elitist criterion regulates the ant which is allowed to update the pheromone trail. It could be the one with better tour so far (global-best ant) or the one with the best solution in the current iteration (iteration-best ant). In general, this algorithm has strong exploration capabilities and it attempts to avoid the stagnation of the colony more effectively. In the next subsection we explain how to use the MMAS method for solving the optimization problem enunciated before.

3.2 Optimization Design Stage

Here, the idea is to generate reasonable mutations using the ACO metaheuristic. Towards this end, it is important to represents solutions. Thus, each ant needs to build a vector having cardinality equal to the number of punctual mutations that will be induced. In other words, as we know the protease is described by 99 amino acids but only a small subset of such positions is related with the protein mutations. For this reason ants will generate solutions having $1 \leq m \leq 99$

components, that is to say, protein sequences with m mutated positions. In this paper these m positions are taken as those that have been previously associated with the drug resistance target. For each drug, selected positions were determined using both numerical and biological perspective [3, 21-24].

Then a solution (to be considered a mutation) is a vector, where each position codifies a specific amino acid expressed by their contact energy. It is relevant to mention that those protein positions that won't be muted preserve the amino acids of the wild sequence. This sequence for protease protein has the form: "PQITLWQRPLVTIKIGGQLKEALL DTGADDTVLEEMNLPGRWKPKMIGGIGGFIKVRQYDQILIE ICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNF".

For better comprehension of this scheme, lets to generate a protease mutation for the drug IDV having 5 mutation points. For this drug, five of the positions directly related with the resistance target are: AA_{10}, AA_{24}, $AA_{46,}$ AA_{71} and $AA_{90.}$ It means that ants will build solutions with five components, that is, with five possible amino acids denoting the protein mutations. Now suppose that the optimal sequence found by the algorithm has the following form: "PQGTL". It implies that the final mutation has the codification: "PQITLWQRP**P**VTIKIGGQLKEAL**Q** DTGADDTVLEEMNLPGRWKPK**G**IGGIGGFIKVRQYDQILI EICGHK**T**IGTVLVGPTPVNIIGRNL**L**TQIGCTLNF".

Notice that the cardinality of the optimization problem is the number of mutations points, where the number of such points is fixed as the number of protein positions previously associated with the resistance. Hence, in the optimization stage, the k-th ant will select the j-th amino acid as the i-th solution component mainly based on the pheromone trail information and the heuristic information. The pheromone trail is learned by ants during the search process, but the heuristic information should be designed by the user.

Actually, the erroneous choice of this information frequently leads to poor solutions. Hence, how to efficiently estimate the heuristic preference for the optimization scheme? As a suitable alternative we use biological knowledge achieved from historical data. The central idea consist in quantify how many a protein position mutates to a specific amino acids. As an illustrative example, analyzing 150 mutations we noticed that the position AA_{30} mutates to the amino acid D in 104 sequences for drug LPV. It means that the heuristic preference of accepting this amino acid in the position AA_{30} will be $\eta_{ij} = 104/150 \approx 0.7$.

However, due to available historical data are frequently imbalanced as was highlighted before, the proposed strategy for estimating the heuristic component instead of benefiting may negatively affect the global convergence rate of the optimization algorithm. More explicitly, the heuristic value of accepting an amino acid in a specific position probably leads to a resistant mutation, since that the heuristic component is estimated using historical data where most sequences are resistant to existing drugs. Despite this, the proposed strategy allows to simulate more naturally the HIV mutation mechanism, consequently reducing the probability of generating no feasible sequences.

In order to complete the optimization design, a function $f: R^n \rightarrow [0,1]$ is required. In this paper we use a previously adjusted FCM to compute the resistance concepts for a given mutation. Hence, once the FCM is trained using historical data, the system behavior could be studied by simply varying the concept's activation values. To do that, each mutated sequence position is directly related with the corresponding contact energy. Next, the FCM inference mechanism is activated and the resistance concept is examined: the closer to zero the resistant concepts is, the more susceptible the artificial mutation is. In next section this methodology is used for generating mutations having low resistant values for existing inhibitors.

4 Simulations and Discussion

With the intention of validating our proposal, in this section we obtain several susceptible *protease* mutations by using the workflow described above. To do that, we use the following as parameter settings: 20 ants, 100 generations, the evaporation constant ρ is set to 0.1, while the transition rule parameters are $\alpha = 2$ y $\beta = 3$.

Table 2. Example of mutation frequencies on each sequence position, extracted from historical data of Loinavir, which constitutes heuristic data for the ACO optimization

AA	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Z
10	0	0	0	0	24	0	0	61	0	11	0	0	0	0	0	0	0	10	0	0
20	0	0	0	0	0	0	0	6	65	0	1	0	0	0	28	0	5	1	0	0
24	0	0	0	0	1	0	0	13	0	92	0	0	0	0	0	0	0	0	0	0
30	0	0	104	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0
32	0	0	0	0	0	0	0	12	0	0	0	0	0	0	0	0	0	94	0	0
33	0	0	0	0	30	0	0	2	0	73	0	0	0	0	0	0	0	1	0	0
36	0	0	0	0	0	0	0	43	0	6	56	0	0	0	0	0	0	1	0	0
46	0	0	0	0	0	0	0	51	0	14	41	0	0	0	0	0	0	0	0	0
47	0	0	0	0	0	0	0	96	0	0	0	0	0	0	0	0	0	10	0	0
48	0	0	0	0	0	96	0	0	0	0	1	0	0	0	0	1	0	8	0	0
54	1	0	0	0	0	0	0	32	0	6	4	0	0	0	0	3	0	60	0	0
63	4	2	0	0	0	0	2	0	0	10	0	0	85	0	1	0	2	0	0	0
64	0	0	0	0	0	0	0	87	0	0	3	0	0	0	0	0	0	16	0	0
70	0	0	0	4	0	0	0	0	101	0	0	0	0	0	1	0	0	0	0	0
71	31	0	0	0	0	0	0	8	0	1	0	0	0	0	0	0	9	57	0	0
72	0	0	0	1	0	0	0	84	1	1	6	0	0	0	3	0	2	8	0	0
73	2	0	0	0	0	86	0	0	0	0	0	0	0	0	0	15	3	0	0	0
77	0	0	0	0	0	0	0	32	0	0	0	0	0	0	0	0	0	74	0	0
82	59	1	0	0	4	0	0	1	0	0	0	0	0	0	0	1	9	31	0	0
84	1	1	0	0	0	0	0	72	0	0	0	0	0	0	0	0	1	31	0	0
88	0	0	4	0	0	1	0	0	0	0	0	100	0	0	0	0	1	0	0	0
90	0	0	0	0	0	0	0	0	0	46	60	0	0	0	0	0	0	0	0	0
93	0	0	0	0	0	0	0	63	0	42	1	0	0	0	0	0	0	0	0	0

Table 3. Percent of coincidence between the generated susceptible mutations and the outputs of the three well-known expert systems publicly available

Drug	ANRS	HIVdb	REGA
ATV	1.0	0.98	1.0
IDV	1.0	0.96	0.99
LPV	0.98	0.90	0.94
NFV	0.98	0.93	0.95
SQV	0.99	0.92	0.99

Table 4. Example of susceptible mutations obtained for each drug

Drug	Amino Acids Code	Nucleotides Code
APV	PQITLWQRPHVTIKIGGQLF EAGTDTGADITVLEEHNLPG RWKPKYIGGIGGCYKVRQY DQIPIEICGHKFIGTVTVGPT PCNSIGRRLCTQKGCTLNF	CCGCAAATTACCCTGTGGCAACGACCCCACGTGACTATCAAGATCGGA GGGCAACTCTTTGAGGCAGGGACTGACACAGGCGCGGATATCACAGT TCTGGAGGAACACAACCTACCCGGTCGATGGAAACCGAAGTACATAGG GGGGATTGGCGGATGCTATAAGGTAAGACAATATGATCAGATTCCTAT CGAGATCTGCGGACATAAGTTCATAGGTACGGTGACTGTAGGCCCTAC ACCATGTAATTCGATCGGGCGTCGCCTGTGCACCCAGAAGGGATGTAC ATTAAACTTT
ATV	PQITLWQRPIVTIKIGGQLKE ALLDTGADDTFSETANLPG RWKPKREGGEGGGIKVRQ YDQIFIEICGHKQIITVLDGPT PVNIIGRNLTTQGGCTLNF	CCACAAATTACACTTTGGCAACGACCGATCGTGACAATCAAGATCGGT GGGCAACTAAAGGAAGCCTTGCTGGATACAGGAGCGGACGACACCTT CAGCGAGACCGCCAACCTCCCAGGACGTTGGAAACCCAAAAGGGAAG GTGGTGAAGGTGGTGGAATTAAAGTACGTCAGTACGACCAAATTTTTAT CGAGATCTGCGGCCATAAACAAATTATAACCGTATTAGATGGTCCCACC CCCGTCAATATTATAGGAAGAAACTTGACCACACAGGGCGGGTGTACT CTTAACTTT
IDV	PQITLWQRPVVTIKIGGQLE EALLDTGADDTVCEEQNLP GRWKPKRIGGYGGFTKVR QYDQILIEICGHKWIPTVLVG PTPANSIGRKLLTQGGCTLN F	CCGCAGATCACACTCTGGCAACGGCCTGTCGTCACGATCAAGATCGG GGGACAACTTGAGGAGGCTTTACTCGACACGGGAGCGGATGATACTG TCTGTGAAGAACAGAACCTTCCCGGAAGATGGAAGCCTAAACGCATAG GGGGCTATGGGGGTTTTACGAAGGTCCGGCAATATGATCAGATTCTCA TAGAGATATGCGGCCATAAATGGATACCTACAGTGCTCGTGGGGCCAA CCCCCGCCAATAGTATTGGTAGGAAGCTTTTAACGCAGGGCGGATGCA CTTTGAACTTC
LPV	PQITLWQRPKVTIKIGGQLA EALLDTGADQTSQEEDNLP GRWKPKDQGGIGGFKKVR QYDQIPYEICGHFGNGTVL MGPTPWNLIGRPLRTQRGC TLNF	CCCCAAATAACCCTTTGGCAACGACCTAAAGTCACCATTAAAATTGGTG GACAACTCGCAGAGGCGCTGTTAGACACTGGTGCCGACCAAACATCTC AGGAAGAAGATAACCTGCCGGGGAGGTGGAAACCTAAAGATCAAGGG GGCATAGGAGGGTTCAAAAAAGTACGGCAATACGATCAGATTCCCTAT GAGATCTGCGGACATTTTGGAAATGGAACGGTACTAATGGGGCCAACA CCGTGGAACCTCATAGGACGTCCTCTGAGAACCCAACGTGGATGCAC GCTTAACTTC
NFV	PQITLWQRPTVTIKIGGQLD EALLDTGADITVLEEPNLPG RWKPKCIGGIGGFFKVRQY DQIPIEICGHKVIWTVLYGPT PKNQIGRWLATQAGCTLNF	CCGCAGATCACCCTCTGGCAACGACCGACCGTGACTATAAAAATCGGT GGCCAGCTCGACGAGGCACTACTAGATACTGGAGCTGATATCACTGTC TTGGAGGAGCCGAATCTACCCGGTCGCTGGAAACCTAAATGCATTGGA GGAATAGGCGGATTCTTTAAAGTACGCCAATACGACCAGATACCGATT GAGATCTGTGGGCATAAAGTGATATGGACTGTGCTGTACGGGCCAACA CCGAAGAATCAAATTGGAAGGTGGCTAGCAACGCAGGCCGGATGCAC TTTGAACTTC
RTV	PQITLWQRPIVTIKIGGQLAE ALLDTGADMTVREEWNLPG RWKPKEIRGIGGFVKVRQY DQIQIEICGHKRIRTVLVGPT PENRIGRPLITQIGCTLNF	CCACAGATCACCCTCTGGCAAAGACCCATTGTTACCATTAAGATAGGA GGGCAGTTGGCCGAGGCGCTCCTAGACACTGGCGCGGACATGACAGT CAGGGAAGAGTGGAATTTGCCCGGACGGTGGAAGCCGAAAGAGATCC GTGGCATAGGCGGTTTTGTGAAGGTGAGGCAATACGACCAAATCCAAA TTGAGATATGCGGGCATAAACGGATCCGAACCGTCCTCGTCGGGCCAA CACCAGAAAACCGTATTGGACGTCCTCTAATTACACAAATCGGGTGTAC TCTCAACTTT
SQV	PQITLWQRPFVTIKIGGQLS EALLDTGADFTVLEEVNLPG RWKPKHIIGIGGFIKVRQYD QIVIEICGHKSYMTVLVGPT PFNPIGRNLVTQMGCTLNF	CCTCAGATAACGTTATGGCAGAGGCCCTTCGTCACCATTAAGATAGGG GGCCAATTGTCGGAGGCTCTGTTAGATACTGGTGCGGACTTTACAGTG CTGGAAGAGGTGAACCTACCGGGCCGCTGGAAGCCAAAACATATCATC GGAATCGGTGGCTTCATTAAAGTGCGTCAGTACGACCAGATCGTGATA GAGATCTGCGGACATAAGTCTTACATGACAGTTCTGGTTGGACCCACA CCTTTCAACCCTATAGGTCGTAACCTCGTGACGCAAATGGGATGTACG TTGAACTTC

The value of the pheromone trail is initialized with τ_{max}, the maximum (τ_{max}) and the minimum (τ_{min}) value of pheromone are calculated according to the expressions $\tau_{min} = \tau_{max}/10 * n$ and $\tau_{max} = (1/1-\rho) * (1-F)$, where F denotes the best solution found so far. The Table 2 illustrates an example of the heuristic information used for generating mutations which are susceptible to Lopinavir, similar frequency matrixes are used for the other antiretroviral drugs.

As a result, we obtain several susceptible mutations for each trained map representing an antiviral drug. Ideally, biological experiments are needed to verify whether our mutations are really susceptible or not, but these experiments are very costly. So, in order to validate the susceptibility of our generated sequences, we compare the output of three well-known experts systems from: ANRS Agence Nationalede Recherches sur le SIDA [25], HIVdb Drug Resistance Interpretation Algorithm [7], and Rega Institute [26].

These experts systems are rules-based algorithms, where rules are Boolean expression and they are frequently updated and widely accessible. Table 3 shows the percent of coincidence between the generated susceptible mutations and the outputs of the expert systems, using 100 possible susceptible sequences for the available drugs in the expert systems. The values obtained illustrate the accuracy of our proposal. As a further result, Table 4 shows seven susceptible mutations expressed in amino acids and nucleotides codes, one for each modeled and learned FCM (see Section 2). Here, the conversion to nucleotides was performed using the sequence conver*sion tool from [27].*

Then, from the HIVdb expert system we extracted some interesting comments about the mutations processed. For example for the LPV susceptible generated mutation, is an interesting fact that it is not just susceptible to LPV, but to all others, including Darunavir and Fosamprenavir, which are non-studied drugs in this work due to the lack of associated historical data. In addition, it could be a subtype B mutation (which is the most common subtype in literature) with probability 0.64. Also, it has no stop codons or frame shifts, since we only simulated missense mutations.

PR Comments

PIMinor

- L10I/V/F/R/Y are associated with resistance to most PIs when present with other mutations. L10I/V occur in 5-10% of untreated persons. L10F is a non-polymorphic mutation which is associated with decreased susceptibility to all PIs except ATV/r, SQV/r, and TPV/r. L10R/Y are rare poorly characterized mutations.
- D30N causes high-level resistance to NFV. D30P is a highly unusual mutation at this position.
- M46I/L decreases susceptibility to IDV/r, NFV, FPV/r, LPV/r, and ATV/r when present with other mutations. M46V is an uncommon PI-selected mutation at this position. M46K is a highly unusual mutation at this position.
- I47V decrease susceptibility to FPV/r, ATV/r, IDV/r, LPV/r, TPV/r, and DRV/r. I47A usually occurs with V32I and in this setting causes high-level LPV/r and FPV/r resistance and decreased DRV/r susceptibility. I47Y is a highly unusual mutation at this position.
- I54V/M/L/A/T/S have diverse effects on PI susceptibility. I54K is a highly unusual mutation at this position.
- V82A/T/F/L/M/S/C have diverse effects on multiple PIs. V82D is a highly unusual mutation at this position.
- L90M reduces susceptibility to NFV, SQV/r, ATV/r, and IDV/r. When present with other mutations it also reduces susceptibility to FPV/r and LPV/r. L90R is a highly unusual mutation at this position.

Other

- K20R/M/I/T/V are associated with resistance to multiple PIs when present with other mutations. K20C is a highly unusual mutation at this position.
- L63P is a common polymorphism that is also selected by PIs.
- N88S causes high-level resistance to NFV and ATV/r and low-level resistance to IDV/r; it increases susceptibility to FPV/r. N88T/G are rare PI-selected mutations that have much less pronounced effects than N88S. N88E is a highly unusual mutation at this position.

Fig. 2. Comments extracted from the HIVdb expert system about the submitted LPV-susceptible mutation

Figure 2 illustrates the output comments of the expert system on the mutation identified as minor (frequent) or other (less frequent). No major mutation was identified for this particular sequence. In the notation, the first letter means the wild amino acid (source), the number represents the sequence position to be mutated and the last letters stands for the changed amino acids (target mutations).

As a result, the susceptible mutations generated by our methodology could be incorporated to knowledge bases in order to enlarge the historical data available, treating the imbalanced distribution of classes and thus facilitating the study of drug resistance mechanisms in HIV proteins.

5 Conclusions

The complex dynamic and high mutation rate of the HIV leads to serious problem on designing more effective drugs. Particularly, it is known that this retrovirus frequently develop resistance to the existing drugs, thus causing the treatment failure. For this reason, other biological, mathematical or computational approaches allowing understanding this virus are required. Several machine learning methods have been applied for solving the related classification problem, but only a few are interpretable (for instance, Recurrent Neural Networks reported good accuracies in terms of classification rate, but we can't explain the underlying interaction among protein amino acids). Attempting to deal with this issue, the authors introduce a novel modeling based on FCM theory, allowing not just efficiently classify new mutations but also numerically computing the causal influence of the amino acids over the drug resistance concepts.

In this work we extend the above mentioned research, now with the goal of generating artificial mutations. To do that, we use a scheme based on ACO meta-heuristic to compute sequences having lower contact energy. Here, the objective value associated to each solution is computed by using a previously adjusted FCM. Besides, we use biological knowledge obtained from historical data to estimate the heuristic preference of mutations in each protein position, which ensures to generate feasible solutions.

At the end, seven new susceptible mutations are reported, hence confirming the reliability of our methodology. It could contribute, in a certain sense, to understand the behavior the HIV resistant mechanism; since most of the available HIV mutations are highly resistant to existing drugs. More generally, the proposed scheme may be easily adapted to other problems in order to modify datasets with highly imbalanced distribution of classes, where a classifier exists but historical data having a desirable property are insufficient for obtaining consistent performance.

Acknowledgements

We would like to thank Prof. Yovani Marrero-Ponce, PhD, from Unit of Computer-Aided Molecular Biosilico Discovery and Bioinformatics Research, UCLV, for fruitful discussions about the original ideas of this research.

References

1. **Rhee, S.Y., Taylor, J., Wadhera, G., Ben-Hur, A., Brutlag, D.L., & Shafer, R.W. (2006).** Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *Proceedings of the National Academic of Sciences of the United States of America*, 103(46), 17355–17360.
2. **Beerenwinkel, N., Schmidt, B., Walter, H., Kaiser, R., Lengauer, T., Hoffmann, D., Korn, K., & Selbig, J. (2002).** Diversity and complexity of HIV-1 drug resistance: a bioinformatics approach to predicting phenotype from genotype. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12), 8271–8276.
3. **Shafer, R.W. (2006).** Rationale and Uses of a Public HIV Drug-Resistance Database. *Journal of Infectious Diseases*, 194(1), 51–58.
4. **López, V., Fernandez, A., del Jesus, M.J., Herrera, F. (2013).** A Hierarchical Genetic Fuzzy System Based On Genetic Programming for Addressing Classification with Highly Imbalanced and Borderline Data-sets. *Knowledge-Based Systems*, 38, 85–104.
5. **Beerenwinkel, N., Lengauer, T., Selbig, J., Schmidt, B., Walter, H., Korn, K., Kaiser, R., & Hoffmann, D. (2001).** Geno2pheno: interpreting

genotypic HIV drug resistance tests. *IEEE Intelligence System*, 16(6), 35–41.

6. **Draghici, S. & Potter, R.B. (2003).** Predicting HIV drug resistance with neural networks. *Bioinformatics*, 19(1), 98–107.
7. **Liu, T.F. & Shafer, R.W. (2006).** Web Resources for HIV type 1 Genotypic-Resistance Test Interpretation. *Clinical Infectious Diseases,* 42(11), 1608–1618.
8. **Bonet, I., Garcia, M.M., Saeys, Y., Peer, Y.V., & Grau, R. (2007).** Predicting Human Immunodeficiency Virus (HIV) Drug Resistance Using Recurrent Neural Networks. *Bio-inspired Modeling of Cognitive Tasks. Lecture Notes in Computer Science*, 4527, 234–243.
9. **Masso, M. (2012).** Prediction of Human Immunodeficiency Virus Type 1 Drug Resistance: Representation of Target Sequence Mutational Patterns via an n-Grams Approach. *2012 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Philadelphia, PA, 1–6.
10. **Grau, I., Nápoles, G., León, M., & Grau, R. (2012).** Fuzzy Cognitive Maps for modelling, predicting and interpreting HIV drug resistance. *Advances in Artificial Intelligence – IBERAMIA 2012. Lecture Notes in Computer Science*, 7637, 31–40.
11. **Nápoles, G., Grau, I., León, M., & Grau, R. (2013).** Modelling, aggregation and simulation of a dynamic biological system through Fuzzy Cognitive Maps. *Advances in Computational Intelligence*. *Lecture Notes in Computer Science*, 7630, 188–199.
12. **Kosko, B. (1986).** Fuzzy Cognitive Maps. *International Journal of Man-Machine Studies*, 24(1), 65–75.
13. **Papageorgiou, E.I. (2012).** Learning Algorithms for Fuzzy Cognitive Maps – A Review Study. *IEEE Transactions on Systems, Man and Cybernetics*, *Part C: Applications and Reviews*, 42(2), 150–163.
14. **Miyazawa, S. & Jernigan, R.L. (1999).** Self-Consistent Estimation of Inter-Residue Protein Contact Energies Based on an Equilibrium Mixture Approximation of Residues. *Proteins,* 34(1), 49–68.
15. **Nápoles, G., Grau, I., & Bello, R. (2012).** Constricted Particle Swarm Optimization based algorithm for global optimization. *Polibits*, 46, 5–11.
16. **Nápoles, G., Grau, I., & Bello, R. (2012).** Particle Swarm Optimization with Random Sampling in Variable Neighbourhoods for solving Global Minimization Problems. *Swarm Intelligence. Lecture Notes in Computer Science*, 7461, 352–354.
17. **Volkenshtein, M.V. (1981).** BiophysicsMoscow: Mir.
18. **Grau, I., Nápoles, G., Bonet, I., & García, M.M. (2013).** Backpropagation Through Time Algorithm for Training Recurrent Neural Networks Using Variable Length Instances. *Computación y Sistemas*, 17(1), 15–24.
19. **Dorigo, M., Maniezzo, V., & Colorni, A. (1996).** Ant System: Optimization by a colony of cooperating agents. *IEEE Transactions on Systems, Man, and Cybernetics*, *Part B: Cybernetics,* 26(1), 29–41.
20. **Stutzle, T. & Hoos, H.H. (2000).** MAX-MIN ant system. *Future Generation Computer System,* 16(8), 889–914.
21. **Shafer, R.W. (2002).** Genotypic testing for Human Immunodeficiency Virus type 1 drug resistance. *Clinical Microbiology Reviews*, 15(2), 247–277.
22. **Johnson, V., Brun-Vézinet, F., Clotet, B., Conway, B., D'Aquila, R.T., Demeter, L.M., Kuritzkes, D.R., Pillay, D., Schapiro, J.M., Telenti, A., & Richman, D.D. (2003).** Drug resistance mutation in HIV-1. *Topics in HIV Medicine*, 11(6), 215–221.
23. **Sing, T., Svicher, V., Beerenwinkel, N., Ceccherini-Silberstein, F., Däumer, M., Kaiser, R., Walter, H., Korn, K., Hoffmann, D., Oette, M., Rockstroh, J.K., Fätkenheuer, G., Perno, C.F., & Lengauer, T. (2005).** Characterization of novel HIV drug resistance mutations using clustering, multidimensional scaling and SVM-based feature ranking. *Knowledge Discovery in Databases. Lecture Notes in Computer Science*, 3721, 285–296.
24. **Woods, M. & Carpenter, G.A. (2007).** Neural Network and Bioinformatics Methods for Predicting HIV-1 Protease Inhibitor Resistance (Technical Report 2007-004). Boston, Massachusetts: Boston University.
25. **Brun-Vezinet, F., Costagliola, D., Khaled, M.A., Calvez, V., Clavel, F., Clotet, B., Haubrich, R., Kempf, D., King, M., Kuritzkes, D., Lanier, R., Miller, M., Miller, V., Phillips, A., Pillay, D., Schapiro, J., Scott, J., Shafer, R., Zazzi, M., Zolopa, A., & DeGruttola, V. (2004).** Clinically validated genotype analysis: guiding principles and statistical concerns. *Antiviral Therapy*, 9(4), 465–478.
26. **Van Laethem, K., De Luca, A., Antinori, A., Cingolani, A., Perna, C.F., Vandamme, A.M. (2002).** A genotypic drug resistance interpretation algorithm that significantly predicts therapy

response in HIV-1–infected patients. *Antiviral Therapy,* 7(2), 123–129.

27. **In-Silico Sequence Conversion Tools.** http://in-silico.net/tools/biology/sequence_conversion.

Isel Grau received the BSc. degree (with honors) in computer sciences from the Universidad Central "Marta Abreu" de Las Villas (UCLV), Cuba in 2011. She has authored/ coauthored 10 papers in conference proceedings and scientific journals. She earned the Cuban National Award for Computer Science Students in 2010, the Cuban National Award for Best Diploma Thesis in 2011, the Cuban Academy of Sciences Award twice (2011 and 2012) and the Best Paper Award Third Place at MICAI 2012 conference. Her research interests include soft computing, bioinformatics, recurrent neural networks, statistics and machine learning.

Gonzalo Nápoles received the BSc. degree (with honors) in computer sciences from the Universidad Central "Marta Abreu" de Las Villas (UCLV), Cuba in 2011. He has authored /coauthored two monographs and 15 referred papers in conference proceedings and scientific journals. He earned the Cuban National Award for Computer Science Students twice (2010 and 2011), the Academy of Sciences Award in 2012, and the Best Paper Award Third Place at MICAI 2012 conference. His research interests include soft computing, fuzzy cognitive maps, metaheuristics, evolutionary computation, machine learning and knowledge discovering.

Article received on 03/04/2013, accepted on 05/07/2013.