ACSESS proceedings

2010

A Random Forest Model for the Analysis of Chemical Descriptors for the Elucidation of HIV-1 Protease Protein-Ligand Interactions

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

AP10-02

SAN DIEGO STATE UNIVERSITY

A Random Forest Model for the Analysis of Chemical Descriptors for the Elucidation of HIV-1 Protease Protein-Ligand Interactions

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

Abstract: A model for the classification of 70 HIV-1 protease crystal structure binding pockets to one of its complexed FDA approved protease inhibitors utilizing Random Forest has been developed. 456 chemical descriptors of the binding pocket of each crystal structure have been computed and are used to develop the classification model. Simulations were performed to determine the optimal Random Forest model parameters. An implicit feature relevance measure for the optimal model was analyzed using the Gini importance measure. The chemical descriptors most influential in classifying the binding pockets of HIV-1 protease to its complexed protease inhibitor were analyzed and interpreted in terms of the binding pocket structure and their protein-ligand interactions. The selected descriptors by the Random Forest model provides insight on the structure of HIV-1 protease which can be used to drive the drug discovery process to design novel HIV-1 protease inhibitors.

Keywords: Random Forest, classification, HIV-1 Protease, crystal structures

1. Introduction

HIV-1 protease is a popular target in the design of new anti-HIV drugs called HIV-1 protease inhibitors, which are designed to inhibit the activity of HIV-1 protease to prevent cleavage of nascent protein polypeptides into active viruses [1]. To aid in the development of new pharmaceutical drugs, hundreds of crystal structures of HIV-1 protease complexed with potential drug candidates have been created and analyzed for their physiochemical properties. Many of these crystal structures are deposited in the Protein Data Bank (PDB) [2], a crystal structure repository database for large biological macromolecules.

Chemical and structural descriptors can be computed from the crystal structures of proteins based on their atomic makeup which are used for the development of QSAR/QSPR (quantitative structure-activity/structure-property relationship) models to establish a mathematical relationship between the molecular structure and chemical properties for the classification of molecules or prediction of biological activity [3]. The number of descriptors can range in the hundreds and usually exceeds the number of samples available. In a chemometrics study by Svetnik et al., Random Forest was found to be one of the top classifiers, being able to handle high dimensional data and ignoring redundant and irrelevant descriptors [4].

In the present study, we investigate the relationship between the HIV-1 protease binding pockets in terms of their chemical descriptors to its complexed protease inhibitors. We focus our study on the HIV-1 protease crystal structures which are complexed to one of the nine FDA approved HIV-1 protease inhibitors: Darunavir (017), Nelfinavir (1UN), Amprenavir (478), Lopinavir (AB1), Atazanavir (DR7), Indinavir (MK1), Ritonavir (RIT), Saquinavir (ROC), and Tipranavir (TPV). These crystal structures were created to study the binding mechanism of the various protease inhibitors with both the wild type and commonly accepted drug resistant HIV-1 protease proteins. Sequence analysis of these crystal structures reveals a wide variation of the amino acid sequences and is representative of both the wild type and commonly accepted drug resistances for each complexed ligand. The binding pocket shape is directly influenced by the ligand complexed to the structure [5], thus it is crucial to match the ligand with its pocket environment. The chemical descriptors of the binding pockets of each crystal structure were computed. We evaluated the use of Random Forest for classifying HIV-1 protease crystal structures. This classification model will provide an understanding of the relationship between the most relevant quantitative chemical descriptors to the conformation of HIV-1 protease caused by the mutations present in the binding pockets and their complexed protease inhibitor.

2. Methods

2.1 Dataset

The PDB was searched for all HIV-1 protease crystal structures complexed with one of the nine FDA approved protease inhibitors. The binding pocket of each crystal structure was identified as any amino acid within a 6 angstrom radius of the complexed ligand. A total of 70 binding pockets were extracted from the crystal structures. The binding pockets were extracted using PyMOL (Figure 1) [6].

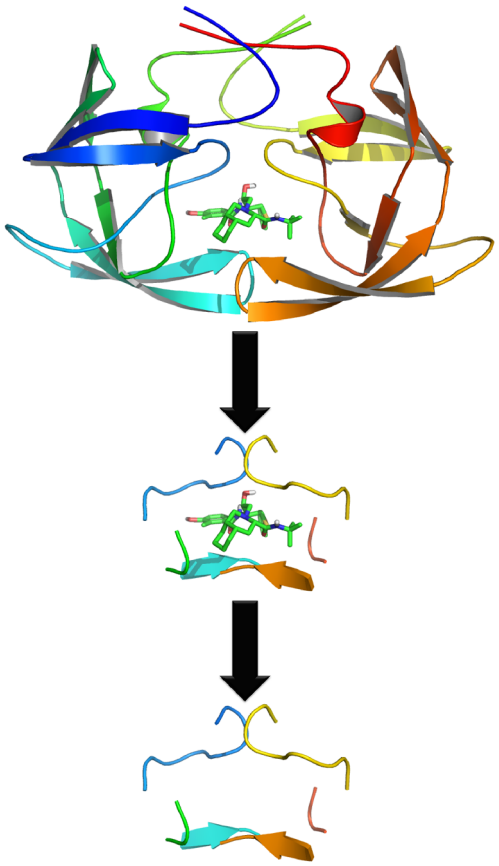


Figure 1. Extraction of binding pocket. All amino acids within a 6 angstrom radius is considered to form the binding pocket and is extracted from the HIV-1 protease crystal structure. The complexed ligand is subsequently removed.

2.2 Descriptor Calculations

To compute the quantum-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 using the atomic coordinates of the crystal structure was computed using AMPAC [7]. The chemical descriptors were then computed using Codessa [8]. A set of 562 constitutional, geometrical, electrostatic, topological, and quantum-chemical descriptors were derived from the molecular structure and AM1 energy calculations of each of the binding pocket structures. Constitutional descriptors describe the non-geometric molecular composition of the structure while geometrical descriptors describe the 3D representation of the molecule. Topological descriptors use graph theory to describe the atomic connectivity of a molecule. Electrostatic descriptors describe the charge distribution of the molecule. Quantum-chemical descriptors use quantum mechanical theory to describe a molecule's electronic and geometrical properties and their atomic interactions.

To reduce the descriptor space, we eliminated any descriptors with null or constant values across the majority of the samples. Null values occur because the descriptor is specific for atoms which are not present in the structure. This resulted in a total of 456 descriptors in the dataset. All descriptor values were recentered to have a zero mean and a standard deviation of one.

2.3 Random Forest

Random Forest as a classification method is a classification tree based ensemble learning technique which consists of a collection of unpruned classification trees used collectively to determine the output class for a given observation [9]. Ensemble learners utilize multiple models in combination which may result in an improved predictive model. A Random Forest classification model is a collection of classification tree predictors

$$\{h(\bar{x}, \Theta_k), k = 1, 2, \ldots, T\}$$

where Θ_k are independent identically distributed random vectors which each cast a vote for a class for a given input vector $\bar{x}$ [9]. Each of the classification tree models are grown fully without pruning as to keep bias at a minimum. In the tree growing steps of Random Forest, a small random sampling of the variables are considered for each nodal split. The Gini

measure of impurity is used to determine the variable selected to make the nodal split. The Gini impurity measure at node t is defined as

$$g(t) = \sum_{j \neq i} p(j \mid t)\, p(i \mid t)$$

where i and j are the categories for the variable. The subsequent Gini criterion for determining a split with variable s at node t is defined as

$$\Phi(s,t) = g(t) - p_L g(t_L) - p_R g(t_R)$$

where p_L and p_R are the proportion of observations in t in the left and right child nodes respectively. The variable s which maximizes $\Phi(s,t)$ is selected for the nodal split.

In the statistical computing environment R [10], there are two major parameters used to train the Random Forest classifier model: n_{Tree}, the number of classification trees to train in the forest classifier, and m_{try}, the number of variables to randomly consider at each node of each tree. As each classification tree is built, an estimate of the Random Forest classifier performance is measured, called the Out-of-Bag (OOB) error. The OOB error measures the classification error over all of the trained classification trees in the Random Forest model.

Random Forest includes an implicit measure of variable importance when determining classification which is obtained by the Gini importance measurement [11]. The Gini importance measures the improvement of each variable in the Gini criterion used to split the classification tree nodes.

Random Forest is used to classify each of the 70 HIV-1 protease binding pockets to one of the nine FDA approved HIV-1 protease. For the determination of the optimal tree size with the lowest OOB error, a Random Forest classifier is trained with parameters n_{Tree} = 20000 and the default parameter of m_{try}. The default parameter of m_{try} in a Random Forest classifier is equal to the square root of the number of descriptors available. In R, as each tree is generated in a single Random Forest model, the OOB error is computed, enabling the determination of the optimal tree size. 40,000 Random Forest models were generated from which the average OOB error is determined at each tree size. The optimal value of m_{try} for the optimal n_{Tree} value is then determined by the smallest average OOB error from a simulation of 10 Random Forest models at each m_{try} value from 1 to 456.

A final Random Forest model is generated with the optimal n_{Tree} and m_{try} parameters. A list of variables deemed to be the most important set of chemical descriptors in building the classifier is determined by the Gini criteria.

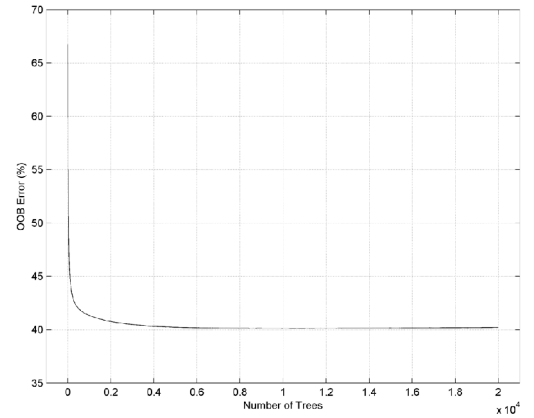
4. Results and Discussion

4.1 Random Forest

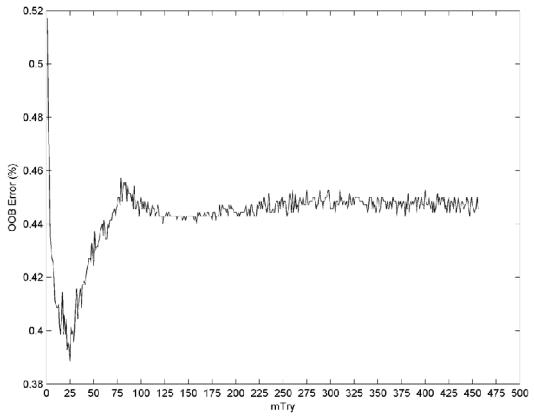
From the simulation of 40,000 Random Forest models, each of the classification trees were built using a default parameter of $m_{try} = \sqrt{456} = 21$. The optimal tree size was determined to be at n_{Tree} = 10586 with an average OOB error of 40.113% (Figure 2a). Next, the optimal m_{try} parameter was determined. Ten Random Forest models were generated using n_{Tree} = 10586 at each varying value of m_{try} from 1 to 456. It was observed that the optimal m_{try} value with the minimal average OOB error occurs near the default value of m_{try} (Figure 2b). This observation was consistent with the observations in the Random Forest simulations by Svetnik et al. [4].

We also observed that the top group of descriptors remain generally the same with varying values of m_{try} near the default value, but varies slightly in its order, which confirms that Random Forest is relatively insensitive to the value of m_{try} except at the extremes. In addition, although the OOB error converges relatively quickly, we observed it was less likely that the top ranked descriptors will deviate when more classification trees were introduced. This indicates a large tree size helps to stabilize the ranking of the top group of descriptors that best influence the classification ability of the Random Forest model.

A final Random Forest model was generated using the parameters n_{Tree} = 10586 and m_{try} = 21. The list of the most important descriptors determined by the Gini importance measurement of this model is shown in Figure 3. Based on the natural break in the elbow curve of the Gini importance plot, the top 12 descriptors are analyzed for their chemical significance of the structure of HIV-1 protease binding pocket and the protein-ligand interactions due to the atomic interactions described by the quantum-chemical descriptors.



(a)



(b)

Figure 2. Random Forest simulation results. (a) The average Out-of-Bag (OOB) error for each tree size in a simulation of 20,000 trees run 40,000 times for the determination of the optimal tree size n_{Tree}. The optimal tree size is n_{Tree} = 10586 with an average OOB error of 40.113%. (b) The average OOB error for each value of m_{try} in a simulation of 10586 classification trees run 10 times for the optimal m_{try} determination. The optimal value of m_{try} with the smallest OOB error occurs near the default parameter of m_{try} = 21.

4.2 Descriptor Interpretations

The top 12 ranked descriptors obtained by the Random Forest variable importance method (Figure 3) are exchange energy + electron-electron repulsion for a C-N bond, max resonance energy for a C-C bond, molecular volume/XYZ box, min >0.1 bond order of a C atom, max total interaction for a C-C bond, relative number of benzene rings, average information content (order 1), exchange energy + electron-electron repulsion for a C-C bond, maximum electron-nuclear attraction for a C-N bond, relative number of C atoms, YZ Shadow/YZ rectangle, and the number of benzene rings.

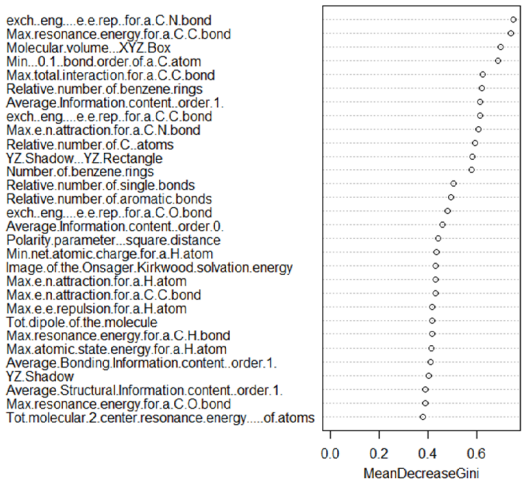


Figure 3. Variable importance measure of the optimal Random Forest classification model using the Gini importance measure. Due to the natural break in the curve, the top 12 descriptors determined by the Gini importance have been selected for chemical interpretation.

The HIV-1 protease binding site is known to be hydrophobic in nature and thus protease inhibitors with hydrophobic side chains have a higher binding affinity [12]. The benzene rings descriptors reflect upon the hydrophobic amino acid residues which contain a benzene ring. The wild type sequence of HIV-1 protease does not contain any residues which contain a benzene ring, and thus mutations resulting in a residue which does contain one may have an effect on the shape of the binding pocket in terms of steric interactions.

The physical shape of the binding pocket is emphasized by the geometrical descriptors YZ Shadow/YZ rectangle and molecular volume/XYZ box. The atomic connectivity of the binding pocket is described by the topological descriptor average information content. Electrostatic energy between the binding pocket and the ligand is revealed by the exchange energy and electron-electron repulsion for C-N and C-C bonds descriptors. The

importance of benzene rings is signified by the C-C bond resonance energy.

5. Conclusions

In this study, we have used Random Forest to build an appropriate classification model for predicting the HIV-1 protease binding pocket structures to its complexed HIV-1 protease inhibitor. We have interpreted the descriptors selected by the Random Forest variable importance measure.

The top ranked descriptors reflected the geometric shape and atomic makeup of the binding site. The quantum-chemical descriptors reflected the energy exchange between the binding pocket and ligand as a result of London-van der Waals interactions in the protein-ligand binding process, specifically between the C-C and C-N atoms. These descriptors provide a means of quantifying the geometric and electronic properties of the HIV-1 protease binding pocket which can be used to design novel HIV-1 protease inhibitors.

6. References

1. E. De Clercq, "Anti-HIV drugs: 25 compounds approved within 25 years after the discovery of HIV," *International Journal of Antimicrobial Agents*, **33**, pp. 307-320 (2009).
2. H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne, "The Protein Data Bank," *Nucleic Acids Research*, **28**, pp. 235-242 (2000).
3. D.A. Winkler, "The role of quantitative structure - activity relationships (QSAR) in biomolecular discovery," *Briefings in Bioinformatics*, **3**, pp. 73-86 (2002).
4. V. Svetnik, A. Liaw, C. Tong, J.C. Culberson, R.P. Sheridan, and B.P. Feuston, "Random Forest: a classification and regression tool for compound classification and QSAR modeling," *Journal of Chemical Information and Computer Sciences*, **43**, pp. 1947-1958 (2003).
5. F. Clavel and A.J. Hance, "HIV drug resistance," The New England Journal of Medicine, **350**, pp. 1023-1035 (2004).
6. The PyMOL Molecular Graphics System version 1.1, Palo Alto, CA: DeLano Scientific LLC, 2009.
7. AMPAC version 8.16.8. Shawnee, Kansas: Semichem, Inc., 2007.
8. Codessa version 2.7.10. Shawnee, Kansas: Semichem, Inc., 2007.
9. L. Breiman, "Random Forests," *Machine Learning*, **45**, pp. 5-32 (2001).
10. The R Project for Statistical Computing version 2.8.1.
11. K.J. Archer and R.V. Kimes, "Empirical characterization of random forest variable importance measures," *Computational Statistics and Data Analysis*, **52**, pp. 2249-2260 (2008).
12. M.D. Kelly and R.L. Mancera, "A new method for estimating the importance of hydrophobic groups in the binding site of a protein," *Journal of Medicinal Chemistry*, **48**, pp. 1069-1078 (2005).

7. Acknowledgements

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