ACSESS proceedings
2009

Data Mining Analysis of HIV-1 Protease Crystal Structures

Gene M. Ko, A. Srinivas Reddy, Sunil Kumar, and Rajni Garg

AP0907

CSRC an interdisciplinary gateway to the future
SAN DIEGO STATE UNIVERSITY

Data Mining Analysis of HIV-1 Protease Crystal Structures

Gene M. Ko[1], A. Srinivas Reddy[2], Sunil Kumar[2], and Rajni Garg[*,1]
[1]Computational Science Research Center, [2]Electrical and Computer Engineering
San Diego State University, San Diego, CA 92182
*Corresponding author: 5500 Campanile Drive, San Diego, CA 92182-1245, rgarg@mail.sdsu.edu

Abstract: A data mining study has been done using HIV-1 protease crystal structures complexed with FDA approved HIV-1 protease inhibitor drugs. Chemical descriptors have been computed for the binding pockets of each crystal structure, yielding approximately 600 constitutional, topological, geometric, elecrotostatic, and quantum mechanical descriptors for each structure. Several supervised (hybrid binary particle swarm optimization-artificial neural network and random forest) and unsupervised learning (locally linear embedding) techniques have been explored for feature selection to determine a quantitative structure-activity relationship (QSAR) model containing the most relevant descriptors needed to cluster each crystal structure according to their bound ligand. This method of computational modeling and screening process would aid in the understanding of the effect HIV mutations have on the binding affinity of various present and future HIV-1 protease inhibitors due to structural changes arising from the mutations.

Keywords: chemical descriptor selection, data mining, QSAR.

1. Introduction

An estimated 33.2 million people are living with the Human Immunodeficiency Virus (HIV). HIV is a retrovirus that can lead to Acquired Immune Deficiency Syndrome (AIDS) [1]. The genetic variation occurring due to its replicatory dynamics within the human host leads to a high rate of mutation which results in drug resistances, frustrating the anti-HIV drug development process. To aid in the understanding of mutations within HIV and their effects on the binding structure of the virus when complexed to an anti-HIV drug, hundreds of crystal structures of various HIV proteins have been developed.

The Protein Data Bank (PDB) houses experimentally derived crystal structures of biological macromolecules [2]. We focused our study on the crystal structures of HIV-1 protease complexed with one of the nine FDA approved protease inhibitors (Darunavir, Nelfinavir, Amprenavir, Lopinavir, Atazanavir, Indinavir, Ritonavir, Saquinavir, and Tipranavir). To develop a quantitative understanding of the mutational effects on the binding nature of HIV, chemical descriptors of the binding pockets on each of the crystal structures have been computed.

As there are many properties available from the descriptor calculation of the binding pockets, it is important to screen them for the best set of properties in determining its action against the drug. Several feature selection techniques have been identified: a hybrid binary particle swarm optimization-artificial neural network scheme, locally linear embedding, and random forest. It is hypothesized that the optimal set of descriptors will lead to a proper clustering of the crystal structures grouped according to their complexed ligand. These optimal sets of descriptors can then be used to build a QSAR model that best describes the shape of the binding pocket based on mutations and inhibitor binding.

2. Methods

A flowchart of the methodology is presented in Figure 1. A dataset is generated using crystal structures deposited in the PDB. Chemical descriptors are generated from these crystal structures. A subset of descriptors are selected using various feature selection techniques which are then used in clustering to validate whether the selected descriptors best correlates the shape of the protein structure with its complexed ligand.

2.1 Dataset Creation

The PDB was searched for all HIV-1 protease crystal structures complexed with FDA approved protease inhibitors. A total of 57 protein crystal structures were found to be HIV-1 complexed with one of the nine FDA approved

HIV-1 protease inhibitors. A list of the protease inhibitors are presented in Table 1.

The binding pocket of each crystal structure was identified as any amino acid within 6Å of the bound ligand. The binding pocket was extracted using the DeepView Swiss-PdbViewer [3]. If multiple ligands were found to be complexed in the structure, then multiple binding pockets will be extracted from the structure. Of the 57 crystal structures, a total of 62 binding pocket structures have been extracted. To compute the chemical descriptors of the binding pocket, calculations of the molecular electronic structure must be computed. The Austin Model 1 (AM1) energy calculation of each pocket structure was computed using Semichem's AMPAC software [4]. The chemical descriptors were then computed using Semichem's Codessa software [4]. Approximately 600 constitutional, topological, geometric, electrostatic, and quantum mechanical descriptors were generated from the AM1 energy calculations. To reduce the number of descriptors, we eliminated any descriptors with samples that contained no values and any descriptors that contained a single constant value in all samples. This resulted in a total of 456 descriptors in the dataset. All descriptor values were rescaled to have a zero mean and a standard deviation of one.

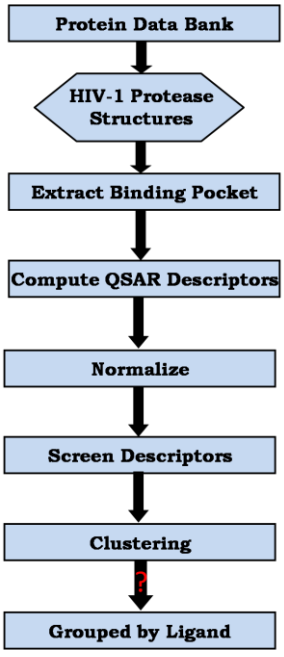

Figure 1. Flowchart of methodology sequence.

Table 1: Number of HIV-1 protease crystal structures in the PDB complexed with one of the FDA approved protease inhibitors.

PDB Ligand ID	Ligand Name	Structures
017	Darunavir	11
1UN	Nelfinavir	6
478	Amprenavir	2
AB1	Lopinavir	4
DR7	Atazanavir	4
MK1	Indinavir	13
RIT	Ritonavir	5
ROC	Saquinavir	7
TPV	Tipranavir	5

The supervised learning techniques must be trained using classification or regression values. For classification, the target values are the ligands with which each crystal structure was complexed to. For regression, we used the binding affinity (K_i) of the crystal structures as the target values. Twenty five crystal structures were found to have their binding affinity values reported in the Binding MOAD database [5].

2.2 Descriptor Screening Techniques

Binary particle swarm optimization-artificial neural network. The binary particle swarm optimization-artificial neural network (BPSO-ANN) algorithm is a hybrid technique involving the use of particle swarms to search the descriptor space to select for the optimal set of descriptors and utilizes an artificial neural network (ANN) as the objective function. We implemented the BPSO-ANN algorithm described by Agrafiotis and Cedeño to discretize the particle swarm process to do a binary selection of descriptors, a technique which computes the probabilities of selecting a descriptor [6]. The roulette wheel selection was used to select a subset of 10 descriptors to be trained by the ANN. For each descriptor subset, two different ANN were trained, using the protease inhibitors as the classification targets and the inhibitor values for the regression targets. To avoid being trapped in a local minimum, each ANN was trained three times with the lowest training error model retained.

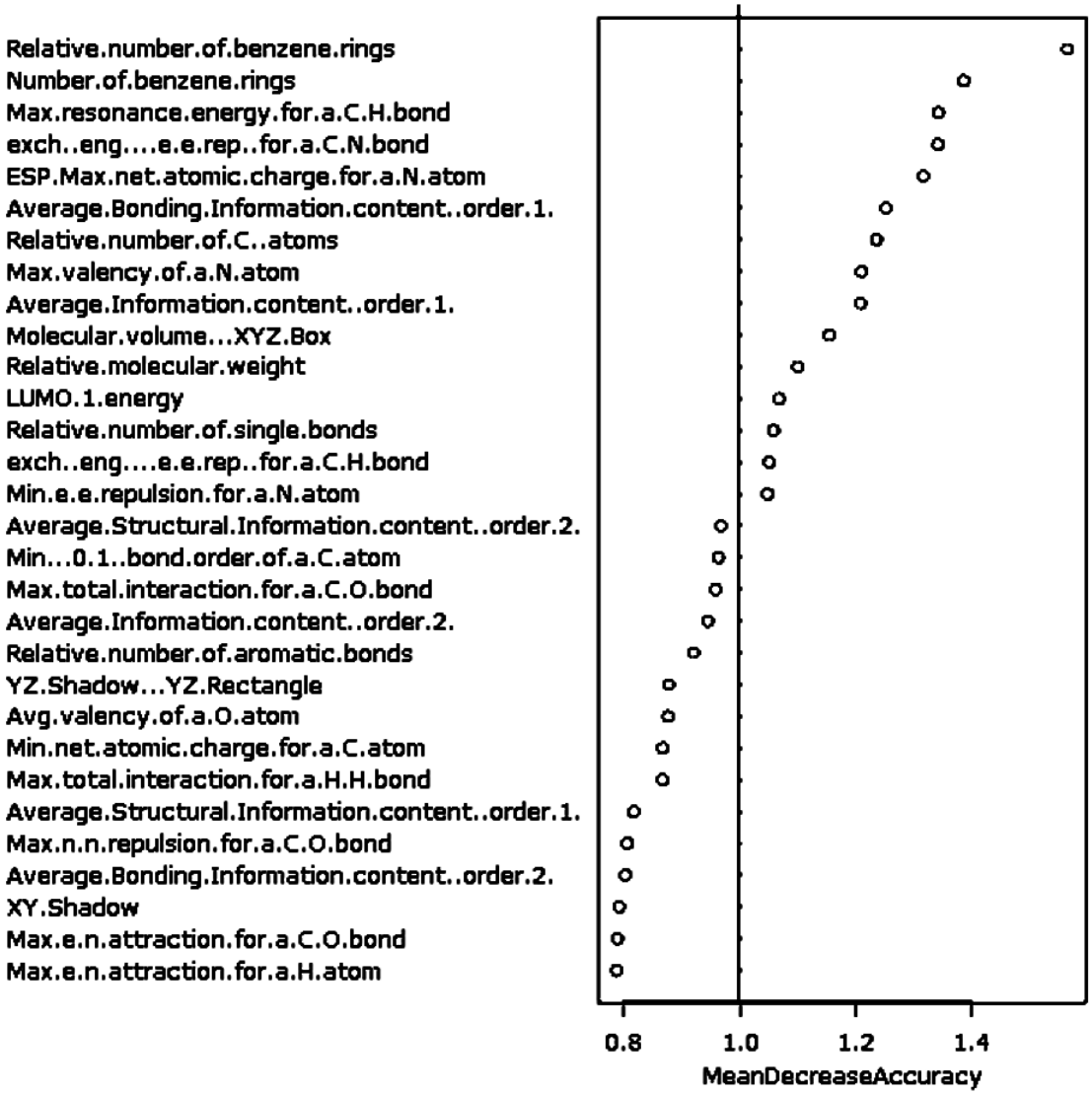

Figure 2. Variable importance plot of the random forest classifier.

The descriptors determined by the neural network model with the lowest training error is then considered for clustering.

Random Forest. Random forest is a supervised tree ensemble machine learning classifier developed by Leo Brieman and Adele Cutler [7]. In addition to being a classifier, random forest includes a measure of descriptor importance in determining classification or regression called the variable importance measure. A random forest model was trained on the dataset to classify each of the binding pockets to their complexed protease inhibitors. A variable importance measurement was computed during the classifier training. The top ranked group of descriptors are then considered for clustering.

Locally Linear Embedding. Locally linear embedding (LLE) is an unsupervised learning technique used to create a non-linear dimensionality reduction representation of a large dataset [8]. LLE maps the high dimension data points X_i to low dimensional vectors Y_i. In the standard LLE algorithm, the Euclidean distance between each data point X_i is computed.

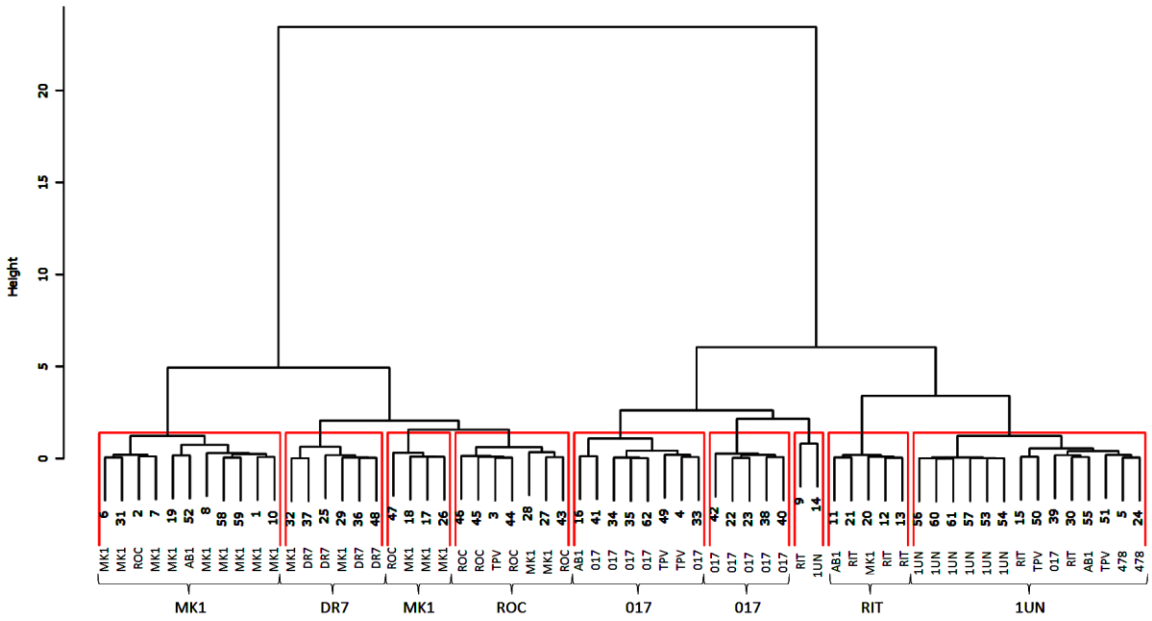

Figure 3. Hierarchical clustering of the 62 binding pockets with the top 15 descriptors determined by random forest.

A linear fitting is performed to compute the weight matrix W that best reconstructs each data point X_i by minimizing the cost function:

$$\varepsilon(W) = \sum_i \left| X_i - \sum_j W_{ij} X_j \right|^2$$

The low dimension mapping vectors Y_i are constructed to d dimensions by using the weight matrix W to minimizing the embedding cost function $\Phi(Y)$ using the bottom d non-zero eigenvectors:

$$\Phi(Y) = \sum_i \left| Y_i - \sum_j W_{ij} Y_j \right|^2$$

LLE was used to map the 456 dimension descriptor dataset to 10 dimensions.

The most suitable set of descriptors selected by the different screening approaches are used by hierarchical clustering to verify that the selected descriptors best correlate the descriptive features of the binding pockets of the HIV-1 protease crystal structures with their bound ligand.

2.3 Clustering

To validate the list of important descriptors, hierarchical clustering of the dataset with the feature selected subsets of descriptors is performed. It is expected that a good set of descriptors would lead to each of the binding pockets clustered together according to their complexed ligands.

Two distance metrics were explored, Euclidean and Pearson distances. These two distances are used by the Ward agglomerative method to build the hierarchical tree. Nine clusters are considered and examined to determine whether the binding pockets are grouped according to ligand.

3. Results and Discussion

We have evaluated three data mining techniques that best assess an optimal set of QSAR descriptors to quantitatively correlate the binding pocket of HIV-1 protease with their bound ligand. Both the BPSO-ANN and LLE techniques produced hierarchical trees which appeared to be clustered randomly. The LLE technique is not suitable for this study as we were unable to determine which chemical descriptors were selected for the low dimensionality mapping. Based on the work of L'Heureux et al., LLE is best used as a preprocessing technique for dimensionality reduction prior to building a model which utilizes large data [9]. Poor clustering results in

the BPSO-ANN and LLE method can be attributed to the small dataset size. ANN do not train well with small training sample sizes and LLE does not appear to capture the non-linearity of the chemical data due to the small sample sizes of each complexed ligand in our dataset.

Initial results with random forest appear satisfactory. The top 15 descriptors shown in the variable importance plot in Figure 2 were considered. Using the Pearson distance measurement with the Ward agglomerative hierarchical tree producing algorithm as shown in Figure 3, we were able to obtain some meaningful clusters. The initial tree was built using the parameters $n_{Tree} = 1000$ and $m_{try} = 21$. We plan to continue our study with random forest by optimizing the parameters n_{Tree} and m_{try} to minimize the classifier training error which should produce the most optimal set of descriptors for clustering. These descriptors will then be used to build a QSAR model which will quantitatively describe the shape of the binding pocket with its affinity for ligand binding.

5. References

1. UNAIDS, 2008 *Report on the Global AIDS Epidemic*, 32. UNAIDS, Switzerland (2008)
2. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., The Protein Data Bank. *Nucleic Acids Research*, **28**, 235-242 (2000)
3. Guex, N. and Peitsch, M.C., SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling, *Electrophoresis*, **18**, 2174-2723 (1997)
4. http://www.semichem.com/
5. Hu, L., Benson, M.L., Smith, R.D., Lerner, M.G., Carlson, H.A., Binding MOAD (Mother Of All Databases), *Proteins*, **60**, 333-340 (2005)
6. Agrafiotis, D.K. and Cedeño, W., Feature Selection for Structure-Activity Correlation Using Binary Particle Swarms, *Journal of Medicinal Chemistry*, **45**, 1098-1107 (2002)
7. Breiman L., Random forests, *Machine Learning*, **45**, 5-32 (2001)
8. Saul, L.K. and Roweis, S.T., An Introduction to Locally Linear Embedding, (2001). Available from http://www.cs.toronto.edu/~roweis/lle/
9. L'heureux, P.J., Carreau, J., Bengio, Y., Delalleau, O., Yue, S.Y., Locally Linear Embedding for dimensionality reduction in QSAR, *Journal of Computer-Aided Molecular Design*, **18**, 475-482 (2004)

6. Acknowledgements

G.M. Ko was supported as a trainee by the NIH RoadMap Initiative award T90 DK07015.