International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 22 (2017) pp. 12060-12066
© Research India Publications. http://www.ripublication.com

Ontological Representation of Data Sampling in HIV1 Reverse Transcriptase

R.Geetha
Research Scholar, Bharath University,
173, New Agaram Rd, Selaiyur, Chennai, Tamil Nadu 600073 India
Orcid Id: 0000-0003-4631-2475

S.Sivasubramanian
Principal, Dhanlakshmi College of Engineering,
Dr. V. P. R Nagar, Off. Tambaram Sriperumbudur Road (via Manimangalam)
Manimangalam Post, Chennai, Tamil Nadu - 601 301

Abstract

Acquired immuno deficiency syndrome (AIDS) is a lifetime aggressive disease of the human immune system shaped by human immune deficiency virus (HIV). A supervised knowledge attitude to support the selection of HIV literature and technologically advanced an overseen knowledge method to support selection responsibilities, by spontaneously abating hypothetically appropriate documents from a list reclaimed by a literature catalogue hunt. Limited non-nucleoside converse transcriptase inhibitors (NNRTIs) need remained appropriate by the United States Food and Drug Administration (US FDA) as drugs for AIDS is presented with its ontological model. To overwhelm the foremost disputes related with the spontaneous literature selection assignment, we assessed the usage of data sampler, feature permutations, and feature choice approaches, generating an aggregate of 105 making ready representations. In demand to enrich therapeutic selections in contradiction of AIDS we inspected innovative herbal compounds of 4-thiazolidinone and its end product that are recognized to need significant antiviral strength and while inspecting the data is to be sampled using corpus and data sampling metrics. This data sampling and corpus are represented ontologically for HIV using various tools. An ontology HIVSamp has been developed to meet the sampling in HIV data which can be used by the future generation for research.

Keywords: HIV, Docking, Simulation, Sampling

INTRODUCTION

The molecular docking and simulation experiments need acknowledged one such herbal molecule recognized as (5E)-3-(2-aminoethyl)-5-benzylidene-1, 3-thiazolidine-2,4-dione that might predicament HIV-1RT through extraordinary attraction to origin noncompetitive reticence. Consequences are furthermore equated through other US FDA appropriate drugs. Long de novo simulations and docking education recommend that the ligand (5E)-3-(2aminoethyl)-5-benzylidene-1, 3-thiazolidine-2,4-dione (CID: 1656714) has sturdy binding interfaces through Asp113, Asp110, Asp185 and Asp186 amino acids, altogether of which be appropriate to one or the other catalytic pockets of HIV-1RT. It is predictable that these interfaces might be hazardous in the inhibitory activity of the HIV-1RT. Consequently, this learning offers an indication for attention of (5E)-3-(2-aminoethyl)-5-benzylidene-1, 3thiazolidine-2,4-dione as an appreciated natural molecule in the behavior and avoidance of HIV- linked sicknesses. The representations yielding the finest consequences remained collected of a Logistic Model Trees classifier, a impartially serviceable exercise set, and feature permutation of Bag-Of-Words and MeSH terms. According to our consequences, the organization appropriately tags the great mainstream of appropriate documents, creating it serviceable to support HIV methodical evaluations to permit investigators to evaluate a superior quantity of documents in fewer periods. OPEN literature depositories are regularly the foremost cause of information used by scientific researchers. Life science and biomedical databases comprise a huge quantity of documents, and remain quickly increasing imitating the stride of scientific periodicals and informal right to use online depositories. The selection of scientific fiction is characteristically accomplished by researchers to classify right trainings for a certain focus and sustenance systematic evaluations. Querying the PubMed[1] database with the string HIV, retrieves over 295k documents, while the query AIDS brings more than 238k documents. A detailed description of the systematic review workflow is given in [2]. The initial step of the process is to define the research problem, and to search for eligible literature by querying scientific databases. The next step is to select studies, a task that requires exhaustive screening of a document list that was retrieved by strategic searches made by researchers. The evaluation of biomedical data is highly relevant to assist the information discovery process in biomedical research (e.g. [3], [4]). Machine learning approaches have been applied to support systematic reviews by performing literature screening (e.g. [5], [6]). Substantial efforts are put into extracting and annotating information on life science related documents [7], [8], with the use of natural language processing approaches [9]. The BioCreative (http://www.biocreative.org/) creative [10] signifies the current extensive effort in the study of approaches to perform biomedical text classification.

Automatic classification of bioliterature was specially evaluated at some of the BioCreative challenges [11], [12]. This project unambiguously addresses the problem bioliterature text classification, and the exact task challenges, by designing a problem-oriented supervised learning model. One of the main issues is the underlying distribution of the data. Given a selected list of documents retrieved by a query search, researchers usually label most of them as excluded, and only a small portion is selected as relevant, and labeled as included. A dataset is considered imbalanced when the difference between the numbers of documents belonging to each class is so severe that it interferes in the machine learning process [13]. Classification algorithms remains apt to make the best of the overall accuracy, therefore favoring the most frequent class while overlooking the least represented class in a document collection [14]. Feature selection methods [15] are strategies applied in classification models to identify a subset of feature that most suits a given task. Feature selection reduces the size of the feature space by keeping only the most relevant features for a specific problem. In this work, we investigated the use of imbalanced learning strategies and feature selection methods applied to text classification with the goal of supporting HIV literature screening.

AIDS is tote up as one of the furthermost severe endemic civic fitness challenges worldwide through undesirable influence, first renowned in United States in 1981 so-called as human immunodeficiency virus (HIV). Approximately 25 million individuals ought to die owing to this infection [16]. The reticence of HIV-1 RT is the significant participant in the contamination cycle of HIV[17] for the treatment of AIDS, since this enzyme [18], [19], [20] is accountable for transformation of the single-stranded viral RNA genome into double-stranded DNA that acquires consequently unified into the host genome [21]. A small number of HIV-1 RT inhibitors, such as, Efavirenz, Nevirapine and Rilpivirine are existing aimed at the suggestive action with aggressive treatment [22]. The 4-thiazolidinone the fourth location derivative show healthier pharmacological activities, such as, antitubercular [23], anti-inflammatory [24], pesticidal [25], anticonvulsant [26], and antimicrobial [27]. The carbonyl group of 4thiazolidinone is nonreactive. The occurrence of N-C-S linkage in 4-thiazolidinone has been revealed to ensure anti-cancer actions [28] and methylene carbon atom at position 5 retains nucleophilic action [29]. The docking study delivered distinct fixed outlook about protein and lignd interfaces, whereas MD (Molecular Dynamics) simulations mark available the range of indications and conformational revisions in the protein domains, catalytic sites and essential

significant improvement in thoughtful ligand interfaces to proteins, and protein relations [30], [31], [32]. (5E)-3-(2-aminoethyl)-5-benzylidene-1, 3-thiazolidine- 2, 4- dione (CID: 1656714), a derivative of 4-thiazolidinone, through means of a probable herbal drug candidate predominantly in contrast to HIV-1 RT. The acquired consequences must be remained associated by that of the US FDA accepted drugs for AIDS treatment and inhibition.

METHODOLOGY

HIV-1 Reverse Transcriptase

In HIV-1 RT, p66 polymerase domain has a right-handed conformation and thumb subdomain residues (244-322) plus liking subdomain residues (323-427). All these domains outline a primer binding cleft with polymerase active site residues (Asp110, Asp185 and Asp186) located in b6-b10-b9 sheet of palm subdomain [33], [34]. The polymerase active site is carboxylated in p66 palm subdomain that binds with two magnesium divalent ions (Mg2þ) for cataltysis and hence, helps in the addition of nucleotide to the growing primer strand known as “primer grip” for correct positioning of 30 end [35], [36]. The binding pocket of non-nucleosides for inhibition is formed primarily by b5-b6 of loop, b6, b9-b10, b12-b13 of hairpin, b15 of p66 and b7-b8 connecting loop of p51. This pocket is also occupied by Gly231, Trp266, Tyr188 and Trp229 side chain residues [36].

Selection of Protein

HIV-1 RT lacking ligand is obtained from RCSB Protein Data Bank Preparation and Binding Sites Prediction PDB structure. The best feasible three-dimensional structure of HIV-1 RT was reached using Swiss-PdbViewer v4.1 [37]. Lipinski’s rule of five is used by the compunds for screening therapeutic properties. The best 3D pharmacophore model is made to intersect with Best Fit’ selection. The molecular position of HIV-1 RT is organized using Electrostatic and structural properties. Using docking and simulation water molecules, ions, co factors, charges, energy minimization and MD simulation are detached. Binding Site analysis module confirms the binding between cavities and protein ligand binding.

Docking

4-thiazolidinone and its derivatives occurred is retrieved from Chembank and NCBIPubChem database. In order to draw the structure of these derivatives Advanced Chemistry Development/Structure Elucidator 12.01 is used and to perform simulation experiments [38] for docking Auto-Dock

4.2 and AutoGrid 4.2 suites are used. To correct blcharge for target protein Polar H-atoms were additionally added to proteins and ligands respectively. Ligands are assigned with rigid roots and out of bonds existing five bonds were made ready to remain “active” or rotatable. AutoDeck calculation needs PDBQT format, so the HIV-1RT which has been modified to 3-dimensional structure of HIV-1RT is converted to PDBQT. Five million energy evaluation and Lamarckian Genetic Algorithm associated with 150 docking population were to be used in docking [38]. AutoGrid4.2 is used to retrieve pre-calculated grid maps. These maps store energy grids which remain on the basis of interaction with ligand atom moves along with receptor targets. In the docked structures based on the similarity in conformations it is clustered into binds as per cluster root-mean-square deviation.

Docking Results

ParDOCK confirms the docking result obtained from AutoDock4.2 which is purely based on Monte Carlo docking protocol [39]. The utmost satisfactory docking pose is the least binding energy conformation that remains in all the clusters. The interacted ligand and receptor and hydrogen bond lengths were examined by using LigPlot [40].

Water Simulation

A high end research tool for studying protein dynamics using molecular dynamics theory [41][42] is the Gromacs 4.5.5 package. This tool is used to study the molecular dynamics simulation study of protein ligand complex under Gromos 43a1 force field in water. HIV-1 RT/CID: 1656714 complex was neutralized and solvated [43][45]. PME [44] is a renowned technique for computing long-range electrostatics. Linear Constraint algorithm is used for setting entire bond lengths [45]. The system was equilibrated and the subsequent phase also involved in the equilibration. The equilibration is for the solute molecules with a stable specification of the solvent molecules. Earlier executing long MD simulations [46] the whole system was equilibrated for 100 ps at 300 K.

Data Sampling

SHARE is an easy-to-hunt and frequently updated warehouse of synthesized research indication addressing topics associated to HIV/AIDS. Documents in SHARE are included those address a topic focused on HIV based on review policy. Presently, the document pool is selfpossessed of 18,703 scientific abstracts accessed from the PubMed database. The ratio of included and excluded abstracts that researches come across when accomplishing literature selection for HIV systematic assessments are in Table 1, remains imbalanced.

Table I: Information on Share

Name of the attribute	**Number**	**%**
Documents(Total Number)	17,694	100%
Documents excluded	16809	95%
Documents Included	1280	7.5%
Distinct words in abstract	25,954	-
Distinct words in title	5,964	-
Distinct MeSH terms	15,966	-

To accomplish supervised learning, fragment the document pool into two parts. Test set is the first part, which signifies _10% of the complete pool, haphazardly selected to evade any unfairness. The aforementioned comprises 1,588 documents (1280 were included and 16809 excluded). The class distribution in the test set is related to the distribution in the whole document pool. To remove documents from the class, five training sets remain created subsequently separating test set documents. Under sampling methodology on the way to create all training sets, haphazardly eradicates documents from the class. Recognize which is most applicable for this task.

Table 2: Undersampling Approach

Set	**Excluded %**	**Included %**
1	990 10%	8,912 90%
2	990 20%	3,963 80%
3	990 30%	2,316 70%
4	990 40%	1,484 60%
5	990 50%	990 50%

Numerous under-sampling features are applied to the dataset, till a balanced distribution is get hold of, i.e. 50% of excluded and 50% of included documents as in Table 2.

EXTRACTION AND SELECTION

Extraction:

By means of the baseline Bag-Of-Words (BOW) to MeSH terms [37] numerous classification models are constructed and compared, and a set of domain keywords are recognized by researchers on the way to work on HIV systematic assessments. PubMed XML is the source for extracting features and classification algorithms is fed with the same. Features extracted from SHARE: Feature #1: Consider words occurring only twice with a length of minimum 3 characters; (Bagof-words) Feature #2: Consider words occurring minimum twice; (MeSH) Feature #3: HIV systematic reviews relevant keywords in source code.

Selection:

Before feeding classification algorithms use feature selection to investigate the needed. Compare the results by means of Odds Ratio and IDF, to screen the fewer discriminative attributes in the classification models. Calculate the inverse document frequency to execute feature selection with IDF as a metric for each feature. IDF value lesser than 1.0 remain discarded.

ALGORITHM FOR CLASSIFICATION:

Three different classification algorithms are encountered: Naıve Bayes (NB), Logistic Model Trees (LMT) and Support Vector Machine (SVM). NB a durable conditional independence of the features, feature vector F, the features f1; :::; fn are expected to be conditionally independent specified a class C, considered as starting position assessment for feature selection approaches. LMT [38] explained by [39] as existing capable to competently handle tasks with imbalanced datasets. It comprises of a permutation of Decision Tree and LogitBoost algorithms, existing as a classification tree, with logistic regression models happening on its nodes. At every node of the decision tree, the LogitBoost algorithm is used to train a data subset for a definite number of iterations, and to describe a logistic regression model for the existing node. A Decision Tree condition at that time is applied to fragment the existing data subset. SVM [40] was likewise suggested by prior work (e.g. [41], [42]) while dealing by means of imbalanced data. SVM calculates the sideline maximum classifier [43], which is the leading radius about a classification frontier, and attempts to detach data points on a dimensional space, to recognize the dissimilar classes to which they fit in.

Metrics for Evaluation

Evaluated the experiment result as precision (P), recall (R), F1, and F2.

P – Precision - quantitative measure of correct predictions by the classifier of respective class, to generate more relevant outputs.

R – Recall - Ratio of relevant predictions d to all existing relevant documents, evaluates how capable the classifier is. Fβ – Weighted harmonic mean between precision and recall. $F\beta = (1+\beta2)*(P+R)/(\beta2*P+R)$

β – Relative weight of recall over precision β = 1, leading to the F1 score. β = 2, leading to the F2 score.

Ontological Representation

HIVSamp ontology represents the classes and sub classes user for Data sampling with HIV-1 Reverse Transcriptase, Selection of Protein, Docking, Docking Results, Water Simulation, Data Sampling, Extraction and Selection, Algorithm for classification and Metrics for evaluation. OWL language is used for representation and protégé is used for visualizing the classes.

RESULTS

1.Experiments was executed for evaluating the under sampling technique, various class of balances are used in the training set in different feature types among the three classifiers.

2. Run new experiments in same under sampled training sets and classifiers since this time applying feature selection remains the best in the feature configurations.

CONCLUSION

Molecular dynamics simulation and molecular docking reveals binding mechanism. The structural conformational modifications in the ligand binding region/pocket comprising the energetic location catalytic triad residues, motif and other important residues have been identified. The current MD simulations support 5(E)-3-(2-aminoethyl)-5-benzylidene-1,3thiazolidine-2,4-dione, a herbal compound, might be a promising minor molecule for the inhibition of HIV-1 RT and might perform as medicine for the drug-therapy of HIV infection. Beyond the part of CID: 1656714 as a probable anti-HIV candidate, it might too lessen AIDS related co-morbidities owing to occurrence of the enchanted moiety in the structure subsequent successful clinical hearings. These outcomes would be valued for more designing non-covalent type inhibitors with great specificity and superior potency. An advanced overseen learning technique to support the HIV literature screening is framed. Only a minor part of the HIV document gathering symbolized the job target. Data under sampling and feature selection remains examined as approaches to overcome this problem the negative effects. Subsequently experimenting 105 classification models, two models are identified for HIV screening uses BagOf-Words and MeSH terms as features. The usage of an automatic methodology to support literature selection can significantly profit specialists working in HIV systematic evaluations. By means of using classification models, widely held number of documents to be possibly comprised in evaluations by researchers can be exactly labeled. The prototype in this paper could be re-used to support dissimilar literature selection jobs elsewhere the one defined now.

REFERENCES

[1] E. W. Sayers, T. Barrett, D. A. Benson, E. Bolton, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, S. Federhen et al., “Database Resources of the National Center for Biotechnology Information,” Nucleic acids research, vol. 39, no. suppl 1, pp. D38–D51,2011.

[2] J. P. Higgins, S. Green et al., Cochrane Handbook for Systematic Reviews of Interventions. Wiley Online Library, 2008, vol. 5.

[3] U. S. Mudunuri, M. Khouja, S. Repetski, G.Venkataraman, A. Che,B. T. Luke, F. P. Girard, and R. M. Stephens, “Knowledge and Theme Discovery across Very Large Biological Data Sets Using Distributed Queries: A Prototype Combining Unstructured and Structured Data,” PLOS ONE, vol. 8, no. 12, p. e80503, 2013.

[4] C. Quan, M. Wang, and F. Ren, “An Unsupervised Text Mining Method for Relation Extraction from Biomedical Literature,” PLOS ONE, vol. 9, no. 7, p. e102039, 2014.

[5] T. Bekhuis and D. Demner-Fushman, “Screening Nonrandomized Studies for Medical Systematic Reviews: A Comparative Study of Classifiers,” Artificial intelligence in medicine, vol. 55, no. 3, pp. 197–207, 2012.

[6] M. Wang, W. Zhang, W. Ding, D. Dai, H. Zhang, H. Xie, L. Chen, Y. Guo, and J. Xie, “Parallel Clustering Algorithm for Large-Scale Biological Data Sets,” PLOS ONE, vol. 9, no. 4, p. e91315, 2014.

[7] C. N. Arighi, C. H. Wu, K. B. Cohen, L. Hirschman, M. Krallinger, A. Valencia, Z. Lu, J. W. Wilbur, and T. C. Wiegers, “BioCreative-IV Virtual Issue,” Database, vol. 2014, p. bau039, 2014.

[8] F. Leitner, M. Krallinger, C. Rodriguez-Penagos, J. Hakenberg, C. Plake, C.-J. Kuo, C.N. Hsu, R. Tsai, H.-C. Hung, W. W. Lau et al., “Introducing Meta-services for Biomedical Information Extraction,” Genome Biol, vol. 9, no. Suppl 2, p. S6, 2008.

[9] L. Hirschman, G. A. C. Burns, M. Krallinger, C. Arighi, K. B. Cohen, A. Valencia, C. H. Wu, A. Chatr-Aryamontri, K. G. Dowell, E. Huala et al., “Text Mining for the Biocuration Workflow,” Database, vol. 2012, p. bas020, 2012.

[10] L. Hirschman, A. Yeh, C. Blaschke, and A. Valencia, “Overview of BioCreative: Critical Assessment of Information Extraction for Biology,” BMC bioinformatics, vol. 6, no. Suppl 1, p. S1, 2005.

[11] S. Matis-Mitchell, P. Roberts, C. O. Tudor, and C. Arighi IV, “BioCreative IV interactive task,” BioCreative IV Proceedings, vol. 1, 2013.

[12] C. N. Arighi, B. Carterette, K. B. Cohen, M. Krallinger, W. J. Wilbur, P. Fey, R. Dodson, L. Cooper, C. E. Van Slyke, W. Dahdul et al., “An overview of the BioCreative 2012 Workshop Track III: Interactive text mining task,” Database, vol. 2013, p. bas056, 2013.

[13] H. He and E. A. Garcia, “Learning from Imbalanced Data,” IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263–1284, 2009.

[14] G. M. Weiss and F. Provost, “The Effect of Class Distribution on Classifier Learning: An Empirical Study,” Technical Report ML-TR-44, August 2, Department of Computer Science, Rutgers University, 2001.

[15] I. Guyon and A. Elisseeff, “An Introduction to Variable and Feature Selection,” The Journal of Machine Learning Research, vol. 3, pp. 1157– 1182, 2003.

[16] [16] T. Ren-rong, L. Qingjiao, and C. Xulin, “Current status of targets and assays for anti-HIV drugs screening,” Virologica Sinica, vol. 22, pp. 476–485, 2007.

[17] H. Jonckheere, J. Ann_e, and E. De Clercq, “The HIV-1 reverse transcription (RT) process as target for RT inhibitors,” Med. Res. Rev., vol. 20, pp. 129–154, 2000.

[18] A. L. Hopkins, J. Ren, R. M. Esnouf, B. E. Willcox, E. Y. Jones, C. Ross, T. Miyasaka, R. T. Walker, H. Tanaka, D. K. Stammers, and D. I. Stuart, “Complexes of HIV-1 reverse transcriptase with inhibitors of the HEPT series reveal conformational changes relevant to the design of potent non-nucleoside inhibitors,” J. Med. Chem, vol. 39, pp. 1589–1600, 1996.

[19] J. Wang, P. Morin, W. Wang, and P. A. Kollman, “Use of MMPBSA in reproducing the binding free energies to HIV-1 RT of TIBO derivatives and predicting the binding mode to HIV-1 RT of efavirenz by docking and MM-PBSA,” J. Amer. Chem. Soc., vol. 123, pp. 5221–5230, 2001.

[20] [20] E. De Clercq, “HIV-chemotherapy and -prophylaxis: New drugs, leads and approaches,” Int. J. Biochem. Cell Biol., vol. 36, pp. 1800–1822, 2004.

[21] S. G. Sarafianos, K. Das, C. Tantillo, A. D. Clark Jr., J. Ding, J. M. Whitcomb, P. L. Boyer, S. H. Hughes, and E. Arnold, “Crystal structure of HIV-1 reverse transcriptase in complex with a polypurine tract RNA: DNA,” EMBO J., vol. 20, pp. 1449–

1461, 2001.

[22] A. N. Phillips, J. Neaton, and J. D. Lundgren, “The role of HIV in serious diseases other than AIDS,” AIDS, vol. 22, pp. 2409–2418, 2008.

[23] A. K. Monian, G. G. Khadse, and S. R. Sengupta, “Synthesis of some 4-thiazolidinone derivatives as antitubercular agents,” Chem Abstr, vol. 30, no. 324–26, 120, pp. 323–342, 1994.

[24] T. Shinji and M. Yoshitaka. Eur. Pat. Appl., EP, Chem Abstr, vol. 149, no. 884, p. 34071e, 1984.

[25] B. Gunether, B. Wilhelm, D. Stefan, and P. Wilfried, “Ger Offen DE 3,” Chem Abstr, vol. 842, no. 790, 113, p. 6330f, 1990.

[26] D. Bernard, R. J. Pierce, H. Patrick, and L. J. Yyes. Eur. Pat. Appl., EP, Chem Abstr, vol. 322, no. 296, 111, p. 232799, 1990.

[27] R. Harode, V. K. Jain, and T. C. Harma. J. Indian Chem. Soc., Chem Abstr, vol. 67, nos. 262–263, 113, p. 132066f, 1990.

[28] N. Solankee, K. P. Patel, and R. B. Patel, “Efficient synthesis and pharmacological evaluation of some new 4-thiazolidinones and 5-arylidenes,” Archit. Appl. Sci. Res., vol. 4, pp. 72–77, 2012.

[29] A. Jacobo-Molina and E. Arnold, “HIV reverse transcriptase structure-function relationships,” Biochemistry, vol. 30, pp. 6351–6356, 1991.

[30] A. Selvan, C. Seniya, S. N. Chandrasekaran, N. Siddharth, S. Anishetty, and G. Pennathur, “Molecular dynamics simulations of human and dog gastric lipases: Insights into domain movements,” FEBS Lett., vol. 584, pp. 4599–4605, 2010.

[31] V. Rajendran and R. Sethumadhavan, “Drug resistance mechanism of PncA in Mycobacterium tuberculosis,” J. Biomolecular Struct. Dyn., vol. 32, no. 2, pp. 209–221, 2014.

[32] R. Purohit, “Role of ELA region in auto-activation of mutant KIT receptor: A molecular dynamics simulation insight,” J. Biomolecular Struct. Dyn., vol. 32, no. 7, pp. 1033– 1046, 2014.

[33] B. A. Larder, D. J. Purifoy, K. L. Powell, and G. Darby, “Sitespecific mutagenesis of AIDS virus reverse transcriptase,” Nature, vol. 327, no. 6124, pp. 716–717, 1987.

[34] K. Das, J. Ding, Y. Hsiou, A. D. Clark Jr., H. Moereels, L. Koymans, K. Andries, R. Pauwels, P. A. Janssen, P. L. Boyer, P. Clark, R. H. Smith Jr., M. B. Kroeger Smith, C. J. Michejda, S. H. Hughes, and E. Arnold, “Crystal structures of 8-Cl and 9-Cl TIBO complexed with wild-type HIV-1 RT and 8-Cl TIBO complexed with the Tyr181Cys HIV-1 RT drug-resistant mutant,” J. Mol. Biol., vol. 264, pp. 1085–1100, 1996.

[35] J. Ding, K. Das, H. Moereels, L. Koymans, K. Andries, P. A. Janssen, S. H. Hughes, and E. Arnold, “Structure of HIV-1 RT/TIBO R 86183 complex reveals similarity in the binding of diverse nonnucleoside inhibitors,” Nat. Struct. Biol., vol. 2, no. 5, pp. 407– 415, 1995.

[36] Y. Hsiou, J. Ding, K. Das, A. D. Clark Jr., S. H. Hughes, and E.Arnold, “Structure of unliganded HIV-1 reverse transcriptase at 2.7 A_resolution: Implications of conformational changes for polymerization and inhibition mechanisms,” Structure, vol. 4, no. 7, pp. 853–860, 1996.

[37] N. Guex and M. C. Peitsch, “SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modelling,” Electrophoresis, vol. 18, pp. 2714– 2723, 1997.

[38] G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell, and A. J. Olson, “Autodock4 and AutoDockTools4: Automated docking with selective receptor flexibility,” J Comput. Chem., vol. 30, no. 16, pp. 2785–2791, 2009.

[39] [39] A. Gupta, A. Gandhimathi, P. Sharma, and B. Jayaram, “ParDOCK: An all atom energy based Monte Carlo docking protocol for protein-ligand complexes,” Protein Peptide Lett., vol. 14, no. 7, pp. 632–646, 2007.

[40] A. C. Wallace, R. A. Laskowski, and J. M. Thornton, “LIGPLOT: A program to generate schematic diagrams of protein-ligand interactions,” Protein Eng., vol. 8, no. 2, pp. 127– 134, 1995.

[41] B. Hess, K. Carsten, D. Van der Spoel, and E. Lindahl, “GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation,” J. Chem. Theory Comput., vol. 4, pp. 435–447, 2008.

[42] E. Lindahl, B. Hess, and D. Van der Spoel, “Gromacs 3.0: A package for molecular simulation and trajectory analysis,” J. Mol. Model., vol. 7, pp. 306–317, 2001.

[43] B. Hess and N. F. A. Van der vegt, “Hydration thermodynamic properties of amino acid analogues: A systematic comparison of biomolecular force fields and water models,” J. Phys. Chem. B, vol. 110, pp. 17616–17626, 2006.

[44] [44] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L.G. Pedersen, “A smooth

particle mesh Ewald potential," J. Chem Phys., vol. 103, pp. 8577–8592, 1995.

[45] B. Hess, H. Bekker, and H. J. C. Berendsen, "Fraaije JGEM. LINCS: A linear constraint solver for molecular simulations," J. Comput. Chem, vol. 18, pp. 1463–1472, 1997.

[46] T. Darden, D. York, and L. Pedersen, "Particle mesh Ewald: An Nlog (N) method for Ewald sums in large systems," J. Chem. Phys., vol. 98, no. 12, pp. 10089–10092, 1993.

[47] C. E. Lipscomb, "Medical Subject Headings (MeSH)," Bulletin of the Medical Library Association, vol. 88, no. 3, p. 265, 2000.

[48] N. Landwehr, M. Hall, and E. Frank, "Logistic Model Trees," Machine Learning, vol. 59, no. 1-2, pp. 161–205, 2005.

[49] E. Charton, M.-J. Meurs, L. Jean-Louis, and M. Gagnon, "Using Collaborative Tagging for Text Classification," Informatics 2014, pp.32–51, 2013.

[50] V. N. Vapnik, "The Nature of Statistical Learning Theory," 1995.

[51] R. Akbani, S. Kwek, and N. Japkowicz, "Applying Support Vector Machines to Imbalanced Datasets," in Machine Learning: ECML 2004. Springer, 2004, pp. 39–50.

[52] A. Mountassir, H. Benbrahim, and I. Berrada, "An Empirical Study to Address the Problem of Unbalanced Data Sets in Sentiment Classification," IEEE Systems, Man, Cybernetics, pp. 3298–3303, 2012.

[53] S. Marsland, Machine Learning: An Algorithm Perspective, 1st ed. Chapman and Hall, 2009.