Silesian University of Technology

Faculty of Automatic Control, Electronics and Computer Science

Doctoral dissertation

Algorithms for the analysis of molecular protein structures and drug-like ligands for modeling and simulation of residence time drug-molecular target

Author: mgr inż. Magdalena Ługowska
Supervisor: prof. dr hab. inż. Marek Kimmel
Co-supervisor: dr inż. Marcin Pacholczyk

Gliwice 2023

Dissertation abstract

"Algorithms for the analysis of molecular protein structures and drug-like ligands for modeling and simulation of residence time drug-molecular target"

mgr. inż. Magdalena Ługowska

1. Introduction

For a drug to be effective, it must not only have the right concentration, but it must also have the ability to bind to the target molecule. One of the most commonly used selection criteria in drug design is the equilibrium binding affinity of a small chemical molecule (the ligand) to its molecular target (the receptor). This affinity is a measure of the persistence of the binding and the strength of the effect of the ligand on the receptor. Ligand potency is an important determinant of ligand activity. This parameter defines the potential of a ligand to efficiently activate a receptor for the production of a strong response in vivo. Thus, affinity is a measure to quantify the efficacy of a drug and to determine the benefit of the interactions that occur between the drug molecule and its target. Therefore, drug design protocols are mainly based on molecules with high binding affinity. However, this approach does not always translate into higher drug efficacy under in vivo conditions because many drugs have non-equilibrium binding properties. Besides pharmacokinetic properties, binding and dissociation rates of drugs can be measured to predict their biological activity profile under in vivo conditions (Copeland 2016, Vauquelin 2016). The concept of drug residence time in a molecular target has also been introduced, which takes into account the conformational dynamics of target molecules that affect drug binding and dissociation. One important observation is that this time is sometimes better correlated with in vivo drug effectiveness than binding affinity (Copeland et al. 2006). Therefore, ligand residence time at the target site (τ) has become a reliable determinant of drug efficacy that is considered along with reaction kinetic parameters in drug discovery programs. However, it is important to keep in mind that the use of residence time as the sole measure of drug efficacy can be a limited picture of reaction kinetics (Folmer 2018).

A substrate-ligand (L), which is a small chemical molecule that is acted upon by an enzyme-receptor (R), begins the series of successive biochemical reactions that occur in the cell. When the ligand (molecule that is the initial stimulus) meets a specific protein receptor, cellular pathways are initiated outside the cell. The molecules come in tightly matched pairs. The receptor recognizes only one (or a few) specific ligands, and the ligand binds only one (or a few) target receptors. When the ligand binds to the receptor, the shape or activity of the receptor changes,

allowing it to transmit a signal or to directly induce a change in the cell. For example, the drug ibuprofen, one of the most widely used analgesics, antipyretics, and anti-inflammatory drugs, is a non-selective inhibitor of cyclooxygenase (COX), the enzyme responsible for converting fatty acids to prostaglandins, and belongs to the NSAID (non-steroidal anti-inflammatory drugs) group of drugs. Prostaglandins are substances involved in the inflammatory process. Inhibition of COX leads to the blockade of prostaglandins. By reducing the production of prostaglandins, ibuprofen is expected to reduce the fever and pain associated with inflammation. However, it should be noted that blocking the COX enzyme has side effects. This enzyme has very important beneficial functions in the body, such as protecting the stomach lining. Therefore, long-term use of COX inhibitors is associated with adverse effects on the stomach and intestines.

Understanding and fully describing receptor-ligand (RL) binding kinetics, as well as the molecular determinants of this fit, is an important part of drug design.

Residence time is a measure of how much time a ligand spends at a protein binding site. In other words, it is the residence time of a drug at a given target site. A drug is pharmacologically active as long as it remains bound to the receptor. Thus, residence time is defined as the inverse of the dissociation rate constant $\tau = 1/koff$. This means that the concentration of a ligand does not affect its residence time in the target, and drugs with long residence times can remain bound even when their concentration falls below the equilibrium dissociation constant Kd. This is particularly important when the drug is cleared from the body, resulting in variable in vivo concentrations. The residence time model is described by a two-step model of binding kinetics that accounts for conformational changes leading to increased molecular complementarity. In this model, a free drug encounters its target in a conformational state defined by a drug-binding pocket that is suboptimal for the structure of the drug molecule. The initial binding step is an association process in which a binding complex (RL*) is formed, which is defined by the association rate constant (k_{on}), dissociation rate constant (k_{off}), and equilibrium dissociation constant (K_d) (Gabdoulline & Wade 1999, 2022). The initial binding is followed by another step in which the system must overcome the energy barriers created by conformational changes in the receptor and ligand so that a new steady state (RL) can be formed in which the binding pocket adopts a structure that is more complementary to that of the drug molecule.

Computer methods to determine binding kinetic parameters

With the increasing interest in residence time and the importance of drug binding kinetics at the target binding site, *in silico* methods are becoming increasingly important. This is especially true when commonly used experimental methods are often time-consuming and costly. In addition, the use of computational methods that predict residence time and characterize reaction kinetics can

support personalized medicine. Patient-specific simulations can speed up a physician's decision to select the optimal drug from several potential candidates. What's more, such calculations can be performed on compounds that have not yet been synthesized, which significantly affects the cost and time of research. It should be noted that the developed *in silico* methods are based on experimental data, which can confirm their reliability.

Computational methods for the estimation of residence time and other kinetic parameters can be divided into two main groups. The first is a set of molecular dynamics methods with improved sampling. The second are methods based on machine learning, often also using molecular dynamics simulations.

2. Motivation and Objectives

An important element in the drug design process is the characterization and understanding of the reaction kinetics of ligand dissociation from the receptor target site. Molecular simulations are key to describing dissociation pathways, predicting kinetic parameters including residence time, and defining structural determinants. To observe the occurrence of rare events during the simulation and to reduce computational complexity, these approaches often use simplifications such as increased sampling. The use of simplifications results in a simulation that is not as detailed as classical molecular dynamics. It does not allow to explain with high accuracy the behavior of the system that occurs on short time scales, such as the rearrangement of atoms in a molecule during the induced fitting step. This can be seen as a certain limitation of these simplifications. On the other hand, classical molecular dynamics allows to understand these rapidly occurring important events, but again it cannot be applied to longer time scales, such as the residence time of a drug in a target, which can range from a few seconds to even hours. Simulations are also mostly an input for machine learning-based algorithms, but their accuracy is not stabilized. The motivation behind the work presented here is the need to develop and apply more efficient and accurate methods to analyze receptor-ligand binding kinetics.

The purpose of this work is to apply and verify solutions to determine the drug residence time to investigate whether they can be used regardless of the size of the molecules or protein family, and to analyze the structural features and interactions in the binding process of InhA protein and its inhibitors. The research presented here aims to answer the following questions:

- Are there and what are the key receptor-ligand interactions that distinguish long- and short-resident ligands?
- Is the τRAMD procedure universal and can it be applied to molecules of different sizes and similarities?
- What is the correlation between relative residence time and experimentally measured?

In addition to answering the above questions, the work has developed two new tools to support the study of ligand-receptor binding kinetics:

- PDBrt kinetic database publicly available at https://pdbrt.polsl.pl/ and
- tool to automatically analyze the interactions occurring between ligand and receptor during simulation.

3. Ligand dissociation trajectories from the receptor binding site by τ Random Acceleration Molecular Dynamics method

τRAMD is an enhanced sampling method for molecular dynamics simulations. It was developed to calculate the relative residence time of pharmacological compounds in their molecular targets, as well as to study the dissociation pathway of ligands from receptor binding sites (Kokh et al. 2018). RAMD simulations are performed for receptor-ligand systems immersed in a solvent, where a small randomly oriented force is applied to the ligand center of mass to accelerate its exit from the receptor active site. After a given time step, the movement of the ligand is checked. If the change in position was less than a predefined threshold distance, a random change in force direction occurs. The simulation ends when the ligand leaves the receptor binding site. This condition is defined by specifying in the configuration file the distance of the ligand from the binding site that corresponds to its release. The simulation time depends on the residence time of the ligand in the target. Ligands with longer residence times take longer to leave the target (simulation time is longer) or require more force to leave the target within a given simulation time. Furthermore, this method does not require prior knowledge of the dissociation path or extensive parameterization. The only parameter that needs to be carefully defined is the magnitude of the force, which should not interfere with the calculated relative residence times. In other words, special care should be taken to ensure that the force set is not too high, as this will force the ligand out of the binding site, and thus the estimated values of the relative residence times for each ligand will be approximate, regardless of the actual dissociation rate. The characteristics of the method described above make τRAMD not only an efficient, but also a relatively simple tool for estimating the relative residence times of drugs for molecular purposes.

In the present study, an analysis of the published results by Kokh et.al was extended to include the molecule geldanamycin in complex with HSP90. The protocol was then repeated for 11 ligands of the InhA receptor and 1 ligand each of the ENR, EGFR and HIV-1 receptors. The purpose of the analysis was to test the versatility and reproducibility of the τRAMD method.

For τRAMD simulations, crystallographic structures of HSP90 inhibitor, HSP90-geldanamycin, InhA inhibitor, ENR-triclosan and EGFR-lapatinib complexes in the bound

state were used as starting structures. The preparation of the molecular models for the simulation included: protonation of the input systems using the PyMOL tool (Schrodinger & DeLano 2020), parameterization using the AmberTools package (Case et al. 2022), including assigning partial charges to individual ligand atoms and changing atom types to those recognized by the Amber package14, calculating the partial atomic charges of the ligands using the AM1-BCC method (Jakalian et al. 2000, 2002), neutralizing the charges (adding Na+ or Cl- ions), and immersing the system in a solvent using the selected force field and generating system topology files. The pmemd tool was then used to perform energy minimization, heating, and equilibration calculations of the systems (Maier et al. 2015). The atomic coordinates of the systems generated in this way were used as input for molecular dynamics simulations performed with NAMD software (Phillips et al. 2020). The 2 ns simulations considered Langevin dynamics for a fixed temperature (300 K) and pressure (1 atm). The resulting atomic coordinates and velocities were then used as input for RAMD simulations. The end of the simulation, and thus the release of the ligand, was observed when the distance between the center of mass of the ligand and the receptor exceeded 40 Å. If no ligand release was observed within 2 ns, the simulation was stopped. Trajectory coordinates were recorded every 100 fs. For each system, molecular dynamics simulation steps were repeated 5 times using NAMD software, which was treated as start files for τRAMD. A set of 10 dissociation trajectories was generated from each start file, resulting in a total of 50 dissociation simulations for each system. Following the published procedure, the residence time was defined as the simulation time required to dissociate the ligand in at least 50% of the trajectories. For each simulated replicate, a bootstrap procedure was used to calculate the transient residence time as the mean of the (t_r) distribution. The correct relative residence time was then estimated as the average of all simulated repetitions for a given system.

Application for HSP90 protein inhibitors

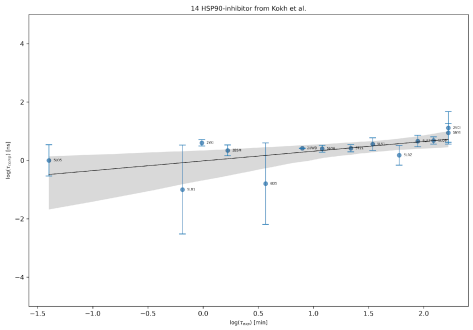
The published application of τ-RAMD to a set of 70 HSP90 ligands with different chemical compositions showed that for 55 of them, there is a strong correlation between the average length of the dissociation trajectory (τ_{comp}) and the experimentally measured residence time. Furthermore, it was observed that there is a correlation between τcomp and the experimentally determined residence time for congeneric, i.e. structurally similar, sets of ligands. It was concluded that τRAMD is an efficient computational method with broad applicability for improving the residence time of a drug target.

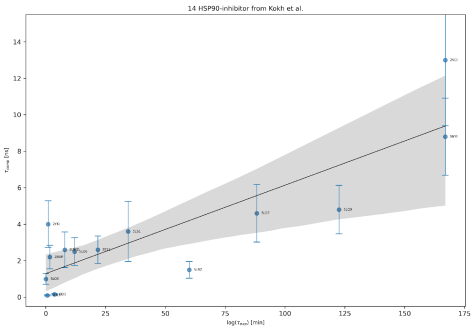
To verify the reproducibility of the results and thus the reliability of the method, analysis using the τ-RAMD protocol was performed for 15 HSP90 inhibitors, 14 of which are included in the published data set and 1 of which was not analyzed by the method authors. All compounds tested

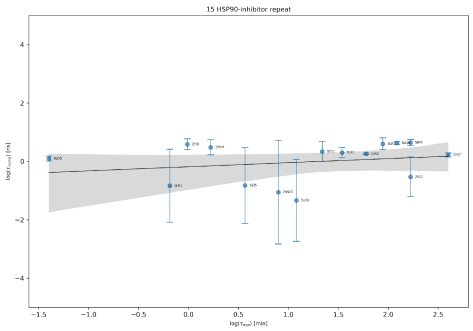
were crystallized in the inhibitor-HSP90 complex and the molecular models are publicly available. This means that the exact same initial structure can be used for testing.

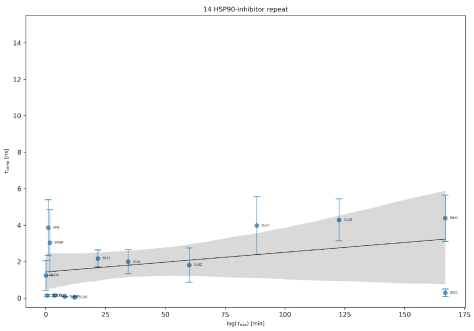
The results obtained are visualized in a scatter plot on a logarithmic and ordinal scale (Figure 1) with the linear fit performed. The black line shows the linear regression of all points except the gray area, which indicates the region within the residual standard deviation of the linear fit calculated with a confidence threshold of 0.95. The error bars show the standard deviations of the calculated residence times.

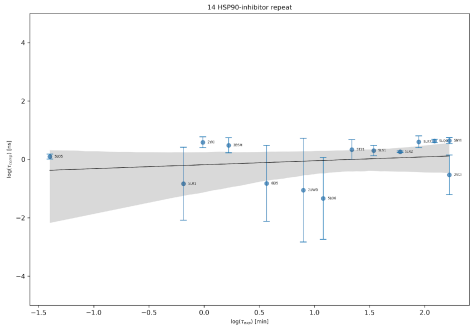
The correlation of the data was determined by Pearson's coefficient. For the published results (Fig. 1 a,b) it is R^2=0.86, indicating a very strong correlation between calculated and experimental data. For repeated simulations of exactly the same systems (Fig. 1 c,d), which were published, the correlation coefficient was R^2=0.4, indicating a moderately positive relationship. In contrast, the data with the additional complex (Fig. 1 e,f) can be described as weakly correlated, since the correlation coefficient is R^2=0.2.











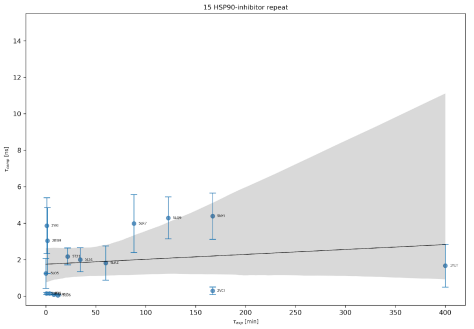


Figure 1: Correlation plot of τ_{comp} with τ_{exp} on a logarithmic (left) and ordinal (right) scale for (a, b) 14 HSP90 inhibitors and published results (c, d) 14 HSP90 inhibitors and simulation repeat results (e, f) 15 HSP90 inhibitors.

Application to InhA protein inhibitors

The τRAMD analysis was performed on a set of 10 InhA protein ligands. The correlation of the data was determined by Pearson's coefficient, which has a value of R^2=0.68, indicating a strong correlation between calculated and experimental data.

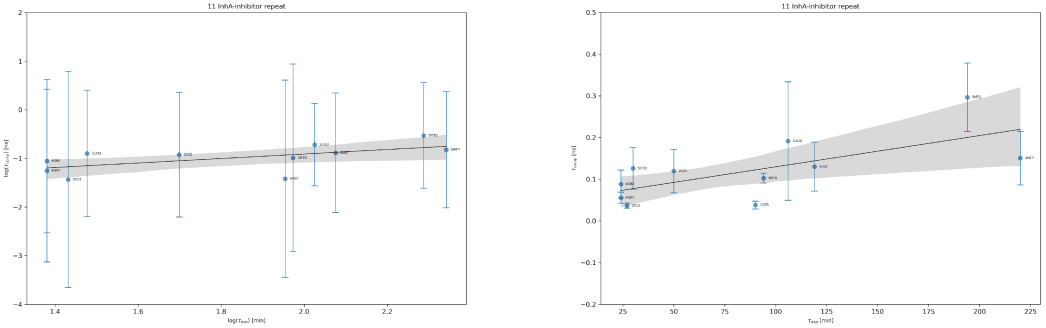


Figure 2: Correlation plot of τ_{comp} with τ_{exp} on a logarithmic (left) and ordinal (right) scale for 10 InhA inhibitors.

Application for ENR, EGFR and HIV-1 protein ligands

The correlation of the data was determined by Pearson's coefficient, which has a value of R^2=0.99, indicating a very strong relationship between calculated and experimental data.

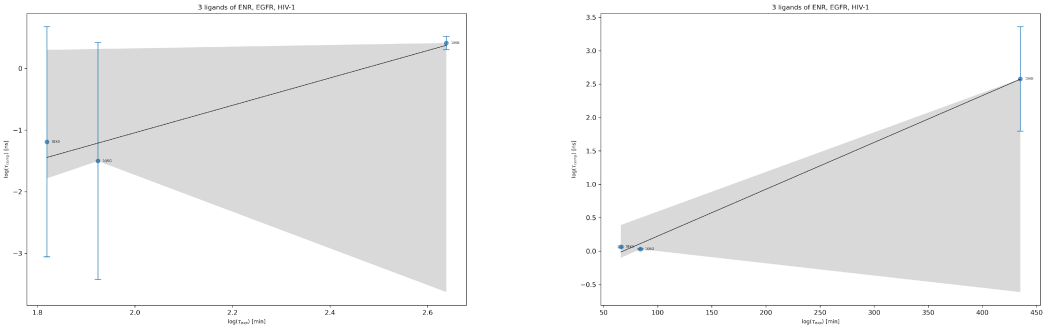


Figure 2: Correlation plot of τ_{comp} with τ_{exp} on a logarithmic (left) and ordinal (right) scale for 3 ligands of ENR, EGFR i HIV-1

Application for all systems studied

To check the universality of the method, the correlation of τ_{comp} against experimentally determined time was checked for all the systems studied. Pearson's coefficient took the value of 0.23 (R^2 =0.23), which is considered a negligible correlation. Therefore, it can be concluded that the accuracy of the method for compounds with greater structural variation is lower.

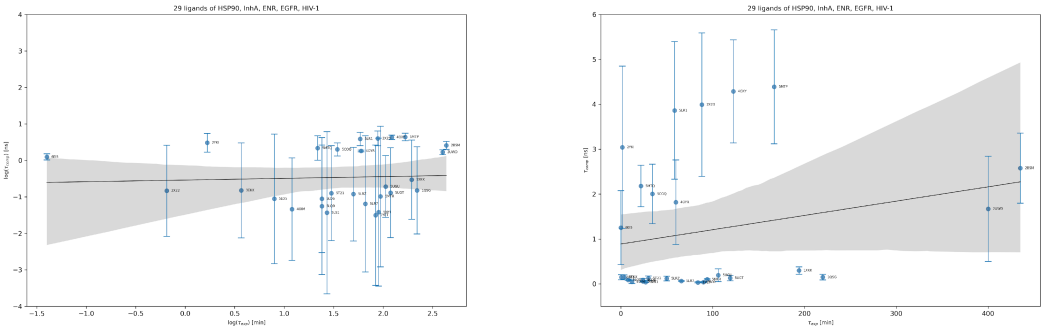


Figure 4: Correlation plot of calculated residence time τ_{comp} (ns) with experimentally determined τ_{exp} (min) for all tested systems, respectively.

4. Ligand properties that affect residence time

An important factor in understanding the kinetics of ligand binding to receptors is the analysis of interactions between them, as well as understanding their structural features. A new approach has been proposed to identify key interactions occurring in the τRAMD dissociation trajectories that affect the residence time of a drug at its molecular target for the studied InhA enzyme inhibitors.

The receptor-ligand interactions were extracted from the τRAMD dissociation trajectory as follows: (i) from each frame (time point) of the trajectory, the position of the ligand center of mass and the coordinates of the atoms constituting the entire system were obtained and stored in separate files using a prepared tcl script for the VMD tool running a Python script; (ii) the obtained coordinates of the position of the atoms in space were used as input to identify ligand-receptor interactions using the RDKit and ProLif libraries of Python (interaction categories: Hydrophobic, π-stacking, π-cation and cation-π, anionic and cationic, and H-bond donor and acceptor); (iii) each interaction was marked as "1" if an interaction was observed or "0", i. e., (iv) repeated interactions were summed to determine the frequency of occurrence of a given contact and averaged over both a single dissociation trajectory and all replicates (trajectories) for a given system (which also eliminated the time dependence of the data); (v) based on the frequency of occurrence of the

interactions, a threshold was defined that allowed the separation of the bound state from the transient and fully released states - for further analysis, the conformations of the system in which an interaction was detected in at least 20% of a single dissociation trajectory were removed.

Standardization of datasets is an important step in the analysis because it removes bias from the original variables. Standardized variables have similar variance. Analysis of the table of interactions of the studied complexes using PCA allowed us to extract the main factors for all the studied InhA inhibitor systems. Figure 5 shows how the main variables correlate with the principal component.

The first two factors model the largest variance in the data, best describing its structure. For the systems studied, these factors describe only about 50% of the total variance in the data. Nevertheless, PCA analysis provides valuable information.

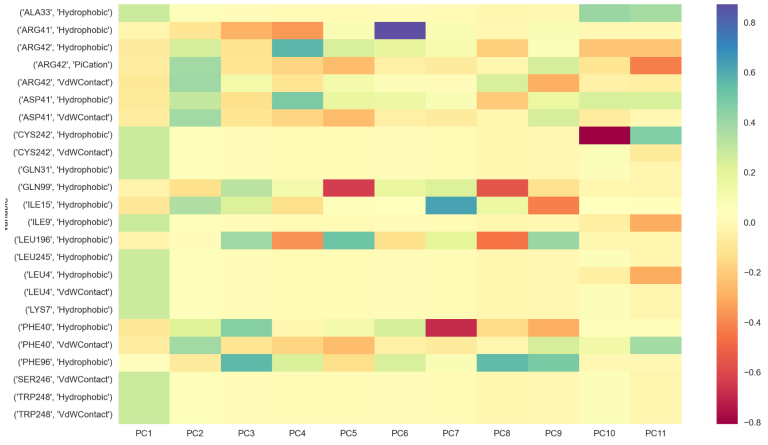


Figure 5: Correlation of features (interactions) with principal components.

A projection of the weights onto the space defined by the first two principal components after k-means clustering of the data is shown in Figure 6 to examine which factors are responsible for the variation in the sample. Each variable (interaction) is represented as a vector whose direction and length determine how much the variable influences each principal component. Thus, it can be concluded that the largest contribution to the formation of the first factor (the values of the coefficients are the highest) is the van der Waals interaction with the amino acid Trp248. Hydrophobic contacts with Leu196, Arg41, Gln99 and Phe96 also contribute to the formation of the first main factor. The latter three together with hydrophobic interactions with Phe40, Arg42, Asp41, Ile15 and van der Waals interactions with Arg42 and Phe40 form the second principal component. The weight projections indicate that hydrophobic interactions with Gln99 and Arg41 are highly correlated. Hydrophobic interactions with the amino acids Phe40, Arg42, Asp41 and Ile15 and van der

Waals interactions with Arg42 and Phe40 are also positively correlated. This set of interactions is negatively correlated with hydrophobic interactions with Gln99, Arg41 and Phe96. The hydrophobic interaction between the ligand and Leu196 is negatively correlated with the van der Waals interaction with the amino acid Trp248, which shows no correlation with the other interactions. The interaction with Trp248 was observed only with the ligand in the 5MTQ complex (residence time of 119 min) and with Leu196 with the ligand in the 5UGU complex (194 min).

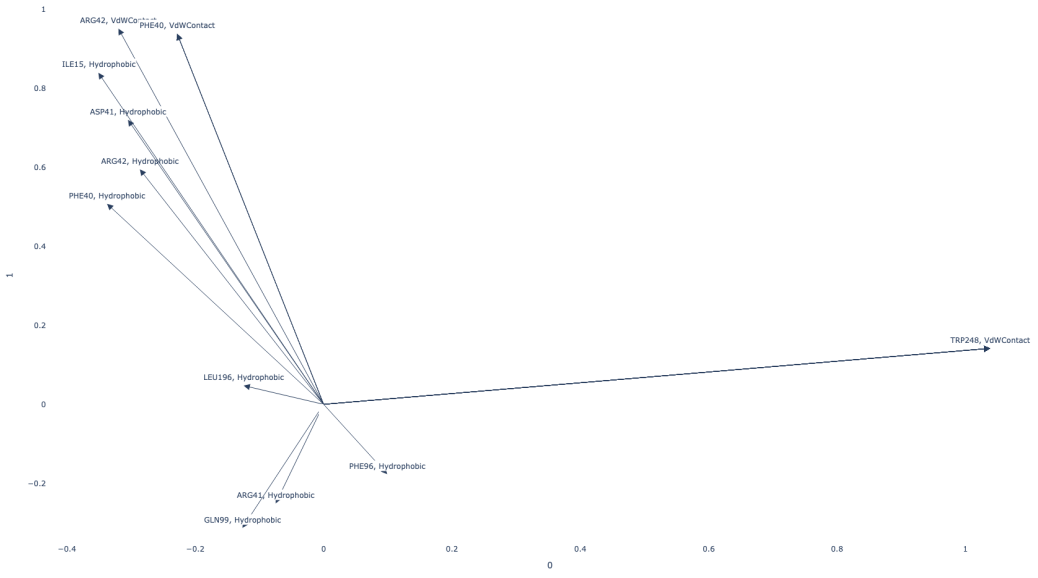


Figure 6: Projection of weights onto the space of the first two principal components.

5. Discussion and conclusions

The aim of the study was to apply and test drug residence time prediction solutions to determine if they can be used regardless of the size of the molecules and the protein family, as well as to identify the molecular properties of the molecules involved in the dissociation of InhA protein inhibitors. The main questions that the presented studies sought to answer are What are the main receptor-ligand interactions that distinguish long- and short-residence ligands? Is the τRAMD method universal and applicable to molecules of different sizes and structural similarities? What is the correlation between the relative residence time and experimental measurements?

Since no database of ligand binding kinetics was available, data were collected from the literature and published in an online database (). literature and published in an online database (https://pdbrt.polsl.pl). The PDBrt database contains 59 ligand entries for 7 different protein families. The database is continuously updated with new data.

The τRAMD method was applied to 5 different protein systems to estimate the relative residence time. To check the reproducibility of the method, the analysis was performed on a set of 15

HSP90 protein inhibitors, 14 of which have been published. On the other hand, to check the versatility, tests were performed on 10 inhibitors of the InhA protein, as well as 3 ligands of the ENR, EGFR HIV-1 proteins. The total set thus consisted of 28 ligands with different structures.

The analysis led to the conclusion that τRAMD is reproducible, but in some cases, more sets of simulations should be performed to reduce the scatter of the data around the mean. It was observed that the τRAMD results show a good or strong correlation with the experimentally determined residence time for compounds of similar structure, i.e. differing by a slight modification, e.g. a shift of the functional group. In contrast, for a structurally diverse set of ligands, the method was found to have poor performance, suggesting a limited application of τRAMD in the drug discovery process.

An approach that takes into account the transition states of the system and the interactions that occur between the protein molecule and the ligand molecule was used to determine the molecular properties of the InhA protein and its inhibitors during the residence time-specific dissociation process. The approach requires no prior knowledge of the binding mechanism and is based on PCA and k-means clustering analysis.

The identified groups form interaction fingerprints specific to ligands with specific residence times, as well as groups that distinguish ligands based on the length of residence time in a molecular target.

Additional analysis of molecular descriptors that provide deeper insight into the molecular properties of the protein-ligand system would likely be an interesting extension of the method. This is an area for further development of the method.

The analysis of the frequency of a given interaction in the complexes allowed us to identify key amino acids that can have a significant impact on the variation of the residence time in the complexes:

- hydrophobic interactions with the amino acid Leu196, as well as van der Waals interactions with Phe40, Asp41, Arg42 and π-cation with Arg42 favor longer residence times, as they were identified only in complexes with ligands with the longest residence times in the studied data set (106, 194 and 220 min),
- for compounds with relatively short residence times (30 and 50 min), the hydrophobic interaction between the ligand and the amino acid Gln99 is characteristic.

Principal component analysis (PCA) identified the factors responsible for the differences in ligand residence times. These factors are as follows:

- van der Waals interactions with the amino acid Trp248,
- hydrophobic interactions with the amino acids Gln99 and Arg41,
- hydrophobic interactions with the amino acids Phe40, Arg42, Asp41, Ile15, and

- van der Waals interactions with Arg42 and Phe40.

For compounds with similar structures, studies have shown that the relative residence time correlates well with the experimentally determined time. The proposed algorithm can be used to identify key molecular features for the rate of ligand dissociation from the binding site for compounds with similar structures.

6. Bibliography

Case, D., Aktulga, H., Belfon, K., Ben-Shalom, I., Berryman, J., Brozell, S., Cerutti, D., III, T. C., Cisneros, G., Cruzeiro, V., Darden, T., Duke, R., Giambasu, G., Gilson, M., Gohlke, H., Goetz, A., Harris, R., Izadi, S., Izmailov, S., Kasavajhala, K., Kaymak, M., King, E., Kovalenko, A., Kurtzman, T., Lee, T., LeGrand, S., Li, P., Lin, C., Liu, J., Luchko, T., Luo, R., Machado, M., Man, V., Manathunga, M., Merz, K., Miao, Y., Mikhailovskii, O., Monard, G., Nguyen, H., O'Hearn, K., Onufriev, A., Pan, F., Pantano, S., Qi, R., Rahnamoun, A., Roe, D., Roitberg, A., Sagui, C., Schott-Verdugo, S., Shajan, A., Shen, J., Simmerling, C., Skrynnikov, N., Smith, J., Swails, J., Walker, R., Wang, J., Wang, J., Wei, H., Wolf, R., Wu, X., Xiong, Y., Xue, Y., York, D., Zhao, S., & Kollman, P. (2022), 'Amber 22'. URL: https://ambermd.org/AmberTools.php

Copeland, R. A. (2016), 'The drug-target residence time model: A 10-year retrospective', Nature Reviews Drug Discovery 15, 87–95. URL: http://dx.doi.org/10.1038/nrd.2015.18

Copeland, R. A., Pompliano, D. L. & Meek, T. D. (2006), 'Drug-target residence time and its implications for lead optimization', Nature Reviews Drug Discovery 5, 730–739

Dahl, G. & Akerud, T. (2013), 'Pharmacokinetics and the drug-target residence time concept'.

Dowling, M. R. & Charlton, S. J. (2006), 'Quantifying the association and dissociation rates of unlabelled antagonists at the muscarinic m 3 receptor', British Journal of Pharmacology 148, 927–937

Folmer, R. H. (2018), 'Drug target residence time: a misleading concept', Drug Discovery Today 23, 12–16. URL: https://doi.org/10.1016/j.drudis.2017.07.016

Gabdoulline, R. R. & Wade, R. C. (1999), 'On the protein-protein diffusional encounte complex'.

Gabdoulline, R. R. & Wade, R. C. (2022), 'Biomolecular diffusional association gabdoulline and wade 205'

Jakalian, A., Bush, B. L., Jack, D. B. & Bayly, C. I. (2000), 'Fast, efficient generation of high-quality atomic charges. am1-bcc model: I. method', Journal of Computational Chemistry 21, 132–146.

Jakalian, A., Jack, D. B. & Bayly, C. I. (2002), 'Fast, efficient generation of high-quality atomic charges. am1-bcc model: Ii. parameterization and validation', Journal of Computational Chemistry 23, 1623–1641

Kokh, D. B., Amaral, M., Bomke, J., Gradler, U., Musil, D., Buchstaller, H. P., Dreyer, M. K., Frech, M., Lowinski, M., Vallee, F., Bianciotto, M., Rak, A. & Wade, R. C. (2018), 'Estimation of drug-target residence times by τ-random acceleration molecular dynamics simulations', Journal of Chemical Theory and Computation 14, 3859–3869.

Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E. & Simmerling, C. (2015), 'ff14sb: Improving the accuracy of protein side chain and backbone parameters from ff99sb', Journal of Chemical Theory and Computation 11(8), 3696–3713. PMID: 26574453. URL: https://doi.org/10.1021/acs.jctc.5b0025

Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kale, L. & Schulten, K. (2020), 'Scalable molecular dynamics with namd'. URL: www.ks.uiuc.edu

Schrodinger, L. & DeLano, W. (2020), 'Pymol'. URL: http://www.pymol.org/pymol

Swinney, D. C. (2004), 'Biochemical mechanisms of drug action: what does it take for success?'. URL: www.nature.com/reviews/drugdisc

Vauquelin, G. (2016), 'Effects of target binding kinetics on in vivo drug efficacy: koff, kon and rebinding', British Journal of Pharmacology pp. 2319–2334