Revista Cubana de Ciencias Informáticas
Vol. 17, No. 1, Mes Enero-Marzo, 2023
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 115-129

Tipo de artículo: Artículo originales
Temática: Bioinformática
Recibido: 07/11/2022 | Aceptado: 27/12/2022 | Publicado: 17/02/2023

Evaluación de descriptores moleculares basados en ENM-NMA de la proteasa del VIH-1 para la predicción de la resistencia a fármacos mediante métodos de aprendizaje automático

Assessing ENM-NMA based molecular descriptors of HIV-1 protease for drug resistance prediction by machine learning methods

Jorge Alejandro Jiménez Garí 0000-0001-9586-5354[1*]

[1]Departamento de Bioinformática, Universidad de las Ciencias Informáticas. Carretera San Antonio Km 2 $\frac{1}{2}$, Reparto Torrens, La Lisa, La Habana. jorgeajg@estudiantes.uci.cu

[*]Autor para correspondencia: (jorgeajg@estudiantes.uci.cu)

ABSTRACT

La resistencia a fármacos es un factor importante en el fracaso de la terapia antirretroviral contra el virus de la inmunodeficiencia humana (VIH). Debido a los altos costos de los ensayos fenotípicos directos para evaluar la resistencia a los medicamentos, los ensayos genotípicos, basados en la secuenciación del genoma viral o parte de él, se usan comúnmente para inferir la resistencia a los medicamentos. Para las pruebas genotípicas, la interpretación de la información de la secuencia es el mayor desafío. La gran cantidad de datos que vinculan la información del genotipo y el fenotipo proporcionó un marco para predecir la resistencia a los medicamentos a partir del genotipo, basado en métodos de aprendizaje automático. Los métodos actuales se basan principalmente en la información de la secuencia de las variantes observadas y todavía fallan en gran medida en la predicción de la resistencia en variantes no observadas previamente. Se supone que la inclusión de información estructural y dinámica mejora las predicciones. El uso de descriptores moleculares basados en información dinámica se ha visto limitado por su costo computacional de cálculo. Este estudio

Revista Cubana de Ciencias Informáticas
Vol. 17, No. 1, Mes Enero-Marzo, 2023
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 115-129

muestra la viabilidad de los descriptores dinámicos derivados del análisis de modos normales en modelos de redes elásticas de la proteasa del VIH-1 para predecir la resistencia a los medicamentos. Se utilizaron datos de secuencias de VIH-1 disponibles públicamente y resultados de ensayos de resistencia a fármacos en 7 antirretrovirales para evaluar el rendimiento de 4 algoritmos de predicción utilizando descriptores clásicos y dinámicos por separado, obteniendo resultados comparables.

Palabras clave: descriptores; ENM-NMA; VIH-1; clasificación; fármacos.

ABSTRACT

Drug resistance is a major factor in the failure of antiretroviral therapy against human immunodeficiency virus (HIV). Due to the high costs of direct phenotypic assays to assess drug resistance, genotypic assays, based on sequencing of the viral genome or part of it, are commonly used to infer drug resistance. For genotypic testing, interpretation of sequence information is the biggest challenge. The large amount of data linking genotype and phenotype information provided a framework for predicting drug resistance from genotype, based on machine learning methods. Current methods rely primarily on sequence information from observed variants and still largely fail to predict resistance in previously unobserved variants. The inclusion of structural and dynamic information is supposed to improve predictions. The use of molecular descriptors based on dynamic information has been limited by their computational cost of calculation. This study shows the feasibility of dynamic descriptors derived from normal mode analysis in elastic network models of HIV-1 protease for predicting drug resistance. Publicly available HIV-1 sequence data and drug resistance assay results for 7 ART drugs were used to evaluate the performance of 4 prediction algorithms using classical and dynamic descriptors separately, obtaining comparable results.

Keywords: descriptors; ENM-NMA; HIV-1; classification; drugs.

Introduction

Human immunodeficiency virus (HIV) infection is the cause of acquired immunodeficiency syndrome (AIDS), a life-threatening condition. In recent decades, this infection has progressed from a life-limiting condition to

Revista Cubana de Ciencias Informáticas
Vol. 17, No. 1, Mes Enero-Marzo, 2023
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 115-129

a chronic disease (Tischendorf et al., 2022). Despite the efforts of the scientific community, no permanent cure has been found, but antiretroviral treatments (ART) can control the infection and prevent the progression of the disease. Most of the antiretrovirals act on functional proteins involved in the replication of the virus, which allows a great reduction of the viral load in the organism (Aquaro et al., 2020). Due to the high mutation rate of HIV, the phenomenon of drug resistance usually appears Menéndez-Arias (2013). Antiretroviral drug resistance testing is key for clinical management and epidemiologic surveillance (see Hosseini et al. (2016) and its references). There are two ways to infer the susceptibility of a mutational variant to an antiretroviral: phenotype and genotype tests. Genotypic tests are used more frequently than phenotypic tests due to their lower cost, greater availability, simplicity, and shorter turnaround time. These tests are based on the determination of the nucleotide sequence of the HIV-1 genes whose protein products constitute the ARV target, and their interpretation requires a predictive algorithm that describes susceptibility to a variety of ARVs (Bonet et al., 2013). The computational interpretation of genotypic information leading to a phenotype prediction is an open field in biological sequence analysis and is beyond the problem of HIV ARV resistance. The limitation of using the sequence directly is that learning from the observed cases cannot be generalized to the universe of possible cases. This is particularly important in the prediction of ARV resistance in HIV, where in medical practice rule-based systems are often used to interpret genotypic information, which fail to assess mutational patterns never before observed. It is still challenging to formulate a biological sequence (such as DNA, RNA, or protein) with a discrete model or vector that can effectively reflect its sequence pattern information or capture its key features in question, because almost all existing machine learning algorithms can only handle vectors but not sequence samples (see Liu et al. (2017) and its references). To avoid completely losing information on the sequence order of proteins, Chou (Chou, 2000) proposed the quasi-sequence order approach or PseAAC (pseudoamino acid composition) from where other variants taking into account also the amino acid chemicalphysical properties have arisen, such as Moran autocorrelation, Geary autocorrelation and Moreau-Broto Normalized autocorrelation (see for example Abdullahi et al. (2022); Hajisharifi et al. (2014); Fernández and Caballero (2006)) have arisen. This kind of representation has been widely used in conjunction with machine learning techniques (Steiner et al., 2020), and several techniques have been tested. Among them, the use of (i) decision trees (Riemenschneider et al., 2016; Ramon et al., 2019), (ii) Support Vector Machine (SVM) (Cai et al., 2021), (iii) Multilayer Perceptrons(MLP) and (iv) Deep Neural Networks stand out among others (Bonet et al., 2007; Sheik Amamuddy et al., 2017).

On the other hand, there are structure-based methods that, from the sequence, model the molecular structure and explore the characteristics of the structure of the target-drug complex and focus on its dynamics, which can potentially be related to resistance to the drugs. The most established estructure based method is molec-

ular dynamics (MD) that explores the conformational space on small scales such as nanoseconds (ns) under energetic constraints that generate valuable information but has the drawback of its high computational cost. MD simulations have been widely used to extract knowledge about protein-ligand interactions in the field of HIV drug resistance (See Bastys et al. (2020); Yu et al. (2021); Sohraby and Aryapour (2021)) Some methods, which are based on coarse graining, have been developed to overcome the computational cost of full atomistic molecular dynamics (see, for example, Hosseini et al. (2016)). An intermediate method to explore the dynamics of protein structures, without reaching the resolution level of dynamic simulations but which allows expressing the main characteristics, is Normal Mode Analysis (NMA). This method is based on the coarse graining of the internal displacements of the structure by force and dynamic procedures without using complex calculations, and rely in the physics used to describe small oscillations that can then be expressed as linear combinations of vectors in an orthogonal space, thus, the modeling can be simplified in a generalized problem of decomposition of eigenvalues and eigenvectors (Cui and Bahar, 2005). Many methodologies have been built to further simplify this modeling, among them, Elastic Network Models (ENM) stands out as an efficient way for capturing essential dynamics (Yang et al., 2008). Returning to the problem of predicting resistance to ARVs, one of the most widely used drug targets is the HIV protease given the important role it plays in virus replication. This protein is an enzyme composed of two subunits of 99 amino acids each responsible for the hydrolysis of characteristic peptide bonds in the gag and gag-pol polyproteins (Weber et al., 2021). There are currently nine protease inhibitors (PI) approved for clinical use: saquinavir(SQV), ritonavir(RTV), indinavir(IDV), nelfinavir(NFV), amprenavir(APV), lopinavir(LPV), atazanavir(ATV), tipranavir(TPV).) and darunavir (DRV) (Esté and Cihlar, 2010). Although there is evidence that the dynamics of the HIV-1 protease influences the phenomenon of drug resistance (Costa et al., 2014; Paulsen et al., 2017; Ferreiro et al., 2022), in the literature, there are few, if any studies, that incorporate dynamic information in the predictions of resistance using genotypic information. The present study aims to show the feasibility of incorporating dynamic information into machine learning algorithms for the prediction of antiretroviral resistance in the HIV protease.

Methods or Computational Methodology

To evaluate the feasibility of using descriptors derived from ENM-NMA in the prediction of ARV resistance in HIV, a general methodology was followed that consisted of: (i) obtaining models of the structure of the mutated protease in conjunction with each of the IPs, to which an ENM-NMA was applied and the molecular

Revista Cubana de Ciencias Informáticas
Vol. 17, No. 1, Mes Enero-Marzo, 2023
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 115-129

descriptors based on dynamic information were calculated; (ii) evaluate the performance of four machine learning algorithms taking as independent variables, in addition to the descriptors obtained, four other Chou PAAC-like descriptors. For the management, analysis and constitution of the data, the programming languages Python v3.9 and R v4.2.0 were used through the environments Jupyter lab v3.3.2 and RStudio v2022.02.1 respectively. As hardware resources, all the processes were executed on a computing surface laptop i5-8250U CPU 1.80GHz 16Gb RAM. For computationally intensive tasks, a cloud computing environment from Google Collaboratory (Bisong and Bisong, 2019) was used in its free version.

Dataset

Genotype-phenotype pairs were sampled from PhenoSense experiments for HIV-1 protease from the HIV Drug Resistance Database - Stanford University (HIVdb) (Soo-Yon Rhee et al., 2003; Liu and Shafer, 2006). In this database, sequences are reported as amino acid mutations along the protease coding region, compared to the sequence of the wild-type reference strain HXB2. The phenotype is represented by a resistance factor based on the concentration that the drug needs to inhibit 50% of the virus, known as IC50, provided in fold change respective to the wild type. In order to avoid bias in the primary data, it was decided to use the hight quality filtered data in HIVdb but a preprocessing step was made. Seven drugs were used as targets for classification with the thresholds: 2.0 for TPV, 3.0 for NFV, SQV, IDV, and ATV, 9.0 for LPV, and 10.0 for DRV (Shen et al., 2016) for the fold resistance. Those that exceed the threshold are classified as resistant.

In the preprocessing step of the data, samples that had missing values, unrecognized data, and for the sake of simplicity, samples containing mutations that affect the backbone of the protein (i.e. insertion and deletion events) were eliminated from the initial dataset. Modeling of mutations in the protein backbone often relies on an intensive folding problem and therefore these samples were excluded. The samples were subsequently grouped according to the information recorded in the database regarding the tests performed so that one dataset per drug were constructed.

Of the 2395 samples, 897 samples were selected according to the proposed preprocessing standards. In most cases, the samples contain unresolved positions. Define unresolved position as positions where there is no full amino acid consensus concerning it, so it is annotated with the most probable amino acids. In this case, each annotated amino acid is equally probable, so the selection of one of them is not possible without experimental verifications that support such selection. One possible solution is to evaluate each combination of these. In

Revista Cubana de Ciencias Informáticas
Vol. 17, No. 1, Mes Enero-Marzo, 2023
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 115-129

order to reduce the complexity of the study, samples with these characteristics were removed from the dataset.

Molecular docking

For each variant selected from the database and each PI, protein-ligand complexes were modeled. Two conformations of the protease, open and closed, were modelled by threading. The reference structures were extracted from the RCSB PDB database (available at https://www.rcsb.org/). For the open conformation, structure with pdb code 1HHP was used, and for each protein-drug complex the following structures were used as references: TPV(1D4Y), IDV(1HSG), LPV(1MUI), NFV(1OHR), DRV(1T3R), ATV(2FXE) and SQV(3OXC). For the structure prediction process, the ROSETTA suite was used through its automation interface for the python language, Pyrosetta. Given the extracted reference structure, an energy minimization step was carried out until a stable conformation was found at or near the minimum. For this, the score ref15 (Alford et al., 2017) function was used. Subsequently, these structures were mutated and their subsequent packing was carried out in the mutated position and the surrounding amino acids with a window lenght of 3. To finish the preprocessing, the protease structure was subjected to global packing to correct collisions and residual atomic inconsistencies generated by previously unanalyzed positions.

Low-resolution docking was performed using the protocol ROSETTALIGAND (Davis and Baker, 2009; Meiler and Baker, 2006). to sample the rotational and translational degrees of freedom of the ligand using Monte Carlo simulation(MCM). With a step size of 0.5 Å, 5 replicates with 500 cycles each consisting of restricted transformations with a box size of 7 Å, and interactions with surrounding residues in a range of 10.0 Å. Nine cycles of high resolution refinement were carried out to sample the conformational space of the protein side chains with packing intervals every 3 cycles on the surrounding residues in the range of 10.0 Å. Finally, the last minimization process was performed on the protein backbone in the residues surrounding the ligand in a range of 11.0 Å.

Normal Mode Analysis

NMA was performed for each structure with the Bio3d R package version: 2.4-3 (B.J. et al., 2006) using an ENM based on all heavy atoms of the input structure, which was obtained by fitting to a local energy minimum of a crambin model derived from the AMBER99SB force field. Conformational fluctuation per residue, provided by the NMA calculation, was adopted as the first type of dynamic descriptor. In this way,

Revista Cubana de Ciencias Informáticas
Vol. 17, No. 1, Mes Enero-Marzo, 2023
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 115-129

a quantitative vector with atomic displacements at the residue level (considering the alpha carbon) of both the conformation of the open model and the closed model for each variant is obtained. In the same way, the dynamic cross-correlation matrix (DSSM), which registers the coupling of atomic displacements also at the residue level, was considered as a descriptor, and the same was obtained for both forms, closed and open.

Computational prediction

Four classifiers (SVM, RF, MLP, and DNN) were tested with each of the PI to predict antiretroviral resistance, using seven separate classes of descriptors, including previously extracted information on fluctuations and dynamic cross-correlation matrices. For comparison purposes, descriptors of the kind of Chou PAAC (including PAAC) were calculated starting from many of the chemical and physical properties of amino acids (see Table 1). The stratified Kfold method was used as cross-validation with k=10. The accuracy metric was used to assess performance in each case. In addition, a feature selection process was performed. For this task, an SVM with an L1-norm loss function and a low regularization parameter (C=0.03) was used.

Descriptor	**No. of properties**	**No. of variables**	**properties**
Amino acid composition(AAC)	1	20	Sequence composition
Pseudo amino acid composition(PAAC)	3	298	Hydrophobicity, hydrophilicity, side chain mass
Moran autocorrelation	8	240	Hydrophobicity scale, average flexibility index, polarizability parameter, free energy of amino acid solution in water, residue accessible surface area, amino acid residue volume, steric parameters, relative mutability.
Geary autocorrelation	8	240	Hydrophobicity scale, average flexibility index, polarizability parameter, free energy of amino acid solution in water, residue accessible surface area, amino acid residue volume, steric parameters, relative mutability.
Normalized Moreau - Broto autocorrelation	8	240	Hydrophobicity scale, average flexibility index, polarizability parameter, free energy of amino acid solution in water, residue accessible surface area, amino acid residue volume, steric parameters, relative mutability.
ENM-NMA residue fluctuation	1	198	c-alpha average fluctuation
ENM-NMA dccm	1	99 * 99	Cross-correlation interatomic displacements by residue

Table 1 - Descriptors used in the study.

It should be noted that the scope of this research is reduced only to the comparison with simple descriptors

Revista Cubana de Ciencias Informáticas
Vol. 17, No. 1, Mes Enero-Marzo, 2023
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 115-129

so that it works as a proof of concept. The strict comparison of improvement with respect to descriptors and complex profiles is defined as outside the scope of the investigation. The comparisons were made with the purpose of serving as a starting point for testing the validity of the NMA-based dynamic information for the specific problem of ARV resistance in HIV protease. For comparison of performance, a Friedman test followed by a pairwise Wilcoxom test with Benjamini-Hochberg correction was carried out (following the methodology suggested in Demšar (2006)), to check if there are statistical differences in the accuracy of the different algorithm when considering the effect of the class of descriptor as a factor. Each pair of drug-algorithm was taken as instances for comparison, when reporting mean accuracy in the Kfold experiment.

Results and Discussion

Table 2 shows the results of the computer experiments. It can be seen the average accuracy for each drug-algorithm-descriptor combination. An additional boxplot representation of the full results of the experiments is provided as supplementary material.

It can be seen in the above table that the performance is generally high, regardless of the algorithm, drug or descriptor, with few exceptions. Figure 1 shows the results of the Friedman and Wilcoxon tests. When looking for statistical significance for the differences observed, the Friedman test reported a p value of 1.11e-16, so it can be accepted that there are differences in the mean rank of classification accuracy when considering the effect of the descriptor used for the representation. Taking into account both the mean rank of precision per descriptor (higher is better), as reported in the Friedman test, and the pairwise comparisons in the Wilcoxon test, it could be concluded that dccm is the more informative descriptor, as it seems to outrank the other descriptors, and the pairwise differences are significant with respect to all the others. For their part, fluctuations seem to behave more poorly. For the class of Chou-type descriptors, it is evident that the more the physical-chemical properties of amino acids are considered as a starting point, the better the final result seems to be, being worse only considering the amino acid composition. Even though the mid-range differences between Moran autocorrelation, Geary autocorrelation, and Moreau-Broto autocorrelation suggest that they behave differently, statistically they appear very similar.

From the previous results, it seems evident that the incorporation of dynamic information in the sequence interpretation process could improve the quality of the predictions. In the present study, by testing in the simplest way the individual effect of each descriptor, no combination was performed, which would be the obvious

Revista Cubana de Ciencias Informáticas
Vol. 17, No. 1, Mes Enero-Marzo, 2023
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 115-129

		AAC	Geary_ Autocorrelation	Moran_ Autocorrelation	Norm_ Moreau_Broto	PAAC	fluctuations	dccm
ATV	**SVM**	0.8404	0.9337	0.9284	0.9337	0.8709	0.9105	0.9534
	RF	0.8281	0.9105	0.9122	0.9374	0.8513	0.8835	0.9355
	MLP	0.8154	0.9158	0.9140	0.9374	0.8583	0.9123	0.9534
	DNN	0.8242	0.9157	0.9212	0.9355	0.8673	0.9176	0.9481
DRV	**SVM**	0.8944	0.9306	0.9333	0.9417	0.8861	0.9222	0.9639
	RF	0.9056	0.9167	0.9222	0.9222	0.9222	0.9111	0.9333
	MLP	0.8917	0.9389	0.9278	0.9333	0.9056	0.9361	0.9389
	DNN	0.8667	0.9333	0.9222	0.9139	0.8889	0.8861	0.9417
IDV	**SVM**	0.8410	0.9329	0.9375	0.9376	0.9069	0.9293	0.9494
	RF	0.8363	0.9164	0.9175	0.9235	0.8740	0.9164	0.9269
	MLP	0.8398	0.9281	0.9223	0.9175	0.8940	0.9188	0.9541
	DNN	0.8386	0.9340	0.9341	0.9387	0.8916	0.9282	0.9517
LPV	**SVM**	0.8661	0.9353	0.9296	0.9296	0.9050	0.9381	0.9554
	RF	0.8459	0.9323	0.9340	0.9309	0.8921	0.9223	0.9309
	MLP	0.8518	0.9310	0.9338	0.9353	0.9007	0.9281	0.9569
	DNN	0.8447	0.9353	0.9294	0.9410	0.8978	0.9295	0.9525
NFV	**SVM**	0.8496	0.9449	0.9380	0.9438	0.9059	0.9391	0.9575
	RF	0.8484	0.9265	0.9277	0.9300	0.8840	0.9254	0.9403
	MLP	0.8359	0.9357	0.9391	0.9265	0.8990	0.9312	0.9518
	DNN	0.8415	0.9380	0.9323	0.9322	0.8967	0.9369	0.9553
TPV	**SVM**	0.8199	0.8879	0.8854	0.8985	0.8294	0.8626	0.9113
	RF	0.8271	0.8523	0.8677	0.8704	0.8369	0.8396	0.8778
	MLP	0.8474	0.8984	0.8956	0.9007	0.8524	0.8603	0.9158
	DNN	0.8065	0.8827	0.8553	0.8602	0.7735	0.8601	0.9110
SQV	**SVM**	0.8545	0.9437	0.9343	0.9426	0.8981	0.9296	0.9437
	RF	0.8441	0.9121	0.9109	0.9179	0.8852	0.9109	0.9296
	MLP	0.8337	0.9356	0.9321	0.9355	0.8840	0.9250	0.9544
	DNN	0.8453	0.9391	0.9366	0.9483	0.8899	0.9238	0.9426

Table 2 - Average accuracy values resultant from the ten fold validation of each drug-model-descriptor combination.

procedure in a non-proof of concept scenario like this. There are some aspects that could affect the validity of the results presented here. Statistical comparisons were made using algorithm-drug pairs as instances, when in fact both are factors in the study. In both cases, they require a second-way approach, since the independence of the results of evaluating the same instances with different algorithms cannot be guaranteed. Furthermore, from the chemical-structural point of view, the inhibitors have quite similar principles of action and in many cases share similar interactions with the target, so it is inevitable to think that the results in different PIs may be associated.

Revista Cubana de Ciencias Informáticas
Vol. 17, No. 1, Mes Enero-Marzo, 2023
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 115-129

(A)

	Friedman Rank
AAC	1.855769
Geary_Autocorrelation	8.598395
Moran_Autocorrelation	7.608652
Norm_Moreau_Broto	9.124196
PAAC	3.587820
fluctuations	5.938460
dccm	11.784131

(B)

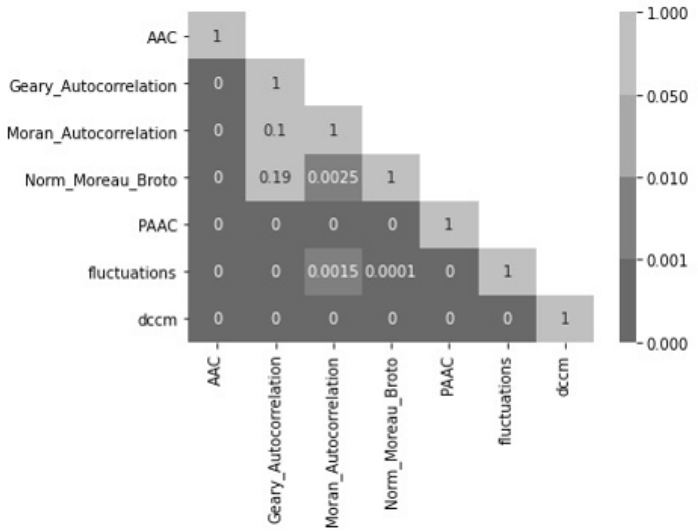

Figure 1 - Friedman test mean range (A) and adjusted p-values related to posthoc Wilcoxon signed rank analysis with Benjamini/Hochberg correction (B).

Conclusions

In this study, the feasibility of two ENM-NMA-based molecular descriptors was examined for the HIV protease drug resistance prediction problem. Different predictive algorithms were tested to demonstrate the reliability of using ENM-NMA derived descriptors to predict HIV protease drug resistance. Of the two descriptors derived from ENM-NMA, only dccm seems to outperform other commonly used descriptors, built according to Chou's logic, obtaining positive results. Although the relationship of fluctuations and atomic displacements in residues relatively distant from the HIV protease binding site is demonstrated, there are still biases regarding the vibrational modes to be taken into account for the study. Therefore, dynamic information from studies using ENM-NMA can be used as a possible source of predictive variables for the problem of antiretroviral resistance prediction in HIV protease.

Supplementary material

Boxplot representation of the full results of the experiments: https://doi.org/10.6084/m9.figshare.22005083.v1

Acknowledgements

We thank Mcs. Mario Pupo Merino for the initial suggestions for carrying out this work, the review of the manuscript, his valuable suggestions and the professional guidance at all times. An important part of the idealization of the research, the methodologies and the academic writing were the result of both jovial and academic conversations carried out with said Bioinformatics science professional.

References

Mustapha Abdullahi, Adamu Uzairu, Gideon Adamu Shallangwa, Paul Andrew Mamza, and Muhammad Tukur Ibrahim. In-silico modelling studies of 5-benzyl-4-thiazolinone derivatives as influenza neuraminidase inhibitors via 2d-qsar, 3d-qsar, molecular docking, and admet predictions. *Heliyon*, 8(8): e10101, 2022. ISSN 2405-8440. doi: 10.1016/j.heliyon.2022.e10101. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9396554.

Rebecca F. Alford, Andrew Leaver-Fay, Jeliazko R. Jeliazkov, Matthew J. Oâ^{TM}Meara, Frank P. DiMaio, Hahnbeom Park, Maxim V. Shapovalov, P. Douglas Renfrew, Vikram K. Mulligan, Kalli Kappel, Jason W. Labonte, Michael S. Pacella, Richard Bonneau, Philip Bradley, Roland L. Jr. Dunbrack, Rhiju Das, David Baker, Brian Kuhlman, Tanja Kortemme, and Jeffrey J. Gray. The rosetta all-atom energy function for macromolecular modeling and design. *J. Chem. Theory Comput.*, 13(6):3031–3048, June 2017. ISSN 1549-9618. doi: 10.1021/acs.jctc.7b00125. URL https://doi.org/10.1021/acs.jctc.7b00125.

Stefano Aquaro, Ana Borrajo, Michele Pellegrino, and Valentina Svicher. Mechanisms underlying of antiretroviral drugs in different cellular reservoirs with a focus on macrophages. *Virulence*, 11(1):400–413, December 2020. ISSN 2150-5594. doi: 10.1080/21505594.2020.1760443. URL https://doi.org/10.1080/21505594.2020.1760443.

Tomas Bastys, Vytautas Gapsys, Hauke Walter, Eva Heger, Nadezhda T Doncheva, Rolf Kaiser, Bert L de Groot, and Olga V Kalinina. Non-active site mutants of hiv-1 protease influence resistance and sensitisation towards protease inhibitors. *Retrovirology*, 17(1):1–14, 2020.

Ekaba Bisong and Ekaba Bisong. Google colaboratory. In *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*, pages 59–64. Apress, Berkeley, CA, 2019. doi: 10.1007/978-1-4842-4470-8_7. URL https://doi.org/10.1007/978-1-4842-4470-8_7.

Grant B.J., Rodrigues A.P.C., ElSawy K.M., McCammon J.A., and Caves L.S.D. Bio3d: An r package for the comparative analysis of protein structures. *Bioinformatics*, 22:2695–2696, Nov 2006.

Isis Bonet, Maria M García, Yvan Saeys, Yves Van de Peer, and Ricardo Grau. Predicting human immunodeficiency virus (hiv) drug resistance using recurrent neural networks. In *Bio-inspired Modeling of Cognitive Tasks: Second International Work-Conference on the Interplay Between Natural and Artificial Computation, IWINAC 2007, La Manga del Mar Menor, Spain, June 18-21, 2007, Proceedings, Part I 2*, pages 234–243. Springer, 2007.

Isis Bonet, Joel Arencibia, Mario Pupo, Abdel Rodriguez, Maria M. Garcia, and Ricardo Grau. Multiclassifier based on hard instances- new method for prediction of human immunodeficiency virus drug resistance. *Curr. Top. Med. Chem.*, 13(5):685–695, 2013. ISSN 1568-0266. doi: 10.2174/1568026611313050011. URL http://dx.doi.org/10.2174/1568026611313050011.

Qihang Cai, Rongao Yuan, Jian He, Menglong Li, and Yanzhi Guo. Predicting hiv drug resistance using weighted machine learning method at target protein sequence-level. *Molecular Diversity*, 25:1541–1551, 2021.

K. C. Chou. Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. *Biochem. Biophys. Res. Commun.*, 278(2):477–483, 2000. ISSN 0006-291X. doi: 10.1006/bbrc.2000.3815. URL https://www.ncbi.nlm.nih.gov/pubmed/11097861.

Mauricio G. S. Costa, Técio G. Benetti-Barbosa, Nathan Desdouits, Arnaud Blondel, Paulo M. Bisch, Pedro G. Pascutti, and Paulo R. Batista. Impact of m36i polymorphism on the interaction of hiv-1 protease with its substrates: insights from molecular dynamics. *BMC Genomics*, 15(7):S5, 2014. ISSN 1471-2164. doi: 10.1186/1471-2164-15-S7-S5. URL https://doi.org/10.1186/1471-2164-15-S7-S5.

Qiang Cui and Ivet Bahar. *Normal mode analysis: theory and applications to biological and chemical systems*. CRC press, 7th edition, 2005.

Ian Davis and David Baker. Rosettaligand docking with full ligand and receptor flexibility. *Journal of Molecular Biology*, 385(2):381–392, January 2009. doi: 10.1016/j.jmb.2008.11.010.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7(1):1–30, 2006. URL http://jmlr.org/papers/v7/demsar06a.html.

José A Esté and Tomas Cihlar. Current status and challenges of antiretroviral research and therapy. *Antiviral research*, 85(1):25–33, 2010.

Michael Fernández and Julio Caballero. Modeling of activity of cyclic urea hiv-1 protease inhibitors using regularized-artificial neural networks. *Bioorg. Med. Chem.*, 14(1):280–294, 2006. ISSN 0968-0896. doi: 10.1016/j.bmc.2005.08.022. URL https://www.ncbi.nlm.nih.gov/pubmed/16202604.

David Ferreiro, Ruqaiya Khalil, María J Gallego, Nuno S Osorio, and Miguel Arenas. The evolution of the hiv-1 protease folding stability. *Virus Evolution*, 8(2):veac115, 2022.

Zohre Hajisharifi, Moien Piryaiee, Majid Mohammad Beigi, Mandana Behbahani, and Hassan Mohabatkar. Predicting anticancer peptides with chou's pseudo amino acid composition and investigating their mutagenicity via ames test. *J. Theor. Biol.*, 341:34–40, 2014. ISSN 0022-5193. doi: 10.1016/j.jtbi.2013.08.037. URL https://www.ncbi.nlm.nih.gov/pubmed/24035842.

Ali Hosseini, Andreu AlibÃ©s, Marc Noguera-Julian, Victor Gil, Roger Paredes, Robert Soliva, Modesto Orozco, and Victor Guallar. Computational prediction of hiv-1 resistance to protease inhibitors. *J. Chem. Inf. Model.*, 56(5):915–923, May 2016. ISSN 1549-9596. doi: 10.1021/acs.jcim.5b00667. URL https://doi.org/10.1021/acs.jcim.5b00667.

Bin Liu, Hao Wu, and Kuo-Chen Chou. Pse-in-one 2.0: An improved package of web servers for generating various modes of pseudo components of dna, rna, and protein sequences. *Nat. Sci. (Irvine)*, 09(04):67–91, 2017. ISSN 2150-4091. doi: 10.4236/ns.2017.94007. URL http://dx.doi.org/10.4236/ns.2017.94007.

Tommy F. Liu and Robert W. Shafer. Web resources for hiv type 1 genotypic-resistance test interpretation. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 42: 1608–18, Jun 2006.

Jens Meiler and David Baker. Rosettaligand: Protein-small molecule docking with full side-chain flexibility. *Proteins: Structure, Function, and Bioinformatics*, 65(3):538–548, 11 2006. doi: 10.1002/prot.21086.

Luis Menéndez-Arias. Molecular basis of human immunodeficiency virus type 1 drug resistance: Overview and recent developments. *Antiviral Research*, 98(1):93–120, 2013. ISSN 0166-3542. doi: 10.1016/j.antiviral.2013.01.007. URL https://www.sciencedirect.com/science/article/pii/S0166354213000223.

Janet L. Paulsen, Florian Leidner, Debra A. Ragland, Nese Kurt Yilmaz, and Celia A. Schiffer. Interdependence of inhibitor recognition in hiv-1 protease. *J. Chem. Theory Comput.*, 13(5):2300–2309, May 2017. ISSN 1549-9618. doi: 10.1021/acs.jctc.6b01262. URL https://doi.org/10.1021/acs.jctc.6b01262.

Elies Ramon, Lluís Belanche-Muñoz, and Miguel Pérez-Enciso. Hiv drug resistance prediction with weighted categorical kernel functions. *BMC bioinformatics*, 20(1):1–13, 2019.

Mona Riemenschneider, Thomas Hummel, and Dominik Heider. Shiva-a web application for drug resistance and tropism testing in hiv. *BMC bioinformatics*, 17(1):1–6, 2016.

Olivier Sheik Amamuddy, Nigel T Bishop, and Özlem Tastan Bishop. Improving fold resistance prediction of hiv-1 against protease and reverse transcriptase inhibitors using artificial neural networks. *BMC bioinformatics*, 18(1):1–7, 2017.

ChenHsiang Shen, Xiaxia Yu, Robert W. Harrison, and Irene T. Weber. Automated prediction of hiv drug resistance from genotype data. *BMC Bioinformatics*, 17(8):278, 2016. ISSN 1471-2105. doi: 10.1186/s12859-016-1114-6. URL https://doi.org/10.1186/s12859-016-1114-6.

Farzin Sohraby and Hassan Aryapour. Comparative analysis of the unbinding pathways of antiviral drug indinavir from hiv and htlv1 proteases by supervised molecular dynamics simulation. *Plos one*, 16(9): e0257916, 2021.

Matthew Soo-Yon Rhee, Rami Gonzales, Bradley Kantor, Jaideep Betts, Robert Ravela, and Shafer. Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Research*, 31 (1):298–303, 1 2003. doi: 10.1093/nar/gkg100.

Margaret C Steiner, Keylie M Gibson, and Keith A Crandall. Drug resistance prediction using deep learning techniques on hiv-1 sequence data. *Viruses*, 12(5):560, 2020.

Jessica S. Tischendorf, James M. Sosman, and Zelalem Temesgen. Chapter 18 - hiv infection. In *A Rational Approach to Clinical Infectious Diseases*, pages 249–267. Elsevier, Philadelphia, 2022. doi: 10.1016/

B978-0-323-69578-7.00018-1. URL https://www.sciencedirect.com/science/article/pii/B9780323695787000181.

Irene T. Weber, Yuan-Fang Wang, and Robert W. Harrison. Hiv protease: Historical perspective and current research, 2021. ISSN 1999-4915.

Lei Yang, Guang Song, Alicia Carriquiry, and Robert L Jernigan. Close correspondence between the motions from principal component analysis of multiple hiv-1 protease structures and elastic network modes. *Structure*, 16(2):321–330, 2008.

Wei Yu, Xiaomin Wu, Yizhen Zhao, Chun Chen, Zhiwei Yang, Xiaochun Zhang, Jiayi Ren, Yueming Wang, Changwen Wu, Chengming Li, et al. Computational simulation of hiv protease inhibitors to the main protease (mpro) of sars-cov-2: Implications for covid-19 drugs design. *Molecules*, 26(23):7385, 2021.

Conflict of interest

The author declares no conflicts of interest regarding this research and the resulting manuscript. The author authorizes the distribution and use of his article for academic and informative approaches.