12-2017

The Role of Protein Structural Ensembles in Thermostability and Ligand Binding

Wilfredo Evangelista Falcón
University of Tennessee

To the Graduate Council:

I am submitting herewith a dissertation written by Wilfredo Evangelista Falcón entitled "The Role of Protein Structural Ensembles in Thermostability and Ligand Binding." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Biochemistry and Cellular and Molecular Biology.

Jerome Baudry, Major Professor

We have read this dissertation and recommend its acceptance:

Gladys Alexandre, Jaan Mannik, Engin Serpersu, Jeremy Smith

Accepted for the Council:

Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

The Role of Protein Structural Ensembles in Thermostability and Ligand Binding

A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Wilfredo Evangelista Falcón
December 2017

DEDICATION

Dedicated to my father, whose lessons keep me moving forward even many years after his departure. To my mother, whose courage and tenacity I inherited.

ACKNOWLEDGEMENTS

I would like to express my gratitude to my Ph.D. advisor, Dr. Jerome Baudry for all his support, guidance, and challenges over the past four and half years. I sincerely thank Dr. Baudry for encouraging my intellectual freedom and critical thinking, leading me to improve my writing and presentation skills.

I want also to express my appreciation to all my committee members: Dr. Gladys Alexandre, Dr. Jeremy Smith, Dr. Engin Serpersu, and Dr. Jaan Mannik for their advice and critical insights, they made me pursue improvements in scientific communication, critical thinking, and academic aspects of my career.

My deepest gratitude to our collaborators: Dr. Xiaomin Jing, Dr. Engin Serpersu for their help and discussion over our project on Thermostability of enzymes; to Dr. Sally Ellingson for her collaboration on the Ensemble-base docking project; and to the members of the Center for Molecular Biophysics at ORNL, their comments, aid, and collaboration made these past years incredibly exciting.

My sincere appreciation to all the faculty, staff, friends and colleagues from the Biochemistry, Cell & Molecular Biology department, they made this journey not just productive, but also enjoyable, I must particular thank to my friend Justin Lindsay for all his support and friendship during my stage at UT; to all my friends from my homeland, their support, advice, and encouragement in the past and present made me who I am.

Finally, my infinity gratitude to my beloved family, without their support and understanding this journey would have been just impossible.

ABSTRACT

The role of protein structural ensembles has been shown to be very important for different physical and chemical properties of proteins. The work presented in this dissertation explores two of these properties:

i) Thermostability, by characterizing, at three different temperatures, the dynamics of aminoglycoside nucleotidyltransferase 4' (ANT). This homodimeric enzyme detoxifies antibiotics. It possess two known variants, D80Y and T130K, with higher melting temperatures than the wild type. These mutations, however, would cause changes in the distributions of conformations in the ensemble and, consequently, on the dynamics of the protein. To test this hypothesis, the wild type and variants were examined by using molecular dynamics simulations and the results were compared with previous experimental information in order to characterize the similarities and differences between the, so-called, thermophilic and thermostable variants of this enzyme.

ii) Ligand binding: Since proteins are in general dynamic structures, it would be expected that the effectiveness of ligand binding varies as the protein's conformation changes. One of the most targeted protein family in the field of drug discovery/design is the G-Protein Coupled Receptor (GPCR) family. Over 30% of approved drugs target this family of proteins. This project examines, via *in silico* experiments, the differences in ligand binding between different conformations of GPCRs. To this end, GPCR ligand structures, actual binding (actives) and non-binding (decoys) ligands, were obtained from public databases, and eight GPCRs structures were selected to generate 5,000 conformational states for each protein. Ensemble-based docking was performed on representative structures of these 5,000 conformers and on a subset of 3,000 conformers from each of the eight proteins. Decoys and statistical analysis were incorporated in the docking simulations to test whether the sampled protein conformations can bind active ligands in greater numbers than the random selection from the pool of active and decoys. The results

show that some conformations bind more ligands than other conformations, random selection, or the crystal structure. Characterizing the entire ensemble of protein conformations can improve the number of bound active ligands identified computationally, compared to random selection of compounds or docking using only a single crystal structure.

TABLE OF CONTENTS

INTRODUCTION..1

PROTEIN STRUCTURE ENSEMBLES 1

ABOUT THE WORK DEVELOPED IN THIS DISSERTATION 2

REFERENCES 4

CHAPTER I RELATIONSHIP BETWEEN STRUCTURE ENSEMBLE AND THERMOSTABILITY OF AMINOGLYCOSIDE NUCLEOTIDYLTRANSFERASE 4′ 6

ABSTRACT..7

INTRODUCTION 7

METHODS..9

Molecular Dynamics...9

Principal Component Analysis 9

Free Energy Landscape of the Structural Ensemble......................................11

RESULTS AND DISCUSSION 13

Principal Component Analysis of the MD trajectories 13

Free Energy Landscape..16

CONCLUSIONS 19

REFERENCES 20

CHAPTER II THE ROLE OF PROTEIN STRUCTURE ENSEMBLE IN GPCR-DRUG BINDING. 22

ABSTRACT..23

INTRODUCTION 24

In silico experiments on Adverse Drug Reaction...24

G-Protein Coupled Receptors...25

Statistical Measurement of Performance 26

METHODS..28

Collecting Structural data 28

Preparation of Protein Structures for MD Simulations....................................28

Sampling conformations via Coarse-Grained modelling & MD simulations ... 28

Pre-docking preparation of receptors and ligands.. 30

Docking on representative structures of clusters 30

Statistical Measurement for Conformation's Performance 31

RESULTS AND DISCUSSION 34

Docking on Clustered Structures 34

Docking on non-Clustered Trajectory.. 39

CONCLUSIONS 45

REFERENCES 46

APPENDIX... 51

CHAPTER III OPPORTUNITY AND CHALLENGES OF ENSEMBLE-BASED DOCKING IN TOXICITY 53

ABSTRACT.. 54

INTRODUCTION 54

CURRENT STATE OF ENSEMBLE-BASE DOCKING AND ADR...................................... 56

INTEGRATION OF MOLECULAR TOOLS AND DATABASES 58

REFERENCES 63

CONCLUSIONS.. 68

VITA.. 70

LIST OF TABLES

Table II-1 Set of GPCR proteins and their number of selected ligands for this work. 33

Table II-2 Number of unique active ligands selected via: Random Selection, docked in crystal structure, in representative structures of clustered frames, and in 3000 frames of 600 ns of MD trajectory. 36

Table II-3. Percentage of the Total Number of Actives bound by the Crystal Structure, Representative structures, and ensemble from trajectory. 37

Table II-4. Number of significant frames found after docking on clustered data and trajectory for each subset. 38

Table II-5. Thresholds to determine outliers for ADORA2A and ADRB2 in random selection distributions. 51

Table II-6. Thresholds to determine outliers for HRH1 and CHRM2 in random selection distributions. 51

Table II-7. Thresholds to determine outliers for HTR1B and HTR2B in random selection distributions. 52

Table II-8. Thresholds to determine outliers for OPRD1 and OPRK1 in random selection distributions. 52

LIST OF FIGURES

Figure I.1 Crystal Structure of variant D80Y. 8

Figure I.2. Principal modes of motion of ANT. 14

Figure I.3. First and second mode of motion of the ANT structure. 15

Figure I.4. Free energy landscape for the three species of ANT. 18

Figure II.1. Crystal Structure of ADORA2A. 27

Figure II.2 Interquartile definition, and Probability distribution for random selection at 5.0% for ADORA2A. 32

Figure II.3 Unique Actives docked in clustered conformations. 35

Figure II.4 Docking results on four GPCRS performed on frames from 600 ns of MD trajectory. 40

Figure II.5 RMSD of the α-Carbons for the four GPCRs as function of simulation time. 44

Figure II.6 Volume of binding site for each of the four systems as function of time 44

INTRODUCTION

Protein Structure Ensembles

The structure-function relationship theory, in which the 3D structure of protein defines function, was bolstered by the availability of high-resolution structures of proteins. Since then, the idea that proteins can be defined by one unique structure, the “right folded state”, prevailed for a few decades. However, the idea of a relationship between energy-landscape and function was proposed by Frauenfelder in the 70's (Austin, Beeson, Eisenstein, Frauenfelder, & Gunsalus, 1975), when this group observed a correlation between the non-exponential kinetics of carbon monoxide and oxygen rebinding to myoglobin and the energy barrier produced by temperature and ligand concentration. Previous observations on myoglobin between the late 1880's and early 1920's, supported this concept, and protein dynamics as a factor affecting its function became more accepted. This led to the idea that structural states are in thermal equilibrium while at the same time solvent and ligands affect the energy landscape, and the population of conformational states (Frauenfelder, Fenimore, & McMahon, 2002; Loncharich & Brooks, 1990). These results on myoglobin and other proteins, e.g. Dihydrofolate Reductase, in which conformational substates modulate the transference rate of hydride (Thorpe & Iii, 2005), led to the idea that proteins exist as ensembles of similar structures. These transient microstates interconvert between each other, and the average of these conformers are resolved by high-resolution techniques, such as crystallography and NMR spectroscopy. The concept of native states as a set of structural conformations or microstates provide a basis to rationalize the physical-chemical properties of proteins such as stability, solubility, affinity, binding, and specificity/promiscuity for ligands as function of these structural ensembles. Thus, biological features of proteins are the result of an energy-weighted contribution from each conformer in the ensemble (Hilser, Garcia-Moreno E., Oas, Kapp, & Whitten, 2006).

The striking improvement of molecular dynamics (MD) simulations since the first reported work in the late 70's (McCammon, Gelin, & Karplus, 1977) has been due to the increase of computational power as well as improvement of potential energy functions. This allows the generation of conformational changes in proteins on time scales that go from femtoseconds to milliseconds. This range of timescales comprises atomic fluctuations, side chains rotations, loop motions, and large domain motions (Henzler-Wildman & Kern, 2007). The combination of all these motions at different time scales will result in many possible conformational structures, which will be part of the ensemble at particular conditions, e.g., temperature, salt concentration, protonation states, bound-ligand, free-ligand. Thus, MD simulations are tools employed to build ensembles of conformational states for any protein for which a structure in available.

About the work developed in this dissertation

Chapter 1 describes the work done on aminoglycoside nucleotidyltransferase 4' (ANT) to study its thermostability. MD simulations at three temperatures were performed for the wild type and two variants, D80Y and T130K. The flexibility of the protein has been analyzed by applying principal component analysis (PCA) to the trajectories obtained from the MD simulations. PCA allows the deconvolution of the main modes of motion, thus, it is possible to find out which regions in the protein have the largest amplitude motions, their directions, and how these motions contribute to the formation of protein structure ensembles (Jing, Evangelista Falcon, Baudry, & Serpersu, 2017).

Chapter 2 describes the development of a protocol to perform ensemble-based docking on G-Protein Coupled Receptors (GPCRs) preventing false positives. This is particularly important in the field of drug discovery and design, because when a new potential drug is tested, this new molecule must be able to bind at least one of the conformers present in the population of the target protein's structure. Otherwise, the binding affinity may be too low to be effective under physiological conditions. Four GPCRs were used in a more exhaustive study of conformational selection. Conformations for each protein ensemble were generated via Coarse-Grained (CG) MD simulations. Molecular docking was then performed by using VinaMPI (Ellingson, Smith, & Baudry, 2013). The most important contribution of this project is the incorporation of the statistical concept of ‘outliers’ as a threshold to determine whether a conformational state is significant beyond random selection its capacity to be selected by the protein ligands.

Chapter 3 proposes future directions in the field of drug discovery and design by using ensemble-based docking simulations as has been previously suggested (Evangelista et al., 2016). The formulation of a reliable protocol to perform such simulations becomes imperative in order to test not only the binding to target proteins, but also off-target proteins. This kind of *in silico* experiments can also be applied to predict adverse drug reactions.

References

Austin, R. H., Beeson, K. W., Eisenstein, L., Frauenfelder, H., & Gunsalus, I. C. (1975). Dynamics of ligand binding to myoglobin. *Biochemistry, 14*, 5355-5373. doi:10.1021/bi00695a021

Ellingson, S. R., Smith, J. C., & Baudry, J. (2013). VinaMPI: Facilitating multiple receptor high-throughput virtual docking on high-performance computers. *Journal of Computational Chemistry, 34*, 2212-2221. doi:10.1002/jcc.23367

Evangelista, W., Weir, R. L., Ellingson, S. R., Harris, J. B., Kapoor, K., Smith, J. C., & Baudry, J. (2016). Ensemble-based docking: From hit discovery to metabolism and toxicity predictions. *Bioorganic and Medicinal Chemistry, 24*, 4928-4935. doi:10.1016/j.bmc.2016.07.064

Frauenfelder, H., Fenimore, P. W., & McMahon, B. H. (2002). Hydration, slaving and protein function. *Biophysical Chemistry, 98*, 35-48. doi:10.1016/S0301-4622(02)00083-2

Henzler-Wildman, K., & Kern, D. (2007). Dynamic personalities of proteins. *Nature, 450*, 964-972. doi:10.1038/nature06522

Hilser, V. J., Garcia-Moreno E., B., Oas, T. G., Kapp, G. K., & Whitten, S. T. (2006). A Statistical Thermodynamic Model of the Protein Ensemble. *Chem. Rev., 106*, 1545-1558.

Jing, X., Evangelista Falcon, W., Baudry, J., & Serpersu, E. H. (2017). Thermophilic Enzyme or Mesophilic Enzyme with Enhanced Thermostability: Can We Draw a Line? *The Journal of Physical Chemistry B, 121*, 7086-7094. doi:10.1021/acs.jpcb.7b04519

Loncharich, R. J., & Brooks, B. R. (1990). Temperature dependence of dynamics of hydrated myoglobin. Comparison of force field calculations with neutron scattering data. *Journal of molecular biology, 215*, 439-455. doi:10.1016/S0022-2836(05)80363-8

McCammon, J. A., Gelin, B. R., & Karplus, M. (1977). Dynamics of folded proteins. *Nature, 267*, 585-590. doi:10.1038/267585a0

Thorpe, I. F., & Iii, C. L. B. (2005). Conformational Substates Modulate Hydride Transfer in Dihydrofolate Reductase. 12997-13006.

CHAPTER I

RELATIONSHIP BETWEEN STRUCTURE ENSEMBLE AND THERMOSTABILITY OF AMINOGLYCOSIDE NUCLEOTIDYLTRANSFERASE 4′

A version of this chapter was originally published by Xioamin Jing, Wilfredo Evangelista Falcón, Jerome Baudry, Engin Serpersu:

Xioamin Jing, Wilfredo Evangelista Falcón, Jerome Baudry, Engin Serpersu. “Thermophilic Enzyme or Mesophilic Enzyme with Enhanced Thermostability: Can We Draw a Line?” *J. Phys. Chem. B*, 2017, 121 (29), pp 7086–7094.

DOI: 10.1021/acs.jpcb.7b04519

This was a collaborative project between Dr. Engin Serpersu’s and Dr. Jerome Baudry’s laboratories. My contributions to this study include: 1) building and performing the computer simulations of this project. 2) Discussions of strategies and results with Dr. Baudry as well as the writing of the computational aspects of this study.

Dr. Xiaomin Jing was a graduate student in Dr. Serpersu’s laboratory and performed all the experimental work done for this project.

Abstract

Aminoglycoside nucleotidyltransferase 4′ (ANT) is a homodimeric enzyme that modifies the C4′-OH site of aminoglycoside antibiotics by nucleotidylation. A few single- and double-residue mutants of this enzyme (T130K, D80Y, and D80Y/T130K) from *Bacillus stearothermophilus* show increased thermostability. Our main interest is to study the structural changes of this enzyme as result of mutations and variation in the temperature. Three systems were prepared based on the crystal structure of the mutant D80Y, WT and T130K. MD simulations on these three systems at 300K, 322K, and 330K were performed for 100 ns each one, in total 900 ns of production time, i.e. three systems at three different temperatures.

Introduction

Thermophilic enzymes are exclusively produced by Archea and Bacteria inhabiting natural hot environments, such as volcanic pools, hot springs, or any other natural hot environment. These organisms' enzymes have a property called thermostability, namely, these enzymes can perform their activity at temperatures higher than 50 °C and up to 110°C. Due to this wide range of temperatures, such enzymes have been classified as thermophilic, performing their function in the at 50-80 °C range and hyperthermophilic in the 80-110 °C range, while regular enzymes, mesophilic, function in the 20-40 °C range (Danson, Hough, Russell, Taylor, & Pearl, 1996; Fields, 2001). ANT belongs to the thermophilic category and has been isolated from mesophile bacteria, *Staphylococcus aureus*, in a genetic study (Lacey & Chopra, 1973). Studies on induced thermostable variants and screening for aminoglycosides resistant mutants were developed in the 80's (Liao, McKenzie, & Hageman, 1986; Matsumura & Aiba, 1985), two single mutants were identified as thermophilic, D80Y and T130K; as well as the double mutant D80Y/T130K.

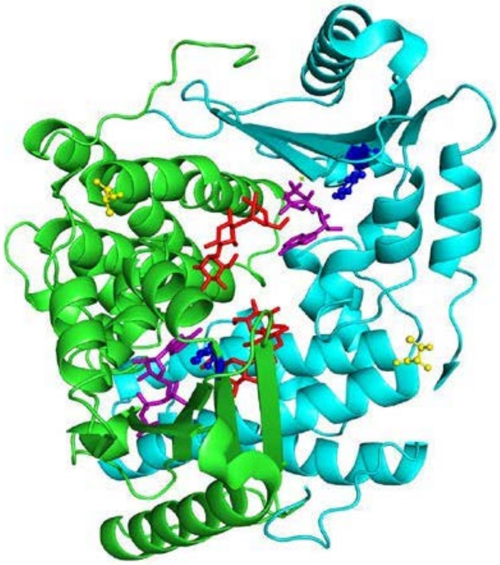


Figure I.1 Crystal Structure of variant D80Y.

PDB structure with bound ligands (Pedersen, Benning, & Holden, 1995), chains are colored in green and cyan. MgATP analog (purple) and kanamycin A (red) are bound to the active site, which is formed at the interface of monomers. Residues D80 (blue) and T130 (yellow) are shown as ball and stick model.

Methods

Molecular Dynamics

Systems were constructed based on the crystal structure of the D80Y species (Protein Data Bank ID: 1KNY) using Molecular Operating Environment (MOE, version 2012, Chemical Computing Group, Ltd, Montréal, Canada). The co-crystalized ligand and cofactor were removed from the model such that the apo form of the wild type, and T130K species were built by performing the correspondent mutations on the crystal structure of D80Y. Each structure was explicitly solvated with the TIP3P water model in a cubic box of 8 nm x 8 nm x 8 nm. Periodic Boundary Conditions in all directions were applied with electrostatic type Fast smooth Particle Mesh Ewald (PME). A18,000-step energy minimization was performed using the steepest decent algorithm. Molecular Dynamics simulations were carried out using the Gromacs 4.6.1 (Berendsen, Vanderspoel, & Vandrunen, 1995; Hess, 2008) simulation engine and the AMBER-f99sb (Hornak et al., 2006) force field. For each species at different temperatures, 300K, 322K, and 330K, a 50 ns equilibration with a 2fs integration timestep was performed in the isothermal-isobaric ensemble (NPT), and 1 bar pressure using the Nosé-Hoover temperature control (Hoover, 1985; Nose, 1984) and the Berendsen weak coupling pressure control (Berendsen, Postma, van Gunsteren, DiNola, & Haak, 1984). Finally, a 100 nanoseconds production run was performed using the Panirello-Rahman algorithm (Parrinello & Rahman, 1981). Atomic coordinates of the trajectory were saved on disk every 5 ps.

Principal Component Analysis

The trajectories obtained from the production run were analyzed using a built-in Gromacs tool, Principal Component Analysis (PCA). This method allows identification of the main modes of motions in the protein, highlighting the different conformational changes in the molecule (Tournier & Smith, 2003). Since these main motions can be described by the first principal components, this method is also called "essential dynamics” (Amadei & Limddrn, 1993). As a first step before applying this method, the trajectory has to be preprocessed; rotation and global translational motions must be removed by carrying out a coordinate root mean square best-fit to a reference structure, for instance, an average structure. Then, the coordinates as a function of time are stored in the matrix R with 3N rows containing the coordinates of the N atoms and M columns holding successive time points of the trajectory. Next, the covariance matrix C is calculated as in equation 1.1, for every atom from the group subject to analysis. Then, matrix C is diagonalized to obtain the set of eigenvectors V and their associated eigenvalues λ_i.

$$C = \frac{RR^T}{3N} \qquad (1.1)$$

$$\mathrm{diag}(\lambda_1, \lambda_2, \ldots,\lambda_n) = V^T C V \qquad (1.2)$$

The 3N-6 eigenvectors in V describe the orthogonal concerted motions of the protein. The variance in the original data is given by the eigenvalues and the direction of this variance by their associated eigenvectors in equation 1.2. Thus, any conformation of the protein whose coordinates are in the cartesian space $R\ (\bar{r}_1, \bar{r}_2, \ldots \bar{r}_N)$ maps onto a point $Q\ (\bar{q}_1, \bar{q}_2, \ldots \bar{r}_{3N-6})$ in the eigenvector space V. The q_i coefficients are the projections of R onto the eigenvectors space. This is just a transformation of the coordinate system, thus, the original and projected coordinates hold back the same information. To map a particular point Q from the eigenvector space to the cartesian coordinate system R it is necessary to calculate the average structure $<S>$ in the R space and add the appropriate linear combination of q_i and the eigenvector basis $\bar{v}_i$, as shown in equation 1.3

$$R =< S > + \sum_{i=1}^{3N-6} q_i . \bar{v}_1 \qquad (1.3)$$

In order to identify the main modes of motion in the protein, the eigenvalues should be sorted in descending order, then, plotting these eigenvalues against their indexes will commonly show that only the few first eigenvectors, those with the largest magnitudes, are responsible for the major motions in the protein. Therefore, the variance of the protein structure and dynamics can be described by just a few modes, the motions along their associated eigenvectors dominate the dynamics of the protein and contain most of the global dynamic information. To capture conformational changes in the secondary or tertiary structure of the protein it is necessary and sufficient to analyze the C_α atoms of the molecule, those atomic coordinates as a function of time are stored in matrix R, as in the equation 1, which is used to generate the covariance matrix C. PCA was performed on the trajectory obtained from the MD simulations to identify the main dynamic modes for the three species, WT, T130K, and D80Y at three different temperatures: 300K, 322K, and 330K.

Free Energy Landscape of the Structural Ensemble

While it is true PCA identifies the main motions of the structures and how they vary with species and temperature, it is not enough to analyze how the protein structure ensemble changes as a result of mutations and/or increase of temperature. In order to examine the change in the ensemble of conformations a method to characterize the different conformations must be selected. Usually conformers can be characterized by using geometric parameters, distances, angles, root mean square deviations, etc., however, this kind of parameters usually misses changes in other regions of the protein that could define a new conformation (Hall, Kaye, Pang, Perera, & Biggin, 2007). Principal components *[v1,v2,…,vN]*, on the other hand, can detect concerted motions, even though they are not parallel or antiparallel, giving an appropriate representation of the protein dynamics. Thus, the full set of eigenvectors over the trajectory can describe the different conformations sampled during the MD simulation. In general, when a set of parameters $S = [s_1, s_2, \ldots, s_n]$, at the temperature T, can depict a particular configuration of the system, it is possible to calculate the probability function $P(s_1, s_2, \ldots, s_n, \mathrm{T})$ from a histogram of the MD trajectory for each combination of values of $\mathrm{s_i}$ at the temperature T. The free energy landscape (FEL) of a system like this is a ***potential of mean force*** (PMF), $\Delta W_{(\mathrm{s1,s2,\ldots,sn,T})}$, (Grubmüller & Tavan, 1994; Rice & Gray Peter, 1965):

$$\Delta W_{(\mathrm{s1,s2,\ldots,sn,T})} = -\, k_{\beta}\; T\, \left[\ln P_{(s1,s2,..,sn,T)} - \ln(P_{max})\right] \qquad (1.4)$$

Where k_{β} is the Boltzmann constant, and P_{max} the maximum probability in the distribution, which is included in the equation as subtracting to make sure that $\Delta W_{(\mathrm{s1,s2,\ldots,sn,T})} = 0$ for the lowest free energy value. Although, the high-dimensional space S is needed to characterize protein conformations, it is possible to reduce these dimensions and only use those parameters that describe the majority of the change in the structure. Thus, in this particular case, given the first two eigenvectors $[e_1,\ e_2]$, the projections of the data r onto them will $[v_1,v_2]$ will be employed as *conformation coordinates* at temperature T to calculate the probability function $P_{(v1,v2,T)}$ and PMF $\Delta W_{(v1,v2,T)}$, Equation 1.5, of the distribution of the conformers sampled by the MD simulations (Grubmüller & Tavan, 1994; Mu, Nguyen, & Stock, 2005; Papaleo, Mereghetti, Fantucci, Grandori, & De Gioia, 2009).

$$\Delta W_{(v1,v2)} = -\, k_{\beta}T\, \left[\ln P_{(v1,v2,T)} - \ln(P_{max})\right] \qquad (1.5)$$

From these equations, it is clear that low values of PMF, ΔW, correspond to high probability configurations, which implies that high probability regions in the conformational space are more stable thermodynamically than regions with low probability. Thus, for this particular case at a given temperature T, a free energy landscape can be generated as a function of the first two eigenvectors. Figure I.4 shows the FEL for the three species at three different temperatures.

Results and Discussion

Principal Component Analysis of the MD trajectories

Analysis of the 300K simulation data suggests that the global dynamics of D80Y is dominated by the first dynamic mode (Figure I.2 right panel), whereas at 322K and 330K the first two modes both contribute significantly to the global protein dynamics. The dynamics of the WT and T130K species are mainly due to the contribution of the first two modes at all temperatures studied here as shown in Figure I.2 (left and middle panels), although the first mode of motion contributes more at 330K. The orientation and amplitude of the first two modes for WT, T130K, and D80Y at 300K are shown by vectors in Figure I.3 (left panel). The origin of these vectors indicates the region of the protein undergoing motion in those particular vectors' direction. The left set in each set of these figures show the motion due to the first mode of the protein. The first mode corresponds to the same dominant movement in the three species: an open/close "breathing" motion. The difference between the dynamics of the three species at different temperatures originates from the second dynamics mode, the WT and T130K variants display similar contributions in magnitude, but the regions impacted by this mode are different as well as the direction of the motion (right sets in each panel of Figure I.3). The D80Y species, exhibit a significant contribution only at 322K and 330K from this mode of motion. These results suggest that the first mode of motion has the same direction at all temperatures for all the three species (left sets in each panel of Figure I.3). The difference seems to arise from the second mode, whose vectors have different origins and directions depending on the species (right sets in each panel of Figure I.3). For instance, at 330 K, D80Y exhibits a very different second dynamics mode, magnitude and directions are different compared to those form WT or T130K species (right sets in the right panel of Figure I.3). These would suggest these point mutations, instead of causing local structural changes, actually affect the global dynamic properties of the enzyme, which in turn characterizes the mesophilic/thermophilic features of this protein.

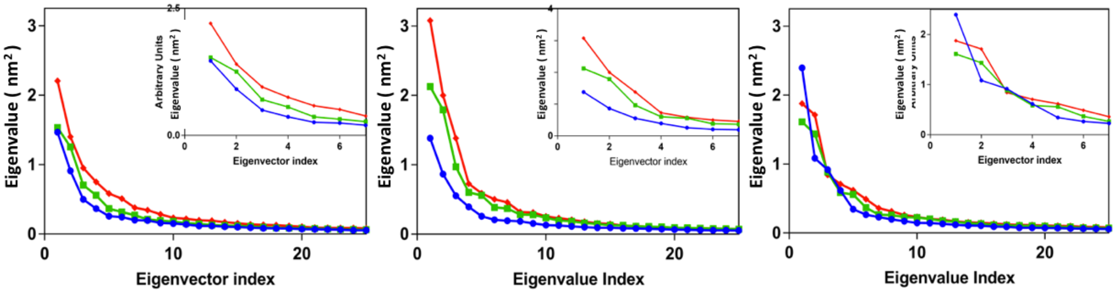
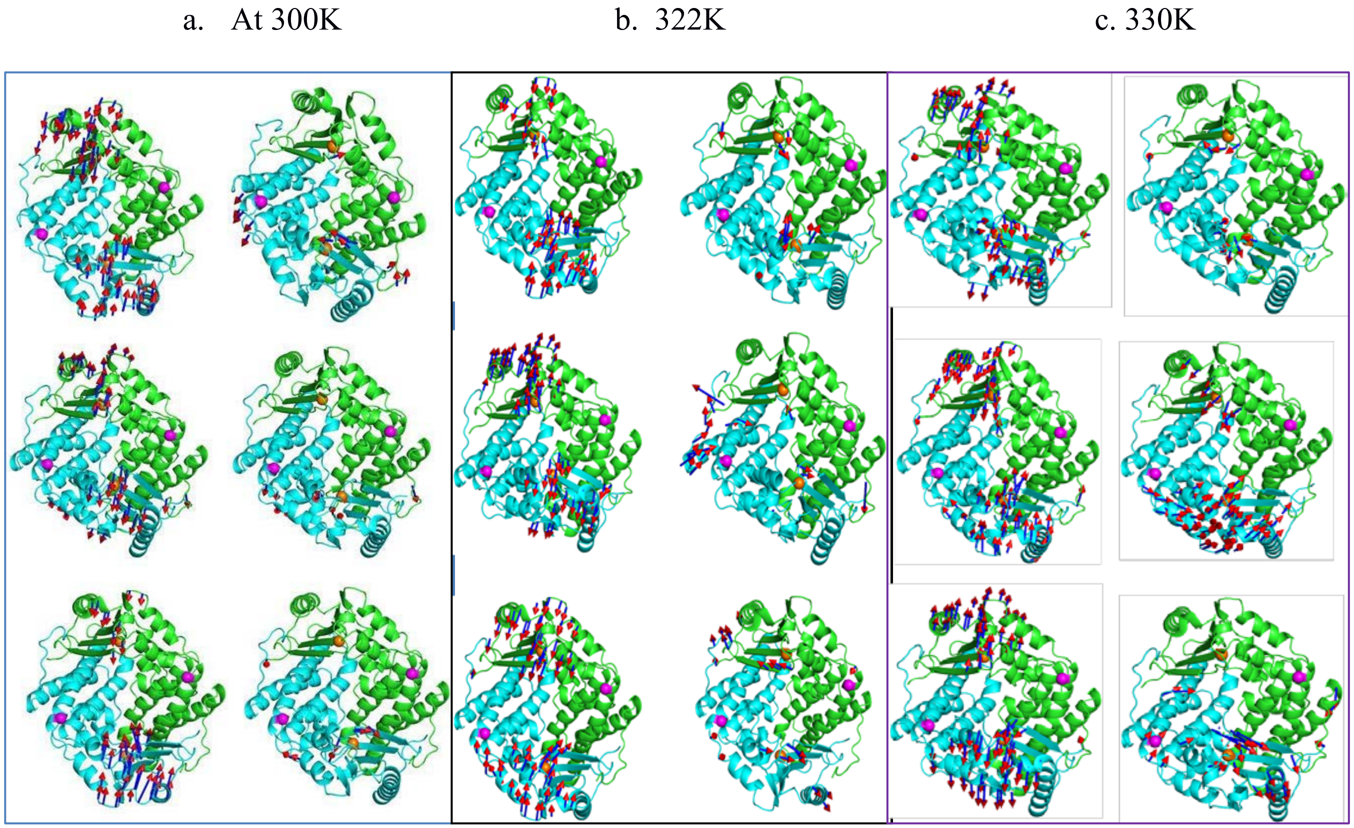


Figure I.2. Principal modes of motion of ANT.

Principal modes of motion projected onto the first 25 eigenvalues calculated from PCA in MD trajectories at different temperatures; Left to right are WT, T130K, and D80Y, respectively. Filled circles (blue), squares (green) and diamonds (red) represent 300K, 322K and 330K respectively. The insets show expanded regions of the several initial modes of motion.



a. At 300K

b. 322K

c. 330K

Figure I.3. First and second mode of motion of the ANT structure.

The arrows, obtained from PCA, represent the first (left panel) and second mode (right panel) of motions respectively. Top to bottom are: WT, T130K, and D80Y. C-α atoms of Asp80 and Thr130 are represented by orange and purple spheres, respectively.

Free Energy Landscape

In the previous section the first two modes of motions, PC1 and PC2, were used to characterize the directions and magnitude of the motion of the enzyme, the different values that these parameters can take are also useful to characterize the different conformational changes that ANT can adopt. These components are useful to obtain a two-dimensional free energy landscape according to Equation 1.5, at different temperatures and for each species analyzed here. Figure I.4 shows how different values of PC1 and PC2 lead to new conformational states of the protein, forming clusters of conformations more thermodynamically stable as temperature increases in the three species. The blue regions represent the most likely conformations at a particular temperature, intermediate states are cyan and green, while red regions are unaccessible states of the protein. WT at 330K shows that the most stable state is around PC1= PC2 = 0, the blue spot, thus the stable conformations are grouped in just one cluster, while at 322K, the breakup of the clusters is already noticeable, and at 330K there are already two main clusters containing the most stable conformations of the protein. The variant T130K, on the other hand, seems to have two clusters at 300K, and this starts splitting into two diffuse clusters at 322K. Finally three less populated stable clusters seem to come up at 330K. The D80Y species shows a specific behavior at 300K. There are three clusters, two of them well populated, the third one less populated, but still well defined. This pattern changes when temperature increases to 322K, with a single cluster centered at PC1 = PC2 = 0, very similar to that of WT at 300K. At 330K this same cluster is visible, centered at the same values of PC1 and PC2, with a slight variation at PC1 = 1.0 and PC2 = -1.0. In other words, for this variant, the number of clusters decreases as temperature increases and seems to be more stable at 322K than it is at 300K.

Analyzing Figure I.4, to investigate the effect of the mutations at a given temperature, it appears that: **at 300K** the D80Y variant has three clusters, stable configurations, while T130K shows two stable clusters, and WT only one cluster representing thermodynamically stable conformations, at this temperature all of those conformations in this cluster are thermodynamically stable. **At 322K**, the WT species shows a cluster of conformations that would be thermodynamically very stable (blue region), according to equation 1.5, however, there is a population of structures that, though less stable, are still part of the ensemble (cyan region in the panel). The T130K species displays one very populated cluster, however, there is also a nascent set of stable conformations that eventually could become populated if the MD simulations were longer. These two species, WT and T130K, seem to split the clusters with respect to their ensemble at 300K probably due to their melting temperatures being 314K and 322K, respectively. This implies that their structures at room temperature will undergo changes when temperature increases to 322K, forming other clusters. On the other hand, the cluster of stable structures is more spread out for D80Y, PC1's range is [-5.0; 4.5] and PC2's [-2.5; 2.5], suggesting that the transition between conformations happen through lower free energy barriers than in the other species.

The situation **at 330K** is quite different, WT displays two well defined clusters, while the T130K species possess one very populated cluster with two additional less populated clusters, suggesting that this mutation enables the protein to keep one stable cluster while shaping two other sets of conformations that may become more populated as the sampling of the MD simulation increases. The case of D80Y is different; the mutation in this case enables the protein to keep a centered cluster very similar to that of WT's at 300K. However, it also shows some nascent separation at PC1=1.8 and PC2= -1.8, which might be explained by the 329K melting temperature of this species.

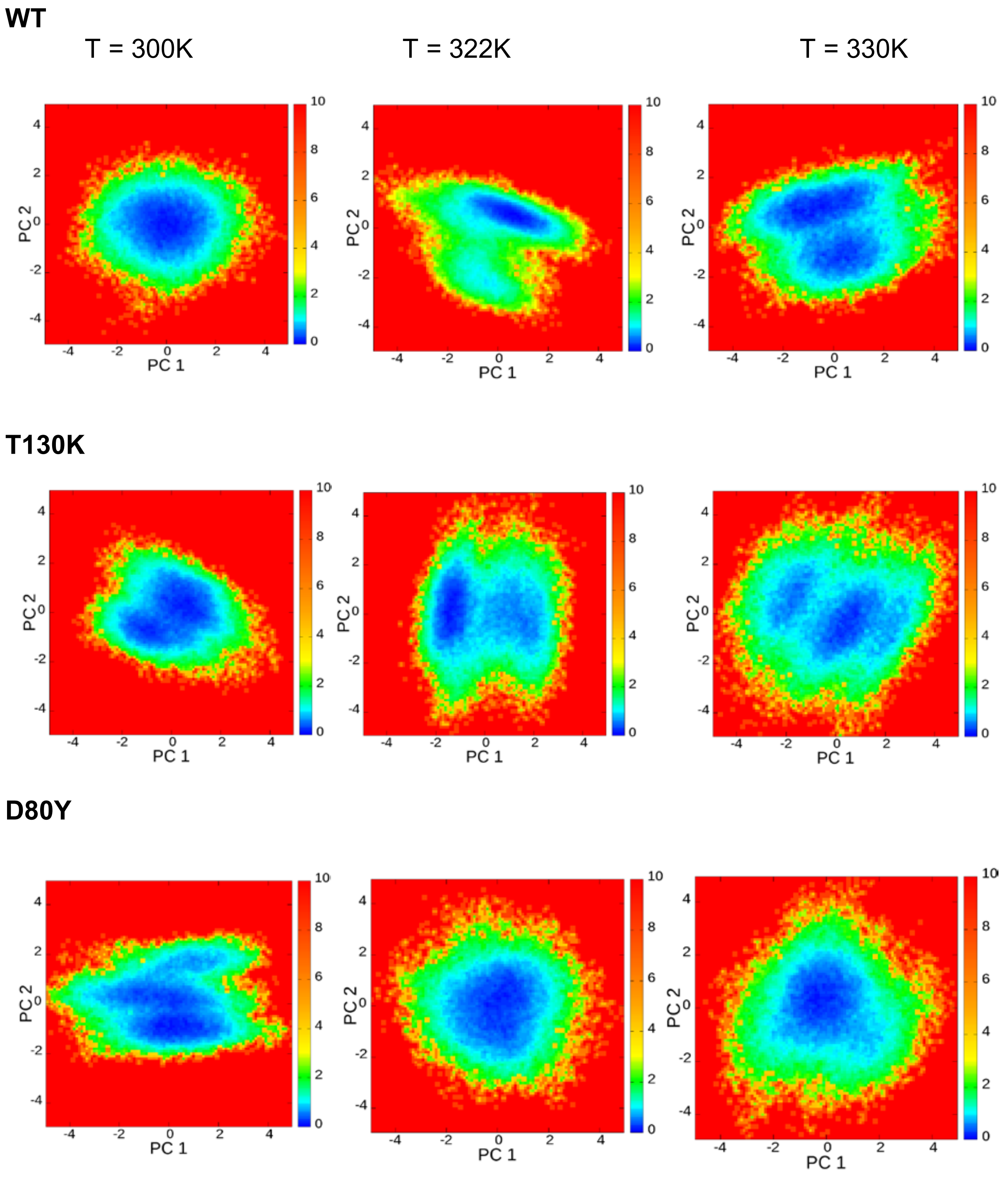


Figure I.4. Free energy landscape for the three species of ANT.

Energy landscape along the first two modes of motion for each species at the three temperatures. Blue regions represent high-probability configurations, and therefore more thermodynamically stable conformations. Yellow and orange regions denote less populated configurations.

Conclusions

Thermophilic proteins use different strategies to reach thermal adaptation. Comparisons between mesophilic and thermophilic proteins have been published, and difference between these enzymes have been attributed to side-chain hydrogen bonds, salt bridges, and internal hydrophobic packing (Dominy, Minoux, & Brooks, 2004; Elcock, 1998; Missimer et al., 2007; Xiao & Honig, 1999). It has also been suggested that water-protein surface (Sterpone, Bertonati, Briganti, & Melchionna, 2009) and protein conformational flexibility (Kalimeri, Rahaman, Melchionna, & Sterpone, 2013) are crucial factors for thermostability. There are also reports focusing on how the type of amino acids would strengthen local interactions and cause thermal stability. At this point, due to the diversity of these features, it is difficult to sketch out a common and unique mechanism that explains how thermophilic proteins keep their stability at temperatures above 300K. Our findings here suggest that mutations bring on global effects in the protein flexibility affecting the distribution of conformations in the ensemble. This change in the distribution is different for each species of ANT and does originate from changes of the global dynamics of the protein rather than from punctual, localized and specific non-bonded interactions. This would correlate with the idea that cooperative networks might be responsible for imposing restrictions to protein flexibility (Henzler-Wildman & Kern, 2007). In such case, residues D80 and T130 are critically positioned nodes of such a network.

References

Berendsen, H. J. C., Vanderspoel, D., & Vandrunen, R. (1995). GROMACS - A message-passing parallel molecular-dynamics implementation. *Computer Physics Communications, 91*(1-3), 43-56. doi:10.1016/0010-4655(95)00042-e

Danson, M. J., Hough, D. W., Russell, R. J., Taylor, G. L., & Pearl, L. (1996). Enzyme thermostability and thermoactivity. *Protein engineering, 9*, 629-630. doi:10.1093/protein/9.8.629

Dominy, B. N., Minoux, H., & Brooks, C. L. (2004). An electrostatic basis for the stability of thermophilic proteins. *Proteins: Structure, Function and Genetics, 57*, 128-141. doi:10.1002/prot.20190

Elcock, A. H. (1998). The stability of salt bridges at high temperatures: implications for hyperthermophilic proteins 1 1Edited by B. Honig. *Journal of Molecular Biology, 284*, 489-502. doi:10.1006/jmbi.1998.2159

Fields, P. A. (2001). Review: Protein function at thermal extremes: Balancing stability and flexibility. *Comparative Biochemistry and Physiology - A Molecular and Integrative Physiology, 129*, 417-431. doi:10.1016/S1095-6433(00)00359-7

Hall, B. a., Kaye, S. L., Pang, A., Perera, R., & Biggin, P. C. (2007). Characterization of protein conformational states by normal-mode frequencies. *Journal of the American Chemical Society, 129*, 11394-11401. doi:10.1021/ja071797y

Henzler-Wildman, K., & Kern, D. (2007). Dynamic personalities of proteins. *Nature, 450*, 964-972. doi:10.1038/nature06522

Hess, B. (2008). P-LINCS: A parallel linear constraint solver for molecular simulation. *Journal of Chemical Theory and Computation, 4*(1), 116-122. doi:10.1021/ct700200b

Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A., & Simmerling, C. (2006). Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins-Structure Function and Bioinformatics, 65*(3), 712-725. doi:10.1002/prot.21123

Kalimeri, M., Rahaman, O., Melchionna, S., & Sterpone, F. (2013). How conformational flexibility stabilizes the hyperthermophilic elongation factor G-domain. *Journal of Physical Chemistry B, 117*, 13775-13785. doi:10.1021/jp407078z

Lacey, R. W., & Chopra, I. (1973). Genetic Studies of a Multi-resistant Strain of Staphylococcus aureus. *Journal of Medical Microbiology, 7(2)*, 235-243.

Liao, H., McKenzie, T., & Hageman, R. (1986). Isolation of a thermostable enzyme variant by cloning and selection in a thermophile. *Proceedings of the National Academy of Sciences of the United States of America, 83*, 576-580. doi:10.1073/pnas.83.3.576

Matsumura, M., & Aiba, S. (1985). Screening for thermostable mutant of kanamycin nucleotidyltransferase by the use of a transformation system for a thermophile, Bacillus stearothermophilus. *Journal of Biological Chemistry, 260*, 15298-15303.

Missimer, J. H., Steinmetz, M. O., Baron, R., Winkler, F. K., Kammerer, R. A., Daura, X., & van Gunsteren, W. F. (2007). Configurational entropy elucidates the role of salt-bridge networks in protein thermostability. *Protein Science, 16*, 1349-1359. doi:10.1110/ps.062542907

Pedersen, L. C., Benning, M. M., & Holden, H. M. (1995). Structural investigation of the antibiotic and ATP-binding sites in kanamycin nucleotidyltransferase. *Biochemistry, 34*, 13305-13311. doi:10.1021/bi00041a005

Sterpone, F., Bertonati, C., Briganti, G., & Melchionna, S. (2009). Key role of proximal water in regulating thermostable proteins. *Journal of Physical Chemistry B, 113*, 131-137. doi:10.1021/jp805199c

Xiao, L., & Honig, B. (1999). Electrostatic contributions to the stability of hyperthermophilic proteins 1 1Edited by J. M. Thornton. *Journal of Molecular Biology, 289*, 1435-1444. doi:10.1006/jmbi.1999.2810

CHAPTER II

THE ROLE OF PROTEIN STRUCTURE ENSEMBLE IN GPCR-DRUG BINDING.

A version of this chapter is being prepared for submission by Wilfredo Evangelista Falcón, Jeremy C. Smith, and Jerome Baudry.

The work and writing presented in this chapter was done by Wilfredo Evangelista Falcón.

Abstract

The development of structural biology has provided enough data to apply *in silico* molecular docking techniques in early stages of drug discovery, drastically reducing the cost and time involved in wet-lab experiments. This has been facilitated by completion of the Human Genome Project, which has uncovered many targetable and off-targetable receptors. In principle, docking techniques can not only be used in potential drug targets, but also to predict the interaction of drug candidates with possible off-target receptors. The goal of predicting adverse drug reactions (ADR) for novel drug candidates is becoming a realistic objective. In this work, we present the basis of a reliable framework for high-throughput ensemble-based docking which allows protein-drug interaction predictions with statistical significance and consequently reduces the amount of false positive and/or negative hits. Crystal structures and representative conformations of clustered trajectories for eight GPCRs were examined, four of these proteins were subjected to a subsequent screening on each conformational state of their respective trajectories; showing that virtual screening is more efficient, when it is performed on a dynamic ensemble of target conformations than on a single crystal structure.

Introduction

In silico experiments on Adverse Drug Reaction

Molecular docking has been a very helpful tool to analyze, validate, and predict binding of small molecules on proteins (Cerqueira et al., 2015; Taboureau, Baell, Fernández-Recio, & Villoutreix, 2012). Its utility has increased with the availability of super computers, protein structural data as well as the availability of large databases of small molecules. The conformational selection paradigm has been the scaffold for ensemble- docking under the premise that ligands will bind specific protein conformations other than the basal crystal or NMR structures (Ellingson, Miao, Baudry, & Smith, 2015; Evangelista et al., 2016; Meng, Zhang, Mezei, & Cui, 2011). In order to obtain these other conformations, molecular dynamics (MD) techniques are frequently employed to sample accessible states with a reasonable computing time. In addition, the same approach could be employed to identify, analyze, or dismiss off-target protein binding. Off-target proteins are responsible for adverse drug reactions (ADR) exhibiting moderate to lethal effects. However, ADR caused by a new drug candidate (Lounkine et al., 2012) typically appear at the pre-clinical or clinical trials, and in a significant number of cases, candidates have to be removed from the market due to reported ADRs (Bender et al., 2007; Pirmohamed, Breckenridge, Kitteringham, & Park, 1998). Identifying off-target interactions in the early stages of the drug discovery process is hence an important goal, even though testing a very large number of proteins involved in the different metabolic pathways is almost impossible. However, a panel of 44 proteins (Bowes et al., 2012) has been proposed as responsible for about 75% of ADRs. This panel contains: 24 GPCRs, eight ion channels, six intracellular enzymes, three neurotransmitter transporters, two nuclear hormone receptors, and one kinase. Bowes' work implies that if a new potential drug is discovered or designed, it must not bind to, or have a very low affinity for these 44 proteins, in order to minimize the rise of ADRs. This set of proteins can be a very good starting point to predict toxicity *in silico* via ensemble-based docking simulations. However, this requires not only knowledge of the 3D structure of the receptors, but also of the ensemble of conformations for each of these proteins. Moreover, it also requires the design of a reliable procedure to measure the performance of the virtual screening. The most common statistical measures to identify the best receptor structures have been summarized and described by Huang et. al. (Huang & Wong, 2016).

G-Protein Coupled Receptors

Most of those 44 proteins do not have a complete experimental structure. The initial testing set in this work comprises the eight GPCR structures listed in Table 1. GPCRs, also named heptahelical receptors, seven-transmembrane (7-TM) receptors, or guanine-nucleotide-binding protein-coupled membrane receptors, are expressed in eukaryotic organisms, and control a high number of regulatory processes. Since GPCRs are located in the plasma membrane, they are accessible to not only their natural ligands, but also to drugs, both antagonists and agonists. An important characteristic of these proteins is the non-uniformity of expression in different cell types and tissues, which provides special model of selectivity (Insel, Tang, Hahntow, & Michel, 2007). These proteins are basically switches that activate various responses after receiving some stimulus. Once the ligand binds the receptor, a conformational change is performed and the GPCR will recruit, through structural changes in the intracellular domains, a G-Protein to the inner leaflet of the cellular membrane. This mechanism of signal transduction is crucial for many processes in different tissues. Their malfunction might result in different kinds of monogenic diseases, single defective genes in the autosomes, or genetic mutations (Insel et al., 2007; Meyer, 2000). Alteration of GPCRs in number or structure/function will lead to disorder in cellular signal transduction: up-regulation/hypersensitivity, Down-regulation/desensitization, or receptor gene mutation. This is why fine-tuning is crucial for GPCR’s functions and why it is one of the main targets of drugs.

To date, all published GPCRs structures share a common 7-TM α-helices domain, as shown on Figure II.1, and are located within the cellular membranes of different organs (Palczewski & Orban, 2013). In spite of their almost identical structure, there are several subtypes based on extracellular domain topology, on the type of G-protein they activate, on activating ligands, on sequence similarity, or on function (Park, Lodowski, & Palczewski, 2008).

Statistical Measurement of Performance

In order to sample conformational states beyond the experimental structures obtained from the Protein Data Bank (PDB) (Berman et al., 2000), Coarse-Grained (CG) Molecular Dynamics (MD) simulations were ran to generate one microsecond trajectories for each of the eight GPCRs listed on Table 1. Docking calculations were then performed on representative clusters for each protein. In a second set of docking calculations, 3000 structures obtained from the trajectories for each protein, i.e. without structural clustering, were used in docking calculations. The total number of docking calculations on representative structures and full trajectories was about 2.3 million and 127.5 million, respectively.

Here, ensemble-based docking is used to screen many protein conformational states against a ligand library of binding molecules (actives) and theoretically non-binding molecules (decoys). The question that emerges is how to discriminate between snapshots with significant contribution to conformational selection against those with no significant contribution. In this work, I propose a statistical metric identify receptor structures that significantly bind more active ligands than a random selection of ligands. This method is based on Exploratory Data Analysis, developed in 1977 by Tukey, to detect outliers based on the interquartile (IQR) values (Tukey, 1977).

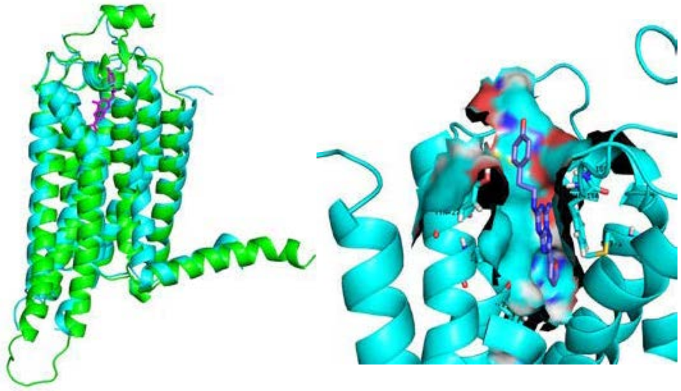


Figure II.1. Crystal Structure of ADORA2A.

Left Panel: In cyan, crystal structure of the ligand-bound form. PDB ID:3EML. In green, modeled structure with intracellular loop and C-terminal completed.

Right Panel: Zoomed-in view of the binding site of the protein, in cyan, surface and ribbon representation; and the ligand in purple bond-stick representation, 4-[2-[(7-amino-2-furan-2-yl[1,2,4]triazolo[1,5-a][1,3,5]triazin-5-yl)amino]ethyl]phenolthe, DrugBank ID: DB0877.

Methods

Collecting Structural data

GPCRs structures were downloaded from the RSCB Protein Data Bank, PDB IDs are listed in Table 1. Ligands structures, actives and decoys, were downloaded from DUD-E E (Mysinger, Carchia, Irwin, & Shoichet, 2012) for ADORA2A and ADRB2. CC-DD database (Gatica & Cavasotto, 2012) was used to obtain the ligands for CHRM2, HTR1B, HTR2B, OPRD1, and OPRK1, as summarized on Table II-1.

Preparation of Protein Structures for MD Simulations

The GPCR structures obtained from PDB are products of chimeric expression for crystallization, domains of the structures that did not belong to the WT GPCR sequence were deleted as well as co-crystallized ligands, and missing loops were modeled and built using MODELER 9.10 (Fiser, Kinh Gian Do, & Sali, 2000). In cases where the missing loop was in the opposite location of the binding site, in the inner side of the membrane, and the number of missing amino acids was more than 20, these loops were not built.

Sampling conformations via Coarse-Grained modelling & MD simulations

In order to form an ensemble of GPCR structures, we sampled at least 5,000 conformations for each protein belonging to one microsecond of MD simulation. To obtain this number of conformations in a reasonable computing time each protein was mapped to coarse grained (CG) models and placed in a bilayer membrane. The main components of the plasma membrane, as suggested by Leventis (Leventis & Grinstein, 2010), were included in both inner and outer leaflets: phosphatidylcholine (POPC), phosphatidylethanolamine (POPE), phosphatidylserine (POPS), and cholesterol (CHOL) at 42%, 25%, 14%, and 19%, respectively. Water and ions were added to equilibrate the system using *martinize.py* v2.5 and *insane.py* scripts available at http://cgmartini.nl/index.php/tools2/proteins-and-bilayers (Ingólfsson et al., 2014; Monticelli et al., 2008; Pierole & Marrink, 2013; Wassenaar, Ingólfsson, Böckmann, Tieleman, & Marrink, 2015) . Each of the systems, protein, membrane, ions, and water, was reduced from ~125,000 atoms to CG ~14,000 particles. Next, MD simulations were performed using Gromacs v5.1.0 (Berendsen, van der Spoel, & van Drunen, 1995) for 1µs, saving frames every 200 ps. The setting parameters for the energy minimization, equilibration, and production time were used as in (Stansfeld et al., 2015). In order to select groups of similar structures from the trajectories, Gromacs clustering tools and its built-in gromos method were used to build clusters based on the Root Mean Square Deviation (RMSD) of the backbone (Daura et al., 1999). The goal was to obtain a number of representative structures such that the docking calculations can be done in an affordable computing time on the Newton high performance computer cluster of The University of Tennessee, Knoxville. According to Table II-1, there is an average of 10,000 active and decoy small molecules per protein, thus, in order to generate 250,000 protein-ligand complexes in a reasonable time, about 25 clusters were calculated for each protein. Different RMSD thresholds were used to accomplish this purpose, for each protein given in Table II-1. Once the representative structures were identified, they were extracted from the trajectories using Gromacs tools, and back mapped to an all-atom model using Backward v0.1 (Wassenaar, Pluhackova, Böckmann, Marrink, & Tieleman, 2014). The second set of ensembles was composed of the all-atom models obtained from back-mapping the entire trajectory, 5,000 conformations, of the four proteins, ADORA2A, ADRB2, OPRD1, and OPRK1. These four proteins were selected because they have a greater number of significant frames in each subset analyzed after the docking on representative structures, except OPRK1, that has zero significant frames in the 1.0% subset, but it contains more representative structures than ADRB2 and OPRD1 in the other subsets (Table II-4). Thus, it was expected to get more significant frames for these proteins than for the other four proteins, HRH1, CHRM2, HTR1B, and HTR2B after the ensemble docking calculations on their 600 ns trajectories. CG-MD and back-mapping, were performed on the Moldyn cluster at the UT/ORNL Center for Molecular Biophysics, Oak Ridge, Tennessee.

Pre-docking preparation of receptors and ligands

VinaMPI, a high throughput docking program efficient on supercomputers and developed in our laboratory (Ellingson, Smith, & Baudry, 2013), requires input files in PDBQT format for both protein and ligands, and scripts from AutoDockTools (ADT) v.1.5.6 (Sanner, 1999) were used to pre-process the conformations obtained from the MD simulations. This pre-processing includes removing any atom other than the protein's and adding polar hydrogens atoms and Kollman charges. Ligands structures were pre-processed, adding hydrogens and charges, and rotamers were set according with the default method of ADT. The configuration files for the virtual screening, receptors and ligands lists, were produced with the Python scripts developed in the laboratory (Ellingson et al., 2013).

Docking on representative structures of clusters

The numbers of structures and ligands tested in this phase are listed in Table II-1. This phase of the project was performed on the Newton high performance computer cluster of The University of Tennessee at Knoxville. VinaMPI produced 2,292,040 combinations of protein-ligand complexes, each complex containing between 1 and 10 poses with corresponding calculated binding energies. The pose docked in the binding site with the most favorable binding free energy is selected as a hit.

Docking on each frame of the trajectory

Four proteins were selected for molecular docking and submitted for ensemble-based virtual screening to the super computer Titan at ORNL. The original idea was to use 5,000 frames of each protein, the same number that was used to obtain the representative structure after clustering, which were obtained after 1.0 us of CG-MD simulations, however, due to computing time limitations, only the first 3,000 conformations were tested instead. To this end, these conformers were prepared in the same way as the representative structures from the previous section. VinaMPI generated 164 million protein-ligand complexes, each with 10 poses and their corresponding binding energy, again the pose with the lowest energy is counted as a hit.

Statistical Measurement for Conformation's Performance

In order to establish a statistical threshold to decide whether docking results on a particular frame are statistically significant, i.e. if it is selected by active compounds beyond a random selection distribution, an outlier detection method was introduced in the analysis. The Intequartile method (Salgado, Azevedo, Proença, & Vieira, 2016; Tukey, 1977) defines the interquartile range (IQR) to set the lower and upper cut-off values Q1-1.5*IQR and Q3+1.5*IQR, respectively (see Figure left panel in II-2). Values below and above these thresholds are defined as outliers. An important feature of this method is that does not depend on the symmetry of the distribution, interqueatiles can be calculated on symmetrical and not symmetrical distributions. Thus, if the number of active ligands bound to a particular frame is higher than the upper cut-off, Q3+1.5*IQR, of a random distribution the docking results on this particular frame will be statistically significant and this frame will be counted as a "significant frame". Values of this upper limit for each set of ligands belonging to their respective proteins are shows in Tables II-3 – II-5 in the appendix section of this chapter. For instance, the ligand library for ADORA2A comprises 844 actives and 10,899 decoys. If 5% of this pool, e.g. 587 molecules, is randomly selected, after many assays, 42 actives and 522 decoys are expected in average, which then implies there is still a significant probability to obtain 50, 60, or 100 active compounds in any of the assays (Figure II.2, right panel).

Once the significant frames have been identified, their list of receptor-ligand complexes are merged together and ranked according to their binding energies. Then, duplicates of active ligands are removed from the list, leaving only those with the lowest free binding energy, ensuring that no duplicates are counted as number of hits in the ensemble.

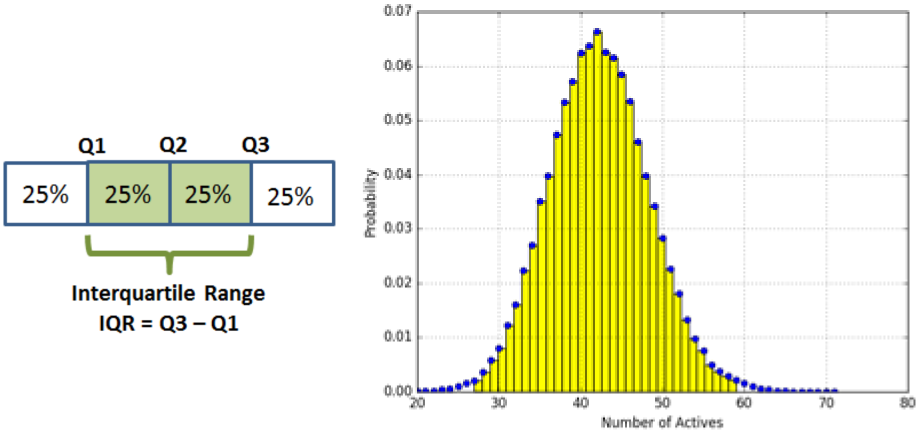


Figure II.2 Interquartile definition, and Probability distribution for random selection at 5.0% for ADORA2A.

Left panel: Any set of data can be divided in four quartiles, containing 25% of the data each. The interquartile range is defined as IQR = Q3-Q1.
Right panel: Random distribution for 5.0% (587 molecules, actives and decoys) of the ligands library of ADORA2A. After many assays, every time 587 molecules are selected from the library, in average 42 active ligands are obtained. The yellow-colored part of the distribution corresponds to a statistical random result and only a number of actives > 59 (Q3 +1.5*IQR) would be deemed as statistical significant of a non-random result.

Table II-1 Set of GPCR proteins and their number of selected ligands for this work.

Protein Name	Gene name	DUD-E		CC-DD		Clustering RMSD (A)	Number of Clusters	Number of Docking
		Actives	Decoys	Actives	Decoys			
Adenosine receptor A2A	ADORA2A	844	10899	443	17277	2	33	399262
β2-adrenergic receptor	ADRB2	447	15255	410	15990	2	18	298338
Histamine H1 receptor	HRH1	------	----------	86	3354	2	21	75680
Muscarinic acetylcholine receptor M2	CHRM2	------	----------	126	4914	1.9	32	166320
5-Hydroxytryptamine receptor 1B	HTR1B	------	----------	113	4407	2.3	35	162720
5-Hydroxytryptamine receptor 2B	HTR2B	------	----------	227	8853	2	36	335960
δ-type opioid receptor	OPRD1	------	----------	377	14703	1.75	32	497640
κ-type opioid receptor	OPRK1	------	----------	307	11973	2.25	28	356120

Results and Discussion

Docking on Clustered Structures

The IQR method was applied on four subsets of the ranked list for each frame, and on the crystal structure: i.e. the top 0.5%, 1.0%, 5.0%, and 10.0% of compounds predicted to bind. Interestingly, **ADORA2A** showed a dramatic improvement in binding active ligands compared to the average of the random selection or crystal structure at any percentage, see Figure II.3. In any subset, the crystal structure binds less or about the same number than a random selection of compounds, while docking experiments bind, by far, more active ligands, this likely because of the number of significant frames found in each subset, 3, 4, ,8, and 16 for 0.5%, 1.0%, 5.0%, and 10.0%, respectively, as shown in Table II-4. For instance, at 1.0% of the ranked list, the average of the random selection is 8 active compounds, while the ensemble binds 99 of them; on the other hand, in 5.0% and 10.0% subsets the enhancement reaches a remarkable 51% and 82% of the total active molecules, respectively, as shown by Table II-3. **ADRB2**'s crystal structure and ensemble bind more active compounds than random selection in any of the subsets as shown by Table II.2 and Figure II-3. However, the major enhance of the ensemble for this protein is achieved in the 10.0% subset, covering 17% and the total number of actives, Table II-3, in spite of there is just one significant frame at this percentage for this protein, Table II-4.

OPRD1, whose ensemble binds more active ligands than expected from the random selection, and as many active ligands as the crystal structure at any of the percentages analyzed, e.g. in the 1.0% subset the average of random selection is 4 actives, the docking experiments bound 13 active compounds. In the 10% subset, the expected random selection is 38, while the ensemble binds 88 active ligands, covering only 23.3% of the total number of active ligands, see Table II-3. In each case there is only one significant frame, Table II-4, and it is the same conformation at 12.8 ns. **OPRK1**'s crystal structure and ensemble represent an interesting case, because the crystal structure binds less active ligands than a random selection in all the subsets; while the ensemble does not contain significant frames in the 1% subset, however, in the 5% and 10% subsets the ensemble binds at least three times the random selection with 2 and 3 significant frames, Table II-4, covering 14% and 33%, respectively, of the total number of active compounds in the pool, as displayed by Table II-3.

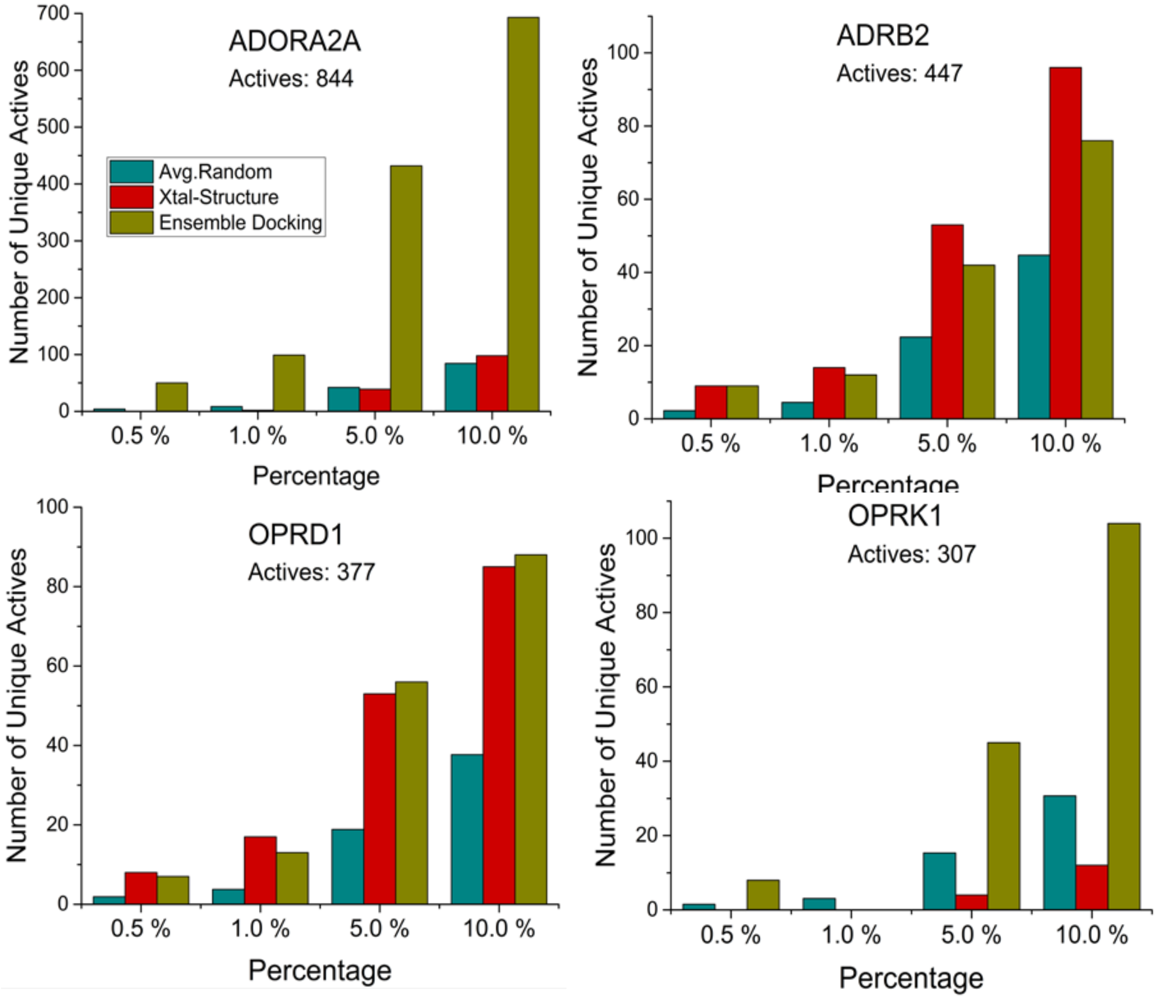


Figure II.3 Unique Actives docked in clustered conformations.

Table II-2 Number of unique active ligands selected via: Random Selection, docked in crystal structure, in representative structures of clustered frames, and in 3000 frames of 600 ns of MD trajectory.

Protein	Ligands			Average of Actives in Random Selection				Actives in Crystal Structure				Unique Actives in significant frames in clustered data				Unique Actives in significant frames in 600 ns of trajectory			
	Actives	Decoys	Total	0.5 %	1.0 %	5.0 %	10.0 %	0.5 %	1.0 %	5.0 %	10.0 %	0.5 %	1.0 %	5.0 %	10.0 %	0.5 %	1.0 %	5.0 %	10.0 %
ADORA2A	844	10899	11743	4	8	42	84	0	2	39	98	50	99	432	693	450	550	802	836
ADRB2	447	15255	15702	2	4	22	45	9	14	53	96	9	12	42	76	56	80	267	392
HRH1	86	3354	3440	0	1	4	9	3	6	19	26	0	6	0	16				
CHRM2	126	4914	5040	1	1	6	13	0	2	9	20	0	0	13	23				
HTR1B	113	4407	4520	1	1	6	11	0	0	4	13	0	0	0	0				
HTR2B	227	8853	9080	1	2	11	23	5	9	24	39	0	9	49	39				
OPRD1	377	14703	15080	2	4	19	38	8	17	53	85	7	13	56	88	79	125	243	276
OPRK1	307	11973	12280	2	3	15	31	0	0	4	12	8	0	45	104	58	69	168	247

Table II-3. Percentage of the Total Number of Actives bound by the Crystal Structure, Representative structures, and ensemble from trajectory.

Protein	Total of Actives	Percentage of Actives Bound by Crystal Structure in each subset				Percentage of Actives Bound by Cluster in each subset				Percentage of Actives Bound by Trajectory in each subset			
		0.5%	1.0%	5.0%	10.0%	0.5%	1.0%	5.0%	10.0%	0.5 %	1.0 %	5.0 %	10.0 %
ADORA2A	844	0.0	0.2	4.6	11.6	5.9	11.7	51.2	82.1	53.3	65.2	95.0	99.1
ADRB2	447	2.0	3.1	11.9	21.5	2.0	2.7	9.4	17.0	12.5	17.9	59.7	87.7
HRH1	86	3.5	7.0	22.1	30.2	0.0	7.0	0.0	18.6				
CHRM2	126	0.0	1.6	7.1	15.9	0.0	0.0	10.3	18.3				
HTR1B	113	0.0	0.0	3.5	11.5	0.0	0.0	0.0	0.0				
HTR2B	227	2.2	4.0	10.6	17.2	0.0	4.0	21.6	17.2				
OPRD1	377	2.1	4.5	14.1	22.5	1.9	3.4	14.9	23.3	21.0	33.2	64.5	73.2
OPRK1	307	0.0	0.0	1.3	3.9	2.6	0.0	14.7	33.9	18.9	22.5	54.7	80.5

Table II-4. Number of significant frames found after docking on clustered data and trajectory for each subset.

	Number of Significant Frames in Cluster				Number of Significant Frames in Trajectory			
Protein	0.5%	1.0%	5.0%	10.0%	0.5%	1.0%	5.0%	10.0%
ADORA2A	3	4	8	16	166	187	508	817
ADRB2	2	1	1	1	20	21	84	128
HRH1	0	2	0	1				
CHRM2	0	0	1	1				
HTR1B	0	0	0	0				
HTR2B	0	1	3	1				
OPRD1	1	1	1	1	33	34	35	41
OPRK1	2	0	2	3	71	23	40	70

Clusters obtained from CG-MD simulations of HRH1, CHRM2, HTR1B, and HTR2B do have at least one subset, in which does not contain significant frames. They were discarded for ensemble docking on trajectory of 3,000 frames.

Clusters obtained from ADORA2A, ADRB2, and OPRD1 have at least one significant frame in each subset; whereas, OPRK1 has none significant frames in 1.0% subset, however, this protein contains in total seven significant frames in the other subsets. These four proteins were then subjected to an ensemble docking on 3,000 conformations of a 600 ns trajectory from CG-MD simulations.

Number of significant frames of the four proteins selected for ensemble docking on the entire 600 ns trajectory increased remarkably compared to number of significant frames obtained in the representative structures of the clustered data.

Docking on non-Clustered Trajectory

Docking experiments were performed on 600 ns of trajectory without structural clustering four proteins: ADORA2A, ADRB2, OPRD1, and OPRK1. In total, 3000 conformations of each of these proteins were tested against their respective number of ligands according with Table II-2. The number of significant frames was dramatically increased with respect to the clustered data, i.e. in the 5.0% subset from only 8 significant frames to 500 as shown in Table II-4. This also leads to an increase in the number of unique actives docked in those significant frames, as shown in Tables II-2 and II-3, and Figure II.4. The **ADORA2A**'s ensemble showed again the best improvement of the ensemble-base docking with respect to a random selection or docking on the crystal structure at any of the percentages (see left upper panel in Figure II.4). The hits obtained in each 0.5%, 1.0%, 5.0% and 10.0% cover the 53%, 65%, 95%, and 99%, respectively, of the entire set of active ligands for this protein. This is pretty remarkable, since the ensemble binds a number of active ligands far higher than a random selection or the crystal structure, this is also due to the increase in the number of significant frames as shown in Table II-4. This implies that a new compound can be tested, and the probability to bind the protein will be assessed much better than using the crystal structure. The second case, **ADRB2** (right upper panel of Figure II-4) shows improvements in every subset respect to the crystal structure and up to 20 times with respect to a random selection. In 0.5% and 1.0% subsets the conformational states provided by the non-clustered trajectory bind 59% and 87% of total active molecules, 447, in the library for this protein, Table II-3.

The hits obtained by the **OPRD1**'s ensemble are by far higher than a random selection and show also improvements with respect to the clustered conformations and crystal structure (left lower panel in Figure II-4). In each of the subsets the ensemble provided by the non-clustered MD simulations covers 21%, 33%, 64%, and 73% of the entire ligands in the library for this protein, versus the 2%, 3%, 15%, and

23% covered by representative structures from the clustered simulation, Tables II-2 and II-3.

The **OPRK1** ensemble (right lower panel in Figure II.4) binds many more actives than the crystal structure at any of the subsets, these subsets cover 19%, 22%, 54%, and 80% of the active ligands library for this GPCR (Tables II-2 and II-3) which is quite remarkable if compared to the percentage of actives bound to the crystal structure or the representative structures provided by clustering.

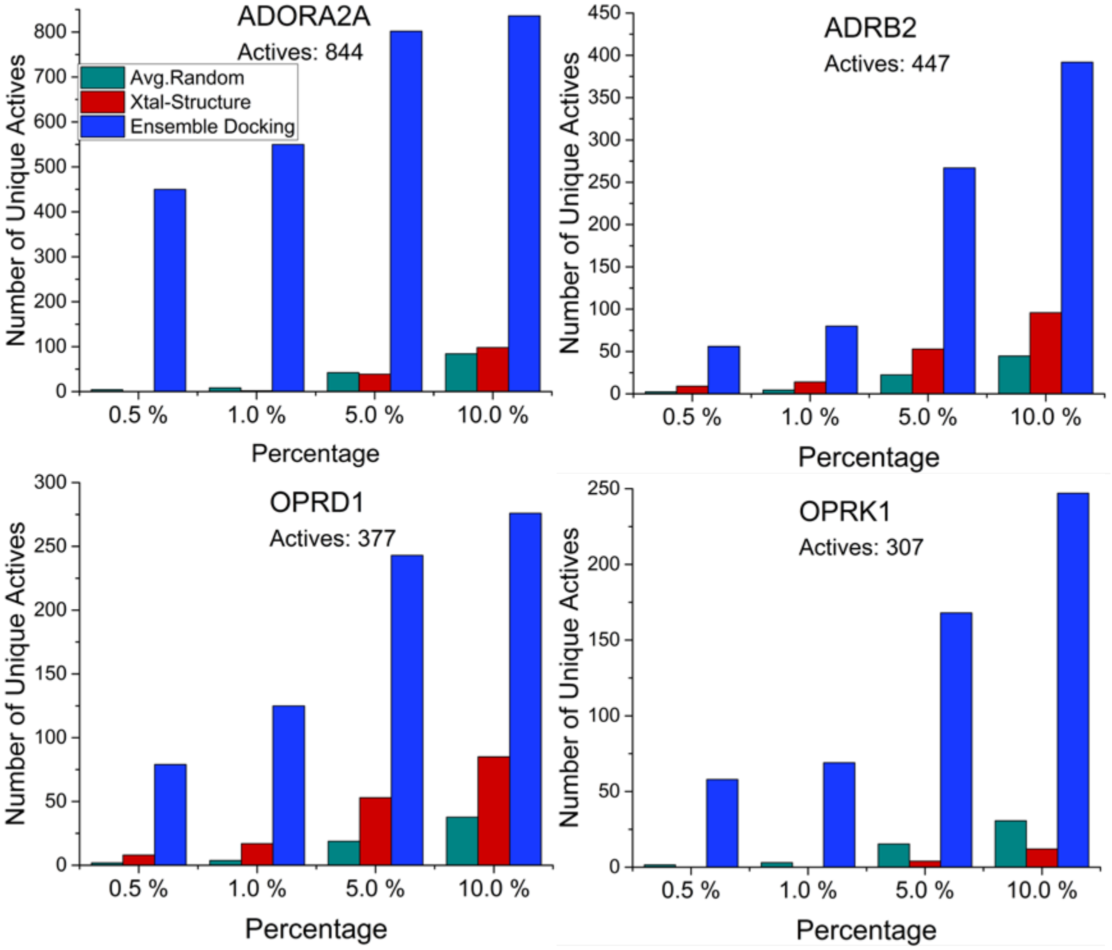


Figure II.4 Docking results on four GPCRS performed on frames from 600 ns of MD trajectory.

These results lead to the question, why are ensemble docking calculations more efficient for some GPCRs? The first explanation comes from the MD trajectories. Depending on the dynamics of each particular protein, the 600 ns of MD trajectory could sample a large variety of conformations, or not. One way to assess this is the atomic root mean square deviation values (RMSD) values of the whole structure in each case, shown in Figure II.5. This figure shows that **OPRD1** does not display as much structural variation as the other three proteins, the highest volume for OPRD1 is about 260 Å^3, while ADORA2A or OPRK1 samples conformations with volumes about 500 Å^3, generating less conformations potentially selected by ligands. This might explain why docking results from representative MD structures of this protein are very similar to results obtained using the crystal structure only, in all the subsets below 24%, (Table II-3). Docking results in the 10% subset for conformations in the trajectory identify 73% of the total actives for this protein. Less conformations sampled, less chances to be selected by ligands, implying that many compounds, about 27% of the total library for this GPCR, in the 10% subset of the ranking did not “find” the right conformation of the protein to form the complex.

ADORA2A, ADRB2, and OPRK1, on the other hand, exhibits RMSD variations greater than OPRD1, hence, the MD is sampling more conformational states. This would explain why these three proteins are selected by more active molecules at any of the percentages analyzed, as more conformational states are sampled, more active compounds will bind the protein. **ADORA2A** results are particularly interesting, the ensemble provided by clustering includes 33 structures and 16 of them are significant (Table II-4), these conformers can bind 82% of the total active ligands in the 10.0% subset. Even better, the trajectory’s ensemble binds 99% of the total ligands library for this protein when the equivalent subset, 10.0%, is analyzed (Table II-3), and 817 significant frames were found in this subset. **ADRB2**’s ensemble coming from clustering is the least efficient of the four proteins, this ensemble’s efficiency is worse than the crystal structure. In the first two subsets, 0.5% and 1.0%, the ensemble built for this protein exhibit 6-times more selected active compounds than a random selection, they only cover 12% and 17% of the entire library of active ligands for this protein. However, in the 10.0% subset, the ensemble obtained for this protein is the second most efficient, covering 87% of its ligands library (Table II-3), and up to 128 significant frames in the 10.0% subset. **OPRK1**'s ensemble presents the highest variation in the RMSD values, Figure II.5, and bind even less active molecules than OPRD1's ensemble for the 0.5%, 1.0%, and 5.0% subsets. Since the RMSD is higher in this case, it is supposed this structure is sampling more conformations, however, frames that were not significant in the mentioned subsets turned out to be significant in the 10.0% subset, reaching to cover 80% of the active molecules library for this GPCR, see Table II-3.

RMSD is an indicator of flexibility of the whole protein, and the lack or abundance of dynamics in the structure might be crucial for GPCR and its affinity for ligands, as suggested by (Lebon, Warne, & Tate, 2012; Shahane, Parsania, Sengupta, & Joshi, 2014). In general, there are several causes for variation of affinity between one GPCR and another, e.g. nature of residues in the binding site (Gether & Kobilka, 1998), chemical nature of ligands, cell membrane constituents (Ghanemi, He, & Yan, 2013). Small structural changes in the binding pocket could also be critical in the receptor function (Deupi & Kobilka, 2010). Thus, to get some insights about the possible contributions to the ligand binding of the volume of the binding site, binding site's volume was calculated on the 600 ns trajectory for the four systems using the trj_cavity software (Paramo, East, Garzón, Ulmschneider, & Bond, 2014). The results (Figure II.6) show that while ADORA2A, ADRB2, and OPRK1 sample conformations with a larger binding site than in their correspondent crystal structures, OPRD1 samples conformational states with a smaller cavity than in the crystal structure.

Another interesting feature about ADORA2A is that the volume of the cavity changes from about 200 Å^3 to 430 Å^3 at 250 ns (upper left panel in Figure II.6). This difference in flexibility could have some contribution to the fact that ADORA2A's ensemble can bind more active ligands than ADRB2 or OPRD1 ensembles in any of the analyzed subsets shows in Table II-3. However, OPRK1 samples the same variation in the volume of the binding pocket, lower left panel in Figure II.6, but even in the 10.0% subset, this ensemble cannot bind as many ligands as ADORA2A's ensemble does, and covers 80% of the total number of its active ligands in the library. Even though the binding site volume does not fully explain the difference in the number of ligands selecting these proteins, it might have some contribution to some of these cases.

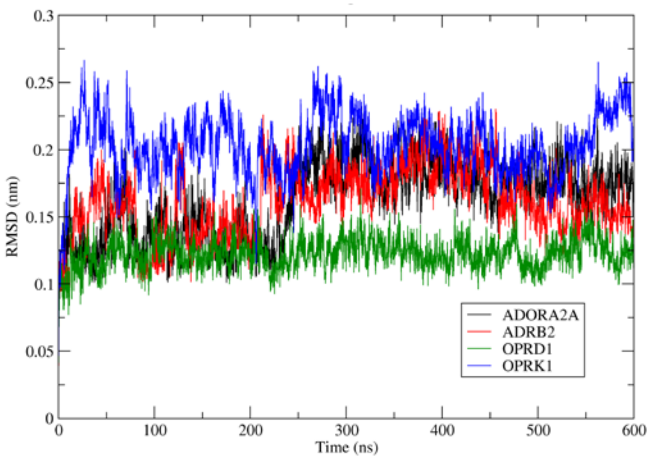


Figure II.5 RMSD of the α-Carbons for the four GPCRs as function of simulation time.

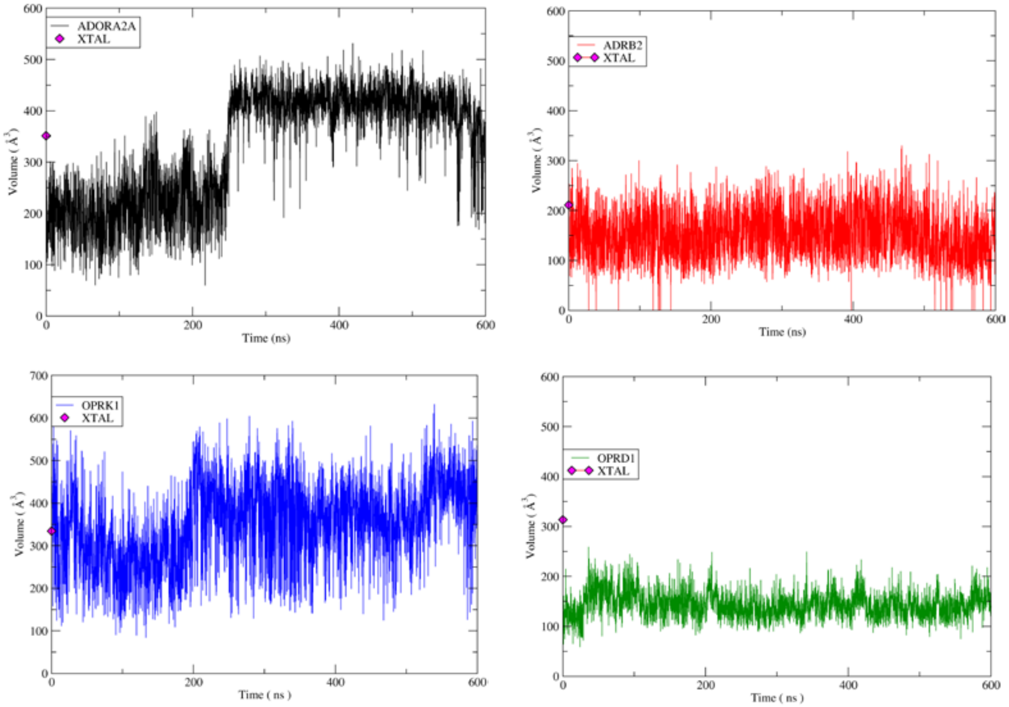


Figure II.6 Volume of binding site for each of the four systems as function of time.

Conclusions

Ensemble-base docking as a tool for virtual screening has been shown to be a promising technique to predict protein-drug interactions. The results here show that ensemble-based docking on conformations from a trajectory of MD simulations leads to improvements in predicting ligand binding with respect to docking calculations on single crystal structures. Importantly, the procedure proposed here discriminates statistically between conformational states are selected by ligands above a random selection of compounds. Thus, prediction of ADR via ensemble-based docking is foreseen as a feasible method once structural significant species are identified. Still yet, its high computational cost might make this technique too much expensive for daily use, e.g. docking calculations on ADORA2A's 3,000 conformations against about 11,7000 ligands required approximately 8 million hours-processors, which is still unaffordable for most of the laboratories working on drug discovery. As discussed in the previous section, active molecules will bind certain conformations more frequently than others. It is imperative to develop a reliable method to find those "magic snapshots" that will be selected by most of the compounds. So far, there are no reports of a trustworthy conformational coordinate that will allow us to cluster the whole set of conformations sampled by the MD simulation, such that the docking calculations be reduced, saving time, money, and extending the search space of off-target proteins. This, however, will be an important future research direction, for rest of the proteins that have known active/decoys and experimental structure available and are part of the 44 protein panel relevant to ADRs, conformations selected by ligands will be identifiable by our statistical method.

References

Bender, A., Scheiber, J., Glick, M., Davies, J. W., Azzaoui, K., Hamon, J., . . . Jenkins, J. L. (2007). Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure. *ChemMedChem, 2*, 861-873. doi:10.1002/cmdc.200700026

Berendsen, H. J. C., van der Spoel, D., & van Drunen, R. (1995). GROMACS: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications, 91*, 43-56. doi:10.1016/0010-4655(95)00042-E

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., . . . Bourne, P. E. (2000). The protein data bank. *Nucleic Acids Research, 28*, 235-242. doi:10.1093/nar/28.1.235

Bowes, J., Brown, A. J., Hamon, J., Jarolimek, W., Sridhar, A., Waldron, G., & Whitebread, S. (2012). Reducing safety-related drug attrition: the use of in vitro pharmacological profiling. *Nature Reviews Drug Discovery, 11*, 909-922. doi:10.1038/nrd3845

Cerqueira, N. M. F. S. A., Gesto, D., Oliveira, E. F., Santos-Martins, D., Brás, N. F., Sousa, S. F., . . . Ramos, M. J. (2015). Receptor-based virtual screening protocol for drug discovery. *Archives of Biochemistry and Biophysics, 582*, 56-67. doi:10.1016/j.abb.2015.05.011

Daura, X., Gademann, K., Jaun, B., Seebach, D., Van Gunsteren, W. F., & Mark, A. E. (1999). Peptide Folding: When Simulation Meets Experiment. *Angew. Chem. Int. Ed, 38*, 38: 236-240. doi:Doi 10.1002/(Sici)1521-3773(19990115)38:1/2<236::Aid-Anie236>3.0.Co;2-M

Deupi, X., & Kobilka, B. K. (2010). Energy landscapes as a tool to integrate GPCR structure, dynamics, and function. *Physiology (Bethesda, Md.), 25*, 293-303. doi:10.1152/physiol.00002.2010

Ellingson, S. R., Miao, Y., Baudry, J., & Smith, J. C. (2015). Multi-conformer ensemble docking to difficult protein targets. *Journal of Physical Chemistry B, 119*, 1026-1034. doi:10.1021/jp506511p

Ellingson, S. R., Smith, J. C., & Baudry, J. (2013). VinaMPI: Facilitating multiple receptor high-throughput virtual docking on high-performance computers. *Journal of Computational Chemistry, 34*, 2212-2221. doi:10.1002/jcc.23367

Evangelista, W., Weir, R. L., Ellingson, S. R., Harris, J. B., Kapoor, K., Smith, J. C., & Baudry, J. (2016). Ensemble-based docking: From hit discovery to metabolism and toxicity predictions. *Bioorganic and Medicinal Chemistry, 24*, 4928-4935. doi:10.1016/j.bmc.2016.07.064

Fiser, A., Kinh Gian Do, R., & Sali. (2000). Modeling Loops in Protein Structures. *Protein Science, 9*, 1753-1773. doi:10.1002/9780470882207.ch13

Gatica, E. A., & Cavasotto, C. N. (2012). Ligand and decoy sets for docking to G protein-coupled receptors. *Journal of Chemical Information and Modeling, 52*, 1-6. doi:10.1021/ci200412p

Gether, U., & Kobilka, B. K. (1998). G protein-coupled receptors. *The Journal of Biological Chemistry, 273*, 17979-17982. doi:10.1074/jbc.273.29.17979

Ghanemi, A., He, L., & Yan, M. (2013). New factors influencing G protein coupled receptorsâ€™ system functions. *Alexandria Journal of Medicine, 49*, 1-5. doi:10.1016/j.ajme.2012.10.002

Huang, Z., & Wong, C. F. (2016). Inexpensive Method for Selecting Receptor Structures for Virtual Screening. *Journal of Chemical Information and Modeling, 56*, 21-34. doi:10.1021/acs.jcim.5b00299

Ingólfsson, H. I., Melo, M. N., Van Eerden, F. J., Arnarez, C., Lopez, C. A., Wassenaar, T. A., . . . Marrink, S. J. (2014). Lipid organization of the plasma membrane. *Journal of the American Chemical Society, 136*, 14554-14559. doi:10.1021/ja507832e

Insel, P. A., Tang, C. M., Hahntow, I., & Michel, M. C. (2007). Impact of GPCRs in clinical medicine: Monogenic diseases, genetic variants and drug targets.

Biochimica et Biophysica Acta - Biomembranes, 1768, 994-1005. doi:10.1016/j.bbamem.2006.09.029

Lebon, G., Warne, T., & Tate, C. G. (2012). Agonist-bound structures of G protein-coupled receptors. *Current Opinion in Structural Biology, 22*, 482-490. doi:10.1016/j.sbi.2012.03.007

Leventis, P. A., & Grinstein, S. (2010). The Distribution and Function of Phosphatidylserine in Cellular Membranes. *Annual Review of Biophysics, Vol 39, 39*, 407-427. doi:DOI 10.1146/annurev.biophys.093008.131234

Lounkine, E., Keiser, M. J., Whitebread, S., Mikhailov, D., Hamon, J., Jenkins, J. L., . . . Urban, L. (2012). Large-scale prediction and testing of drug activity on side-effect targets. *Nature, 486*, 361-367. doi:10.1038/nature11159

Meng, X.-Y., Zhang, H.-X., Mezei, M., & Cui, M. (2011). Molecular Docking: A powerful approach for structure-based drug discovery. *Current Computational Aided Drug Design, 7*, 146-157. doi:10.1038/nature13314.A

Meyer, U. a. (2000). Pharmacogenetics and adverse drug reactions. *Lancet, 356*, 1667-1671. doi:10.1016/S0140-6736(00)03167-6

Monticelli, L., Kandasamy, S. K., Periole, X., Larson, R. G., Tieleman, D. P., & Marrink, S. J. (2008). The MARTINI coarse grained force field: extension to proteins. *J. Chem. Theory Comput., 4*, 819-834. doi:10.1021/ct700324x

Mysinger, M. M., Carchia, M., Irwin, J. J., & Shoichet, B. K. (2012). Directory of useful decoys, enhanced (DUD-E): Better ligands and decoys for better benchmarking. *Journal of Medicinal Chemistry, 55*, 6582-6594. doi:10.1021/jm300687e

Palczewski, K., & Orban, T. (2013). From Atomic Structures to Neuronal Functions of G Protein–Coupled Receptors. *Annual Review of Neuroscience, 36*, 139-164. doi:10.1146/annurev-neuro-062012-170313

Paramo, T., East, A., Garzón, D., Ulmschneider, M. B., & Bond, P. J. (2014). Efficient characterization of protein cavities within molecular simulation trajectories: Trj-cavity. *Journal of Chemical Theory and Computation, 10*, 2151-2164. doi:10.1021/ct401098b

Park, P. S.-H., Lodowski, D. T., & Palczewski, K. (2008). Activation of G Protein–Coupled Receptors: Beyond Two-State Models and Tertiary Conformational Changes. *Annual Review of Pharmacology and Toxicology, 48*, 107-141. doi:10.1146/annurev.pharmtox.48.113006.094630

Pierole, X., & Marrink, S.-J. (2013). Biomolecular Simulations. *Biomolecular Simulations: Methods and Protocols, 924*, 617. doi:10.1007/978-1-62703-017-5

Pirmohamed, M., Breckenridge, A. M., Kitteringham, N. R., & Park, B. K. (1998). Adverse drug reactions. *BMJ: British Medical Journal (International Edition), 316*, 1295-1298.

Salgado, C. M., Azevedo, C., Proença, H., & Vieira, S. M. (2016). Noise Versus Outliers. In *Secondary Analysis of Electronic Health Records* (pp. 163-183). Cham: Springer International Publishing.

Sanner, M. F. (1999). Python: a programming language for software integration and development. *Journal of molecular graphics & modelling, 17*, 57-61. doi:10.1016/S1093-3263(99)99999-0

Shahane, G., Parsania, C., Sengupta, D., & Joshi, M. (2014). Molecular insights into the dynamics of pharmacogenetically important N-terminal variants of the human β2-adrenergic receptor. *PLoS computational biology, 10*, e1004006. doi:10.1371/journal.pcbi.1004006

Stansfeld, P. J., Goose, J. E., Caffrey, M., Carpenter, E. P., Parker, J. L., Newstead, S., & Sansom, M. S. P. (2015). MemProtMD: Automated Insertion of Membrane Protein Structures into Explicit Lipid Membranes. *Structure, 23*, 1350-1361. doi:10.1016/j.str.2015.05.006

Taboureau, O., Baell, J. B., Fernández-Recio, J., & Villoutreix, B. O. (2012). Established and emerging trends in computational drug discovery in the structural genomics era. *Chemistry and Biology, 19*, 29-41. doi:10.1016/j.chembiol.2011.12.007

Tukey, J. W. J. W. (1977). Exploratory data analysis.

Wassenaar, T. A., Ingólfsson, H. I., Böckmann, R. A., Tieleman, D. P., & Marrink, S. J. (2015). Computational lipidomics with insane: A versatile tool for generating custom membranes for molecular simulations. *Journal of Chemical Theory and Computation, 11*, 2144-2155. doi:10.1021/acs.jctc.5b00209

Wassenaar, T. A., Pluhackova, K., Böckmann, R. A., Marrink, S. J., & Tieleman, D. P. (2014). Going backward: A flexible geometric approach to reverse transformation from coarse grained to atomistic models. *Journal of Chemical Theory and Computation, 10*, 676-690. doi:10.1021/ct400617g

Appendix

Table II-5. Thresholds to determine outliers for ADORA2A and ADRB2 in random selection distributions.

ADORA2A					
Actives:	844	**Decoys:**	10899	**Total:**	11743
#Percentage%	0.5 %	1.0 %	5.0 %	10.0 %	
Expected sample size	59	117	587	1174	
Expected Actives.	4	8	42	84	
#Upper Value ITQ	9	17	59	107	

ADRB2					
Actives:	447	**Decoys:**	15255	**Total:**	15702
#Percentage%	0.5 %	1.0 %	5.0 %	10.0 %	
Expected sample size	79	157	785	1570	
Expected Actives.	2	4	22	45	
#Upper Value ITQ	7	11	35	63	

Table II-6. Thresholds to determine outliers for HRH1 and CHRM2 in random selection distributions.

HRH1					
Actives:	86	**Decoys:**	3354	**Total:**	3440
#Percentage%	0.5 %	1.0 %	5.0 %	10.0 %	
Expected sample size	17	34	172	344	
Expected Actives.	0	1	4	9	
#Upper Value ITQ	3	3	11	15	

CHRM2					
Actives:	126	**Decoys:**	4914	**Total:**	5040
#Percentage%	0.5 %	1.0 %	5.0 %	10.0 %	
Expected sample size	25	50	252	504	
Expected Actives.	1	1	6	13	
#Upper Value ITQ	3	6	13	23	

Table II-7. Thresholds to determine outliers for HTR1B and HTR2B in random selection distributions.

HTR1B					
Actives:	113	**Decoys:**	4407	**Total:**	4520
#Percentage%	0.5 %	1.0 %	5.0 %	10.0 %	
Expected sample size	23	45	226	452	
Expected Actives.	1	1	6	11	
#Upper Value ITQ	3	6	12	20	

HTR2B					
Actives:	227	**Decoys:**	8853	**Total:**	9080
#Percentage%	0.5 %	1.0 %	5.0 %	10.0 %	
Expected sample size	45	91	454	908	
Expected Actives.	1	2	11	23	
#Upper Value ITQ	6	7	20	36	

Table II-8. Thresholds to determine outliers for OPRD1 and OPRK1 in random selection distributions.

OPRD1					
Actives:	377	**Decoys:**	14703	**Total:**	15080
#Percentage%	0.5 %	1.0 %	5.0 %	10.0 %	
Expected sample size	75	151	754	1508	
Expected Actives.	2	4	19	38	
#Upper Value ITQ	7	10	32	55	

OPRK1					
Actives:	307	**Decoys:**	11973	**Total:**	12280
#Percentage%	0.5 %	1.0 %	5.0 %	10.0 %	
Expected sample size	61	123	614	1228	
Expected Actives.	2	3	15	31	
#Upper Value ITQ	4	8	26	45	

CHAPTER III

OPPORTUNITY AND CHALLENGES OF ENSEMBLE-BASED DOCKING IN TOXICITY

Abstract

The increase of computational power and production of data lead the scientific community to exponential progress in many fields of science. One of the most impressive advances comes from the increase of computational power through supercomputing. This has improved and expanded molecular simulations in structural biology. Nowadays it is possible to run long single molecular dynamics simulations, one microsecond, of about 100 thousands atoms in only 24 hours on a supercomputer. Similar improvements have been achieved in molecular docking; massive dockings can screen now millions of compounds in the same period of time. This represents an excellent opportunity to study and predict protein-drug interactions on a scale not reachable previously. However, this progress also brings a challenge; how to manage the immense amount of data generated by these simulations, and how to correlate this data with the experimental data. So far, techniques from data mining and machine learning have been useful in the integration and information crossing of all these sources. These techniques and the treatment of information will continue to be important and crucial for next years in order to get solutions and insights of many biological problems.

Introduction

Adverse Drug Reactions (ADR) can have different mechanisms: polymorphism in genes that code enzymes involved in drug metabolism (Meyer, 2000), immune and non-immune mechanisms producing hypersensitivity responses (Dao, Su, & Chung, 2015; Riedl & Casillas, 2003), or other mechanisms reviewed elsewhere (Edwards & Aronson, 2000). However, in fine, the molecular mechanisms mediating ADR always involves protein-drug interactions. This represents the most difficult challenge in the drug design/discovery field. This is because any new drug candidate has a potentially high chance to bind any off-target protein. How to predict whether a drug candidate will bind an off-target protein? An important advance was provided in 2012, when *in vitro* drug screening led to proposing a panel of 44 proteins as a set to test ADR (Bowes et al., 2012). Even though this is great progress, it is still too expensive and time consuming to test the, typically thousands of, drug candidates against this panel of proteins. The alternative is, of course, *in silico* screening. However, this task faces a challenging biological fact, proteins do not exist in only one conformation, and actually, such a conformation might differ from the experimental structure obtained by NMR or X-Ray diffraction (Hilser, Garcia-Moreno E., Oas, Kapp, & Whitten, 2006). Thus, in order to examine whether any molecule would bind a particular protein, it is necessary to test as many conformational states as possible, providing better insights on the protein-drug/candidate interaction, as described in the previous chapter. Efforts to predict general protein-drug binding has been extensive, numerous software applications have been developed to this end, for example, Autodock, Vina, VinaMPI, Gold, Dock, and some others (Pagadala, Syed, & Tuszynski, 2017). The current challenge is to use any of these applications to perform massive high-throughput docking and build a reliable model to decide whether or not a molecule, potential drug candidate, will bind the protein. This is difficult, since at the molecular level there are many factors involved: protonation states of both ligand and receptor, contributions of solvent, conformational entropy, ligand's flexibility, building of energy function, score function, etc. Additionally, one must consider how to rank the massive output obtained from these *in silico* experiments, and make sure that the data provided is reliable beyond the random distribution. The next sections will describe what I think the future directions on ensemble-based docking as virtual screening are, and their role in toxicity predictions, and addressing a particular challenge: how to share and learn from our and others experiences taking into account the increasing amount of information generated every day all over the world. The tools to handle this massive amount of information are data mining and learning machine technologies, whose principles and applications are also described.

.

Current State of Ensemble-base Docking and ADR

In silico experiments of protein-drug interactions have found their most common applications in docking simulations in which a protein is used in a structure-based approach to computationally identify the molecules from a collection predicted to have the most favorable binding energies, and hence to bind more strongly to the target. This is an approach used in the initial steps of drug discovery, where new hits/leads are needed. *In silico* screening thus helps to prioritize chemicals for experimental assays (Jorgensen, 2009). Our laboratory has also used this approach to identify compounds capable of modulating the interactions between proteins responsible for coagulation cascade, Factor Xa (FXa) and Factor Va (FVa). Drugs that inhibit enzymatic functions of FXa already exist, nonetheless, their safety profile is extremely narrow, and have shown to be difficult to use even in hospitals. This project has led to the discovery of novel molecules that can bind to FXa's surface and modulate the interactions with FVa, without affecting FXa's enzymatic functions (Kapoor et al., 2016).

However, narrowing search space or validating potential drug candidates have only been the initial applications of docking. For a long time, the target-centric approaches ignored the physiological context and the cellular composition, which made docking calculations less useful, particularly when the predicted results were could not be validated by biochemical assays or pre-clinical and clinical trials (Iskar, Zeller, Zhao, van Noort, & Bork, 2012). Ensemble-based docking can now be part of a more integral pharmacological strategy, including high-throughput virtual screening of not just the many conformations of the target protein, but also of all the possible proteins responsible for a toxic response. For instance, during the latest Ebola outbreak, an integrated strategy was employed to seek a potential drug. Among 1,766 drugs approved by the FDA, 259 experimental drugs were screened looking for a potential compound that could inhibit Ebola virulence or replication. A protocol including three computational approaches was used: proteome-wide ligand binding site comparison

(Xie & Bourne, 2007), molecular docking, and MD simulations, resulting in identification of Indivinavir, a known HIV protease inhibitor, as a likely reducer of Ebola virulence, and some other anti-fungal and anti-viral drugs as potential RNA polymerase inhibitors (Z. Zhao, Martin, Fan, Bourne, & Xie, 2016). The same year another study predicted emodin-8-beta-D-glucoside from the Traditional Chinese Medicine Database as a potential inhibitor of viral protein 40 (VP40) from Ebola after an integrated work carried out using molecular docking, public databases of toxicity like Protox (Drwal, Banerjee, Dunkel, Wettig, & Preissner, 2014), and MD simulations to validate docking findings (Karthick et al., 2016). These experiences show that ensemble-based docking has already been used as part of an integrated set of computational tools to identify/propose compounds to treat a particular disease, either in repurposing FDA-approved drugs, or filtering experimental drugs that could result in adverse reactions.

A different strategy to predict/test ADR is proposed by LaBute et.al (Labute et al., 2014), in his work in which combined information from DrugBank (Law et al., 2014; Wishart, 2006), Side Effect Resource (SIDER) (Kuhn, Campillos, Letunic, Jensen, & Bork, 2010) and molecular docking were used to train a machine learning model with 906 small molecules and 409 protein targets, resulting in predictions that are comparable in quality to those obtained using the same model over publicly available and experimentally-derived drug-protein interaction data. The combination of molecular docking software and machine learning systems has been also employed to integrate structure-based drug design and quantitative structure-activity relationships (QSAR) into a learning model to improve the performance of binding site recognition, using to this end 139 different kinases and 33 inhibitors. The inclusion of machine learning enhanced the binding prediction of the molecular docking software and therefore the identification of potential targets (Hsin, Ghosh, & Kitano, 2013). What these examples show is that machine learning methods are already part of different studies on the way to enhance not just molecular docking scores, but also to identify off-target proteins responsible for ADR. Currently, virtual and experimental high-throughput screenings of chemical libraries have become the major tools not only to identify on-target hit compounds, but also off-target hits. Additionally, relatively new techniques have been incorporated to the pool of methods in drug discovery: next-generation sequencing to identify new targets, biomarkers, polypharmacology to identify networks by modulating multiple targets. Thus, effective drugs can be developed potentiating on-target hits and avoiding off-target binding (Anighoro, Bajorath, & Rastelli, 2014; Taboureau, Baell, Fernández-Recio, & Villoutreix, 2012).

Integration of Molecular Tools and Databases

The inclusion of MD simulations in a docking strategy could be an important contribution for drug discovery if it were less computationally expensive (H. Zhao & Caflisch, 2015). However, GPU technology makes MD simulations more affordable in terms of computing time (Kutzner et al., 2015; Salomon-Ferrer, Götz, Poole, Le Grand, & Walker, 2013). Thus, GPU-based MD software like Gromacs and Amber are already being used in research on different fields, such as solid-liquid phase transition (Nomura, Oikawa, Kawai, Narumi, & Yasuoka, 2014), MD simulations of the DNA duplex (Galindo-Murillo, Roe, & Cheatham, 2015), and protein folding (Bermudez, Mortier, Rakers, Sydow, & Wolber, 2016). GPU-based MD simulations will be extremely useful in validating findings from docking calculations in a faster way. Similarly, in the same way VinaMPI and VinaLC have boosted virtual screening in drug discovery through parallelization and scalability in clusters and super computers, it is expected that molecular docking software based on GPU technology will increase the performance of ensemble-base docking. In a similar field, there is already GPU-based docking software for protein-protein interactions, for instance, Megadock-GPU (Shimoda, Ishida, Suzuki, Ohue, & Akiyama, 2013), PIPER14 (Landaverde & Herbordt, 2014). Nonetheless, there is not yet a GPU-based docking software for virtual screening of ligand libraries. Since ensemble-based docking is now used not only to identify/validate on-target protein-ligand interactions, but also to identify possible off-target protein binding, developing this kind of software should be one of the top priorities in the field. The availability of GPU technology will make high-throughput molecular docking more affordable, in contrast to the current situation where performing a virtual screening of thousands of compounds over thousands of conformational states, generating millions of receptor-ligand complexes, requires computational power only found in supercomputers.

Developing a clustering method to find significant frames in the conformational space is still a pending task; an efficient conformation coordinate will characterize changes in the structure of the protein during the sampling process. This, of course, will reduce the search space for molecular docking calculations. Root Mean Square Deviation (RMSD) is the most common parameter to measure variations in the protein structure. Depending on the region or domain in the structure, the RMSD can be a measurement for any particular set of atoms, whether it's the binding site, secondary structure, or the entire protein. Typically, C-α atoms are used as a metric to measure changes in the structure, however, this metric can miss some changes. For instance, two conformations of the same protein can exhibit a RMSD of 2.3 Å with respect to a reference structure. However, the first conformation could have those 2.3 Å of variation in a loop, while the second conformation gets the same RMSD due to an alpha helix domain 20 Å apart from the loop. Even though the RMSD values are identical, the conformational changes will originate from totally different conformations. Markov state models have been used to characterize protein conformations (Chodera & Noé, 2014; Lane, Bowman, Beauchamp, Voelz, & Pande, 2011), and principal component analysis (Balsera, Wriggers, Oono, & Schulten, 1996; Papaleo, Mereghetti, Fantucci, Grandori, & De Gioia, 2009; Sittel, Jain, & Stock, 2014), or some particular geometric parameter can also be used. Therefore, an appropriate conformational coordinate is necessary to reduce the number of conformational states subject of molecular docking, otherwise, this problem will stay intractable.

Progress in any field of science has been achieved due to the sharing of information. Biological sciences, in particular, have experienced rapid progress in different disciplines: structural biology, genomic, proteomic, genetics, drug discovery, simulation software, etc. Most of the databases and tools used to share and search data of interest are integrated by the National Center for Biotechnology Information (NCBI). On the other hand, some public databases, which are very important for research, are not part of NCBI: Protein Data Bank, The Cambridge Structural Database, Database of Useful Decoys-Enhanced (DUD-E), etc. Since this work is particularly interested in adverse drug reactions, I found particular interest and promise in some of these databases:

DrugBank: Database of drug and drug targets with chemical and pharmacological information (Wishart, 2006).
https://www.drugbank.ca/

DUD-E: directory of active ligands and decoy molecules for each ligand and their respective targets (Mysinger, Carchia, Irwin, & Shoichet, 2012).
http://dude.docking.org/

SIDER: Information on marketed medicines and their adverse drug reactions; including side effect frequency, side effect classification, and in some cases drug-target relations (Kuhn, Letunic, Jensen, & Bork, 2016).
http://sideeffects.embl.de/
Two interesting initiatives, but apparently not well maintained sites are:

PDBbind: Experimentally measured binding affinity data for protein-ligand complexes from the Protein Data Bank (Wang, Fang, Lu, Yang, & Wang, 2005).
http://sw16.im.med.umich.edu/databases/pdbbind/index.jsp

Platinum: Structural database of experimentally measured effects of mutations on

protein-ligand complexes (Pires, Blundell, & Ascher, 2015).
http://biosig.unimelb.edu.au/platinum/

These databases contain important information for drug discovery research. However, besides the common format for structures of molecules, there is not a standard format to share other information related to protein-ligand complexes, such as affinity constants, dissociation constants or clinical and toxicity information. To integrate all these web servers is, of course, utopic. Nonetheless, it would be extremely helpful for the scientific community to define standard formats to share this kind of data. Moreover, since these databases only gather information, it is crucial to fill them with experimental and computational data, which assists investigators in the different stages of the drug discovery/design process, and even later, in the pre-clinical and trial phases (Taboureau et al., 2012). Of particular importance is the expansion of the experimental data beyond the panel proposed by Bowes (Bowes et al., 2012). This will likely help in the identification of other proteins responsible for adverse drug reactions.

For the scientific community to take advantage of the integration of all these different sources of information, it is necessary to use two essential technologies: data mining and machine learning. Data mining is the automatic harvesting of information from large databases in order to find unknown patterns or knowledge (Lavecchia, 2015). This usually leads to models to explain a particular phenomenon. However, when data increases in amount and type, the model needs adjustments and/or reformulations. Dealing with this kind of problem manually is time and resource consuming, Machine Learning is a technology that is capable of addressing these sorts of problems, this is a field of study that gives computer the ability to learn without being explicitly programmed (Samuel, 1959). As more data become available, machine learning can be used to process new types or data, making the machine learning more favorable than manual programming in terms of time, and consequently cost (Domingos, 2012). This technique has already been employed to deal with increasing amounts of complex data in applications such as web searches, spam filters, stock trading, drug design, as well as in proteomics and genomics (Lavecchia, 2015; Li, Wu, & Ngom, 2016).

References

Anighoro, A., Bajorath, J. r., & Rastelli, G. (2014). Polypharmacology: Challenges and opportunities in drug discovery. *Journal of Medicinal Chemistry, 57*, 7874-7887. doi:10.1021/jm5006463

Balsera, M. a., Wriggers, W., Oono, Y., & Schulten, K. (1996). Principal Component Analysis and Long Time Protein Dynamics. *The Journal of Physical Chemistry, 100*, 2567-2572. doi:10.1021/jp9536920

Bermudez, M., Mortier, J., Rakers, C., Sydow, D., & Wolber, G. (2016). More than a look into a crystal ball: protein structure elucidation guided by molecular dynamics simulations. *Drug Discovery Today, 21*, 1799-1805. doi:10.1016/j.drudis.2016.07.001

Bowes, J., Brown, A. J., Hamon, J., Jarolimek, W., Sridhar, A., Waldron, G., & Whitebread, S. (2012). Reducing safety-related drug attrition: the use of in vitro pharmacological profiling. *Nature Reviews Drug Discovery, 11*, 909-922. doi:10.1038/nrd3845

Chodera, J. D., & Noé, F. (2014). Markov state models of biomolecular conformational dynamics. *Current Opinion in Structural Biology, 25*, 135-144. doi:10.1016/j.sbi.2014.04.002

Dao, R.-L., Su, S.-C., & Chung, W.-H. (2015). Recent advances of pharmacogenomics in severe cutaneous adverse reactions: immune and nonimmune mechanisms. *Asia Pac Allergy, 5*, 59-67. doi:10.5415/apallergy.2015.5.2.59

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM, 55*, 78. doi:10.1145/2347736.2347755

Drwal, M. N., Banerjee, P., Dunkel, M., Wettig, M. R., & Preissner, R. (2014). ProTox: A web server for the in silico prediction of rodent oral toxicity. *Nucleic Acids Research, 42*, 53-58. doi:10.1093/nar/gku401

Edwards, I. R., & Aronson, J. K. (2000). Adverse drug reactions: definitions, diagnosis, and management. *Lancet (London, England), 356*, 1255-1259. doi:10.1016/S0140-6736(00)02799-9

Galindo-Murillo, R., Roe, D. R., & Cheatham, T. E. (2015). Convergence and reproducibility in molecular dynamics simulations of the DNA duplex d(GCACGAACGAACGAACGC). *Biochimica et Biophysica Acta - General Subjects, 1850*, 1041-1058. doi:10.1016/j.bbagen.2014.09.007

Hilser, V. J., Garcia-Moreno E., B., Oas, T. G., Kapp, G. K., & Whitten, S. T. (2006). A Statistical Thermodynamic Model of the Protein Ensemble. *Chem. Rev., 106*, 1545-1558.

Hsin, K. Y., Ghosh, S., & Kitano, H. (2013). Combining machine learning systems and multiple docking simulation packages to improve docking prediction reliability for network pharmacology. *PLoS ONE, 8*. doi:10.1371/journal.pone.0083922

Iskar, M., Zeller, G., Zhao, X. M., van Noort, V., & Bork, P. (2012). Drug discovery in the age of systems biology: The rise of computational approaches for data integration. *Current Opinion in Biotechnology, 23*, 609-616. doi:10.1016/j.copbio.2011.11.010

Jorgensen, W. L. (2009). Efficient drug lead discovery and optimization. *Accounts of Chemical Research, 42*, 724-733. doi:10.1021/ar800236t

Kapoor, K., McGill, N., Peterson, C. B., Meyers, H. V., Blackburn, M. N., & Baudry, J. (2016). Discovery of Novel Nonactive Site Inhibitors of the Prothrombinase Enzyme Complex. *Journal of Chemical Information and Modeling, 56*, 535-547. doi:10.1021/acs.jcim.5b00596

Karthick, V., Nagasundaram, N., Doss, C. G. P., Chakraborty, C., Siva, R., Lu, A., . . . Zhu, H. (2016). Virtual screening of the inhibitors targeting at the viral protein 40 of Ebola virus. *Infectious diseases of poverty, 5*, 12. doi:10.1186/s40249-016-0105-1

Kuhn, M., Campillos, M., Letunic, I., Jensen, L. J., & Bork, P. (2010). A side effect resource to capture phenotypic effects of drugs. *Molecular Systems Biology, 6*, 1-6. doi:10.1038/msb.2009.98

Kuhn, M., Letunic, I., Jensen, L. J., & Bork, P. (2016). The SIDER database of drugs and side effects. *Nucleic Acids Research, 44*, D1075-D1079. doi:10.1093/nar/gkv1075

Kutzner, C., P??ll, S. r., Fechner, M., Esztermann, A., De Groot, B. L., & Grubm??ller, H. (2015). Best bang for your buck: GPU nodes for GROMACS biomolecular simulations. *Journal of Computational Chemistry, 36*, 1990-2008. doi:10.1002/jcc.24030

Labute, M. X., Zhang, X., Lenderman, J., Bennion, B. J., Wong, S. E., & Lightstone, F. C. (2014). Adverse drug reaction prediction using scores produced by large-scale drug-protein target docking on high-performance computing machines. *PLoS ONE, 9*. doi:10.1371/journal.pone.0106298

Landaverde, R., & Herbordt, M. C. (2014). GPU Optimizations for a Production Molecular Docking Code*.

Lane, T. J., Bowman, G. R., Beauchamp, K., Voelz, V. A., & Pande, V. S. (2011). Markov State model reveals folding and functional dynamics in ultra-long MD trajectories. *Journal of the American Chemical Society, 133*, 18413-18419. doi:10.1021/ja207470h

Lavecchia, A. (2015). Machine-learning approaches in drug discovery: Methods and applications. *Drug Discovery Today, 20*, 318-331. doi:10.1016/j.drudis.2014.10.012

Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A. C., Liu, Y., . . . Wishart, D. S. (2014). DrugBank 4.0: Shedding new light on drug metabolism. *Nucleic Acids Research, 42*, 1091-1097. doi:10.1093/nar/gkt1068

Li, Y., Wu, F.-X., & Ngom, A. (2016). A review on machine learning principles for multi-view biological data integration. *Briefings in Bioinformatics*, bbw113. doi:10.1093/bib/bbw113

Meyer, U. a. (2000). Pharmacogenetics and adverse drug reactions. *Lancet, 356*, 1667-1671. doi:10.1016/S0140-6736(00)03167-6

Mysinger, M. M., Carchia, M., Irwin, J. J., & Shoichet, B. K. (2012). Directory of useful decoys, enhanced (DUD-E): Better ligands and decoys for better benchmarking. *Journal of Medicinal Chemistry, 55*, 6582-6594. doi:10.1021/jm300687e

Nomura, K., Oikawa, M., Kawai, A., Narumi, T., & Yasuoka, K. (2014). GPU-accelerated replica exchange molecular simulation on solid–liquid phase transition study of Lennard-Jones fluids. *Molecular Simulation, 7022*, 1-7. doi:10.1080/08927022.2014.954572

Pagadala, N. S., Syed, K., & Tuszynski, J. (2017). Software for molecular docking: a review. *Biophysical Reviews, 9*, 91-102. doi:10.1007/s12551-016-0247-1

Papaleo, E., Mereghetti, P., Fantucci, P., Grandori, R., & De Gioia, L. (2009). Free-energy landscape, principal component analysis, and structural clustering to identify representative conformations from molecular dynamics simulations: The myoglobin case. *Journal of Molecular Graphics and Modelling, 27*, 889-899. doi:10.1016/j.jmgm.2009.01.006

Pires, D. E. V., Blundell, T. L., & Ascher, D. B. (2015). Platinum: a database of experimentally measured effects of mutations on structurally defined protein–ligand complexes. *Nucleic Acids Research, 43*(D1), D387-D391. doi:10.1093/nar/gku966

Riedl, M. A., & Casillas, A. M. (2003). Adverse drug reactions: types and treatment options. *American Family Physicians, 68*, 1781-1790.

Salomon-Ferrer, R., Götz, A. W., Poole, D., Le Grand, S., & Walker, R. C. (2013). Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh ewald. *Journal of Chemical Theory and Computation, 9*, 3878-3888. doi:10.1021/ct400314y

Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development, 3*, 210-229. doi:10.1147/rd.33.0210

Shimoda, T., Ishida, T., Suzuki, S., Ohue, M., & Akiyama, Y. (2013). MEGADOCK-GPU: Acceleration of Protein-Protein Docking Calculation on GPUs. *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics - BCB'13*, 883-889. doi:10.1145/2506583.2506693

Sittel, F., Jain, A., & Stock, G. (2014). Principal component analysis of molecular dynamics: On the use of Cartesian vs. internal coordinates. *Journal of Chemical Physics, 141*. doi:10.1063/1.4885338

Taboureau, O., Baell, J. B., Fernández-Recio, J., & Villoutreix, B. O. (2012). Established and emerging trends in computational drug discovery in the structural genomics era. *Chemistry and Biology, 19*, 29-41. doi:10.1016/j.chembiol.2011.12.007

Wang, R., Fang, X., Lu, Y., Yang, C. Y., & Wang, S. (2005). The PDBbind database: Methodologies and updates. *Journal of Medicinal Chemistry, 48*, 4111-4119. doi:10.1021/jm048957q

Wishart, D. S. (2006). DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research, 34*, D668-D672. doi:10.1093/nar/gkj067

Xie, L., & Bourne, P. E. (2007). A robust and efficient algorithm for the shape description of protein structures and its application in predicting ligand binding sites. *BMC Bioinformatics, 8*, S9. doi:10.1186/1471-2105-8-S4-S9

Zhao, H., & Caflisch, A. (2015). Molecular dynamics in drug design. *European Journal of Medicinal Chemistry, 91*, 4-14. doi:10.1002/ijch.201400009

Zhao, Z., Martin, C., Fan, R., Bourne, P. E., & Xie, L. (2016). Drug repurposing to target Ebola virus replication and virulence using structural systems pharmacology. *BMC Bioinformatics, 17*, 90. doi:10.1186/s12859-016-0941-9

CONCLUSIONS

Computational approaches have become an important tool to understand molecular mechanism of many properties in proteins. Particularly remarkable has been the role of MD simulations of full-atom models or coarse grained models. This technique combined with the use of supercomputers has increased our knowledge of protein dynamics and how different conformations of proteins coexist instead of only one single structure. This ensemble can also vary as interactions with other molecules occur. Chapter I presented a case, Aminoglycoside nucleotidyltransferase 4', where changes in the environment modifies the structures distribution, e.g. as temperature is increased the distribution of conformational states changed. However, these changes are not trivial to identify. A reduction of dimensionality, via principal component analysis, was performed to characterize these modifications in the ensemble. Two main principal components were identified and used to characterize the protein structural ensemble. Potential of mean force, using the first two principal components as conformation coordinates, was then calculated to visualize a free energy landscape of the ensemble at three different temperatures, these energy landscapes show how the ensemble change as the temperature increases, or as point mutations are introduced in the system at a given temperature. The results also suggest that mutations bring on global effects in protein flexibility affecting the distribution of conformations in the ensemble.

Chapter II describes the work done to build a reliable method to analyze results coming from ensemble-based docking calculations. The use of 'outliers', as defined by the Exploratory Data Analysis field, has been useful to discriminate conformations that are selected by more ligands than by a random selection. For this purpose four G-Protein Coupled Receptors were used as model systems, these four proteins have been reported as part of a set of 44 proteins responsible for about 75% of adverse drug reactions. Ensembles of representative structures were examined as well as ensembles containing 3,000 conformations provided by coarse grained molecular dynamics simulations. The results indicate that the method developed in this work is more efficient than a random selection in all the cases. Application of this method on docking results from representative structures produced a remarkable improvement respect to random selection or from using single crystal structure for two proteins. When this technique was performed on 3,000 conformations for each protein, results showed a dramatically improvement with respect to random selection, of using a single crystal structure, and docking on representative structures, correctly identifying up to 73%, 80%, 87%, and 99% of the chemicals that are known to bind these proteins off-targets, leading to adverse reactions.

A review of the current state and future directions in the field of molecular docking and other resources as tools to predict adverse drug reactions is presented in Chapter III. The integration of data and algorithms has been a strategy that has allowed progress in numerous fields of biology. In particularly, public databases with information about drug toxicity are now available and they can be used together with structural biology data to speed up the prediction of such effects in the drug discovery process. Ensemble-based docking in this context represents not only an option to speed up the drug design/discovery process but a necessity. However, its high computational cost still makes this technique affordable only for laboratories with access to supercomputing resources and expertise. Additionally, the generation of such massive amount of data implies also innovation in other areas of science, such as data mining and machine learning. Proficiency in managing this amount data will determine the future of *in silico* prediction of adverse drug reactions.

VITA

Wilfredo Evangelista Falcón was born in Peru. He completed the B.S. degree in Physics at Universidad Nacional de Ingeniería, Peru. He joined the graduate school at Universidad Peruana Cayetano Heredia, Peru. He worked as visiting scholar at the University of Texas Medical Branch, Texas, between 2009 and 2010. In 2012, he joined the doctoral program in the department of Biochemistry, Cellular, and Molecular Biology in the University of Tennessee at Knoxville, where he met his advisor Dr. Jerome Baudry. In 2013 he joined the Center for Molecular Biophysics at Oak Ridge National Laboratory where he worked as graduate research assistant. He is graduating in December 2017.