MRBF: A METHOD FOR PREDICTING HIV-1 DRUG RESISTANCE

Anantaporn Srisawat and Boonserm Kijsirikul
Computer Engineering Department, Chulalongkorn University, Thailand

Abstract*:* This paper presents the MRBF network, a new algorithm adapted from the RBF network, to construct the classifiers for predicting phenotypic resistance on 6 protease inhibitors. The performance of the prediction was measured by 10-fold cross-validation. The results show that MRBF gives the lowest average mean square error (MSE) when compared with the traditional RBF network and multiple linear regression analysis (REG). Moreover, it provides the best average predictive accuracy when compared with HIVdb, REG, and Support Vector Machines (SVM).

Key words: RBF Network, RReliefF, predicting HIV-1 drug resistance

1. INTRODUCTION

Nowadays, there are seventeen approved antiretroviral agents: seven drugs for Nucleoside Reverse Transcriptase Inhibitor (NRTI), three drugs for Non-Nucleoside Reverse Transcriptase Inhibitor (NNRTI), and seven drugs for Protease Inhibitor (PI), but HIV-1 therapies are still not very successful. The limit of treatment success is the decrease of the viral sensitivity to the drug called drug resistance. The cause of drug resistance is the mutations in the reverse transcriptase (RT) and protease enzymes of HIV-1. In addition, it has been estimated that every possible single point mutation occurs between 10^4 and 10^5 times per day in an untreated HIV-1 infected individual and that double mutants also occur commonly [1]. Thus resistance testing is an important role in management of HIV infections.

Currently there are two methodologies for resistance testing: genotyping and phenotyping [2]. For genotyping, resistance testing can be performed by scanning the viral genome for resistance-associated mutations, where phenotyping can be performed by measuring viral activity in the presence and absence of drug. The advantages of genotyping are faster and cheaper than phenotyping. On the other hand, phenotypic results are easier to interpret than genotypic results because the phenotypic results are represented by a single number for each drug called fold change.

The fold change refers to the fraction between 50% inhibitory drug concentration value (IC_{50}) of the patient's virus to the IC_{50} value of the standardized wild type virus ($IC_{50(\text{patient})}$/ $IC_{50(\text{reference})}$). If the fold change is above a certain value called cutoff the virus is resistant to that drug.

To overcome the drawbacks of genotyping and phenotyping methods, the advantage of genotyping and phenotyping are combined by using genotypic data to predict phenotypic results. This paper proposes a new method, called Multi-RBF (MRBF) network. This method applied the Radial Basis Function (RBF) network for predicting the fold change of 6 protease inhibitors (PI): saquinavir (SQV), indinavir (IDV), ritonavir (RTV), nelfinavir (NFV), amprenavir (APV), and lopinavir(LPV).

Since the number of amino acid positions of HIV-1 protease gene is quite large (99 positions), we also used the RReliefF algorithm [3], a feature subset selection technique, to select the amino acid positions that are considered to be relevant to the drug susceptibility and eliminate irrelevant amino acid positions.

2. RELATED WORKS

A variety of techniques have been applied to predict phenotype from genotype such as rule-based, statistical analysis, and machine learning techniques. The phenotypic results from these techniques are classified into two or more classes of drug susceptibility depending on the certain cutoff values.

Rule-based algorithms such as HIVdb [4], ANRS [5], Rega [6], and VGI [7] contain the rules encoding information from the medical literature as the knowledge base. The HIVdb system used the mutation scoring tables to calculate a score from each sequence and interpreted drug susceptibility into one of five classes ranging from susceptible to high-level resistant.

For statistical analysis, multiple linear regression analysis (REG) was applied to construct a separate regression model for each drug [8]. In the model, the dependent variable is the logarithm of the IC50 fold change, while the independent variables are dummy variables corresponding to mutations. In addition, this technique used the stepwise regression method to optimize the parameters for each independent variable.

Besides rule-based and statistical analysis, machine learning is the most popular approach applied to predict phenotype from genotype. Many supervised learning algorithms have been used to deal with this problem such as decision trees [9, 10], support vector machines (SVMs) [9, 11], and artificial neural networks (ANNs) [12]. These algorithms classify drug susceptibility into one of two classes:

susceptible or resistant. Furthermore, the self-organizing map (SOM), an unsupervised learning algorithm, was used to classify drug susceptibility into one of three classes: high, medium, or low resistant [13].

3. RADIAL BASIS FUNCTION (RBF) NETWORK

The RBF network is an approach for function approximation that is closely related to distance-weighted regression and also to artificial neural networks [14, 15, 16]. The construction of the traditional RBF network involves three layers with entirely different roles as illustrated in Figure 1.

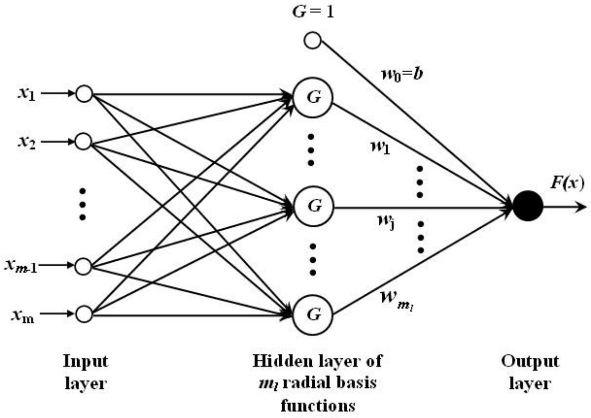

*Figure 1.*The RBF Network

As shown in Figure 1, the RBF network consists of three layers [17]. The first layer is composed of input nodes whose number is equal to the dimension of the input vector. The second layer is a hidden layer. This layer consists of nonlinear units that are connected directly to all of the nodes in the input layer. The activation functions of the individual hidden units are defined by a Gaussian function. The output layer consists of a single linear combination unit, being fully connected to the hidden layer. In this approach, the value of the output unit is a function given in (1).

$$F(x) = w_0 + \sum_{i=1}^{m_1} w_i G(\|x - t_i\|) \tag{1}$$

where m_1 is the number of centers, vector t represents the center points, vector w is the weights in the output layer, and G is the Gaussian function (see Figure 1).

In training step, the weight vector w in the output layer of the network will be calculated by matrix computation as shown in (2).

$$w = G^+ d \tag{2}$$

Where G^{+} is the pseudo inverse of matrix G defined in (3) and *d* is the desired response vector in the training set.

$$G^{+} = (G^{T}G)^{-1}G^{T} \tag{3}$$

where
$$G = \{g_{ji}\} \tag{4}$$

$$g_{ji} = \exp\left(-\frac{\left\| x_j - t_i \right\|^2}{2\sigma_i^{\,2}} \right) \tag{5}$$

where i=j=1,2,…,m_1, x_j is the *j* th input vector of the training sample and t_i is the *i* th vector of the center and σ denotes the width of the Gaussian function.

4. MULTI RBF (MRBF) NETWORK

To improve the performance of the RBF network in estimating the IC50 fold change (FC) for predicting HIV-1 drug resistance, we present a new approach called Multi RBF (MRBF) network. The idea of MRBF is to separately construct the RBF networks class by class to increase the ability of estimating the output value. This method consists of three RBF networks: RBF-all, RBF-S, and RBF-R for estimating the IC50 fold change. The construction of an MRBF network is shown in Figure 2.

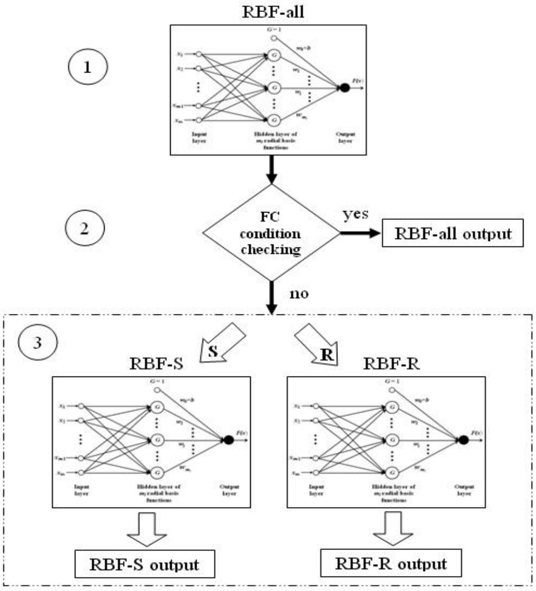

*Figure 2.*The MRBF Network

In the MRBF learning step, the calculation of vector *w* of each network of the MRBF is the same as the RBF network described in section 3. The center vectors in a hidden layer of 3 networks have to be determined in a different way. The centers of the RBF-all network are whole training examples whereas the RBF-S uses only the training examples belonging to the susceptible class and vice versa for the RBF-R network.

There are three parts in the testing process of the MRBF network. First, an RBF-all network (see Figure 2) roughly estimates the FC of an instance *x* , then it uses the logarithm of the cutoff value to classify the FC of the instance *x* into one of two classes: susceptible or resistance.

The second part is called FC condition checking. This step checks the output of the instance *x* from the RBF-all network with two criteria. The criteria are following.
1. FC falls in the boundary zone.
2. The class labels between RBF-all and kNN algorithm are different.

If the FC satisfies both criteria, the final FC of the instance *x* is the output from the RBF-all network, and the testing process is terminated. If any of them is not satisfied, the instance *x* will be fed into the RBF-S or RBF-R network.

The boundary zone in the first criterion has the value between cutoff-*bound* and cutoff+*bound*, where *bound* is calculated by (6).

$$bound = \sqrt{\frac{\sum_{i=1}^{n}(FC_i - cutoff)^2}{n}} \tag{6}$$

where *n* is total training instances, FC_i is target value of the training instance *i*.

For the second criterion, kNN classification is used to measure the confidence of the prediction from the RBF-all network. In another word, if the output of RBF-all is the same as that of kNN, it is probable that RBF-all gives the correct classification result, and thus the instance *x* is fed to the third part to estimate more precise FC. On the other hand, if the output of RBF-all and kNN are not the same, the testing process is terminated since RBF-all may misclassify. This condition prevents feeding the instance *x* into the wrong network in the third step. As the distance between the training instances are already computed in the learning process of RBF-all, kNN is a suitable technique for checking the confidence of the RBF-all network.

In the third part, the instance *x* is fed again into another RBF network: RBF-S or RBF-R depending on its class label. If the FC from RBF-all of the instance *x* is labeled as susceptible, the instance *x* will be an input of the RBF-S network. On the other hand, it will be fed into the RBF-R, if its output is labeled as resistant. Finally, the final FC of instance *x* is the output from the corresponding network.

5. FEATURE SELECTION PROCESS

Since the total amino acid positions of HIV-1 protease gene are 99 and some of them are irrelevant or redundant, these attributes may decrease the performance of the learning algorithm. To solve this problem, we used feature selection techniques to select important attributes in the preprocessing step.

Moreover, the time complexity of learning the RBF network is depended on the number of dimension of the input. It takes O(mn^2) for calculating Gaussian functions of the hidden layer, where m is the number of input nodes and n is the number of training instances. When an amino acid position is transformed to a vector of the input node of the RBF network, each amino acid position provides 20 binary input nodes (there are 20 variables of amino acids which may occur in any position). As there are 99 positions in the sequence of an HIV-1 protease gene, the number of input node is 1980. Thus, it takes a lot of time in the MRBF learning step. For that reason, if we use feature selection techniques to select some important amino acid positions (or attributes) to construct the MRBF model instate of using all positions, the number of the input nodes will be reduced significantly. Subsequently the learning time is decreased.

In this paper, three different approaches of feature selection techniques are used in the preprocessing process: Allmutant, Rule-based, and RReliefF. In Allmutant approach, the attribute with only one value on the total transactions of each drug was eliminated. For Rule-based, we selected the important attributes, recommenced by Stanford HIV Drug Resistance Database and [18]. For RReliefF, we ran RReliefF, a classical feature estimation algorithm, to select important attributes for each drug. The main idea of RReliefF is to estimate the weight of each attribute according to how well its value distinguishes between instances that are near to each other [3]. Using RReliefF, we selected the attributes, which have the weight higher than or equal to θ, where θ was set to 0.01. Table 1 shows the number of input attributes selected by each feature selection approach.

Table 1. The number of input attributes for each feature selection approaches

Feature selection approach	Number of input attributes for each drug					
	LPV	APV	NFV	IDV	SQV	RTV
Allmutant	72	77	78	78	78	78
Rule-based	19	19	18	20	16	19
RReliefF	33	27	22	27	28	26

6. EXPERIMENTS

6.1 Data source

In the experiments, genotype-phenotype data for 6 protease inhibitors were downloaded from Stanford HIV RT and Protease Sequence Database (http://hivdb.stanford.Edu/cgi-bin/PR_Phenotype.cgi) with the ViroLogic Susceptibility test method. The total cases and the cutoff value for each PI drug are shown in Table 2. The phenotypic results were assigned into one of two classes: susceptible or resistant according to the cutoff value of each PI. Each case of genotype-phenotype data was compared with HIV reference strain NL4-3 to create the full strain of HIV-1 protease gene as the input of the MRBF network.

Table 2. Total case and cutoff value for each PI drug

PI drug	LPV	APV	NFV	IDV	SQV	RTV
Number of cases	319	541	626	595	606	573
Cutoff value	10.0	2.0	2.5	2.1	1.7	2.5

6.2 Prediction by RBF and MRBF

For constructing the classifiers using the RBF network, each training example was represented as a center in the hidden layer and σ for each center was set to the same value. Thus the number of hidden nodes is equal to the number of total training examples. The target value is the logarithm of the IC50 fold change. We used the logarithm of IC50 fold change because the distributions of IC50 fold change are usually highly range because of a few highly resistant variants.

For constructing the classifiers using the MRBF network, we also set the value of σ as same as in the RBF network. In the preprocessing process of the MRBF network, we selected one of three feature selection approaches which has the highest predictive accuracy of the RBF network. Since the outputs of the MRBF network are real values representing the fold change, to evaluate the performance of the MRBF network and other algorithms, the outputs are classified into two classes: susceptible and resistant according to the logarithm of the cutoff values described in section 6.1 and used 10-fold cross-validation to assess the predictive accuracy.

6.3 Prediction by other algorithms

The predictive accuracy of the MRBF network was compared with the HIVdb system, REG, SVM, and the original RBF network. The phenotypic prediction by HIVdb was done through the HIVdb version 4.1.2 online system (http://hivdb6.stanford.edu/asi/deployed/HIVdb.html). For REG, the prediction of this technique was done through the statistics software SPSS version 12. For SVM classification, a linear kernel function was used to construct the classifier. In addition, RReliefF was used for selecting the relevant attributes in the preprocessing step of SVM. The performance of the prediction of REG, SVM, and the conventional RBF were measured by 10-fold cross-validation on the same datasets as those of the MRBF network.

7. RESULTS

In the learning step, the classifiers of the three models with different input attributes depending on each feature selection approach in Table 1 were constructed by the conventional RBF network. The predictive accuracy of each model is shown in Table 3.

Table 3. The predictive accuracy for three feature selection approaches

Drug	Allmutant	Rule-based	RReliefF
LPV	**88.70**	87.16	88.06
APV	**89.82**	**89.82**	88.34
NFV	92.97	**93.29**	**93.29**
IDV	91.58	92.60	**93.93**
SQV	88.93	89.25	**90.91**
RTV	92.32	93.89	**94.94**
Average	90.72	91.00	**91.58**

The result in Table 3 shows that most of the drugs, using RReliefF in the preprocessing step, have the highest accuracy, compared with the other feature selection approaches. In addition, RReliefF also has the highest average accuracy of 6 drugs. From this result, it confirmed that RReliefF efficiently selected attributes important to data classification. Thus, in the following experiments, we used the important attributes that were selected by RReliefF for the MRBF network and SVM.

Table 4. The comparison of MSE with various algorithms

Drug	REG	RBF+RReliefF	MRBF+RReliefF
LPV	0.148	0.189	**0.121**
APV	0.163	0.157	**0.116**
NFV	0.180	**0.110**	0.113
IDV	0.135	**0.092**	0.094
SQV	0.184	0.154	**0.144**
RTV	0.118	0.114	**0.105**
Average	0.155	0.136	**0.115**

The result in Table 4 shows that most of the drugs of MRBF have the smaller mean square error (MSE) than the RBF network. It indicates that MRBF improved the performance of the traditional RBF network in estimating the fold change values. Furthermore, MRBF provided the lower MSE than REG for all drugs.

Table 5. The comparison of the accuracy for various algorithms

Drug	HIVdb	REG	SVM+RReliefF	MRBF+RReliefF
LPV	73.98	83.35	88.09	**89.01**
APV	85.58	85.01	87.79	**88.16**
NFV	**94.25**	92.82	93.30	93.93
IDV	92.10	90.74	92.43	**93.77**
SQV	86.80	88.44	88.94	**90.75**
RTV	94.24	93.54	94.41	**95.46**
Average	87.83	88.98	90.83	**91.85**

As shown in Table 5, MRBF with RReliefF has the highest average accuracy when compared with HIVdb, REG, and SVM. Furthermore the predictive accuracy of MRBF also outperforms the others in 5 drugs expect for NFV.

8. CONCLUSION

This paper presents MRBF network, a new method adapted from the RBF network, to predict HIV-1 phenotypic resistance from genotypic data. The main idea of MRBF is to separately construct the RBF networks class by class to increase the ability of estimating the phenotypic value (FC). The MRBF network consists of three RBF networks: RBF-all, RBF-S, and RBF-R. In the first step of testing MRBF, an RBF-all network roughly estimates the FC. Then, FC from the RBF-all network is checked with two criteria. Finally, if any of the criteria is not satisfied, the RBF-S or RBF-R network is used to estimate more precise FC.

To enhance the performance of the classifier, we also used three different feature selection approaches for selecting the relevant attributes in the preprocessing step. Experimental results on the RBF network show that RReliefF gives the highest average accuracy for 6 drugs compared with other feature selection techniques. Then we used RReliefF in the preprocessing step of MRBF and SVM.

The results indicate that MRBF improves the ability of RBF in estimating the fold change values. In conclusion, MRBF has high ability in predicting HIV-1 drug resistance since it provides the highest predictive accuracy for 5 drugs except for NFV when compared with other techniques such as HIVdb, REG, and SVM.

ACKNOWLEDGEMENTS

This work was supported by National Center for Genetic Engineering and Biotechnology (BIOTEC), Thailand.

REFERENCES

1. J.M. Coffin, “HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy”, *Science*, vol. 267, 1995, pp. 483-489.
2. L. Demeter, R. Haubrich, “Phenotypic and genotypic resistance assays: methodology, reliability, and interpretations”, *Journal of Acquired Immune Deficiency Syndromes*, vol. 26, 2001, pp. S3-S9.
3. M. Robnik Sikonja and I. Kononenko, “An adaptation of relief for attribute estimation in regression”, *Machine Learning: Proceedings of the Fourteenth International Conference (ICML’97)*, 1997, pp. 296-304.
4. R.W. Shafer, D.R. Jung, B.J. Betts, “Human immunodeficiency virus type 1 reverse transcriptase and protease mutation search engine for queries”, *NAT Med*, vol. 6, 2000, pp. 1290-1292.
5. JL. Meynard, M. Vray, L. Morand-Joubert, et al, “Phenotypic or genotypic resistance testing for choosing antiretroviral therapy after treatment failure: a randomized trial”, *AIDS*, vol. 16, 2002, pp. 727-736.
6. K. Van Laethem, A. Ke Luca, A, Antinori, et al. “A genotypic drug resistance interpretation algorithm that significantly predicts therapy response in HIV-1 infected patients”, *Antiviral Ther*, vol. 7, 2002, pp. 123-129.
7. C. Reid, R. Bassett, S. Day, et al, “A dynamic rules-based interpretation system derived by an expert panel is predictive of virological failure”, *Antiviral Ther*, vol. 7, 2002, pp. s91.
8. K. Wang,, E. Jenwitheesuk, , R. Samudrala, J.E. Mitter, “Simple linear model provides highly accurate genotypic predictions of HIV-1 drug resistance”, *Antivir. Ther*, vol. 9, 2004, pp. 343-352.
9. N. Beerenwinkel, B. Schmidt, H. Walter, R. Kaiser, T. Lengauer, D. Hoffmann,K. Korn, and J. Selbig, “Geno2pheno: interpreting genotypic HIV drug resistance test”, *IEEE Intellig. Syst*, vol. 16, 2001, pp. 35-41.
10. N. Beerenwinkel, B. Schmidt, H. Walter, R. Kaiser, T. Lengauer, D. Hoffmann, K. Korn, and J. Selbig, “Diversity and complexity of HIV-1 drug resistance: a bioinformatics approach to prediction phenotype from genotype”, *Proceedings of Natl Acad. Sc*, USA, 2002, pp. 8271-8276.
11. N. Beerenwinkel, M. Daumer, M. Oette, K. Korn, D. Hoffmann, R. Kaiser, T. Lengauer, J. Selbig, and H. Walter, “Geno2Pheno: estimating phenotypic drug resistance from HIV-1 genotypes” , *Nucleic Acids Research*, vol. 31, 2003, pp. 3850-3855.
12. D. Wang and B. Larder, “Enhanced prediction of lopinavir resistance from genotype by use of artificial neural networks”, *Infectious Disease*, vol. 188, 2003, pp. 653-660.
13. S. Draghici and B. Potter, “Predicting HIV drug resistance with neural networks”, *Bioinformatics*, vol. 19, 2003, pp. 98-107.
14. M. Powell, “Radial basis function for multivariable interpolation: A review”, *Algorithms for approximation*, 1987, pp. 143-167.
15. D. S. Broomhead and D. Lowe, “Mutivariable functional interpolation and adaptive networks”, *Complex System 2,* 1988, pp. 321-355.
16. J. Moody and C. J. Darken, “Fast learning in networks of locally-tuned processing units”, *Neural Computation*, 1(2), 1989, pp. 281-294.
17. S. Haykin, *Neural Networks: A comprehensive foundation*, Prentice Hall, New Jersey, 1999, pp. 256-312.
18. S. W. Robert, “Genotypic Testing for Human Immunodeficiency Virus Type 1 Drug Resistance”, *Clinical Microbiology Regviews*, vol. 15, 2002, pp. 247-277.