Accounting for side-chain flexibility in protein-ligand docking:

3D Interaction Homology as an approach of quantifying side-chain flexibility of Tyrosine Rotamers

Proteins are the most common and fundamental molecules found in cells and come with a wide variety of functions determined by their structure. A major factor in a proteins ability to fulfill its function is its ability to bind to other molecules like drugs. A drug is a small molecule that can interact, bind and control the function of biological receptors and aid in fighting or curing a disease. Since drugs are molecules that interact with receptors of proteins, they are considered ligands. The interactions between ligands and proteins, and signal transmission trough molecular complementarity are crucial for all processes happening in living organisms.

Protein-ligand docking is a [molecular modelling](https://en.wikipedia.org/wiki/Molecular_modelling) technique with a goal to predict the position and orientation of a [ligand](https://en.wikipedia.org/wiki/Ligand_(biochemistry)) when it is bound to a [protein](https://en.wikipedia.org/wiki/Protein) receptor or enzyme.1

Three different models, the “lock-and-key” “induced fit” and “conformational selection” have been proposed to explain the protein–ligand binding mechanisms. Both the lock-and-key and the induced fit models treat the protein as a single and stable conformation under given experimental conditions. This is not the case for all proteins. In reality, most proteins are inherently dynamic. The conformational selection model (Fig. 1) takes into account this inherent flexibility considering that the native state of a protein does not exist as a single, rigid conformation but rather as a vast ensemble of conformational states that coexist in equilibrium with different substrates, and that the ligand can bind selectively to the most suitable conformational state, ultimately shifting the equilibrium towards this state/substate.2



Figure 1. Illustration of the conformational selection protein(gray)-ligand(black)

binding mechanism

Several studies have shown that better sampling of motions during docking improves the ability of protein-ligand complementarity scoring functions to detect the most accurate docking. For example: when using the docking and screening tool SLIDE (Screening for Ligands with Induced-fit Docking, Efficiently) to dock 42 known thrombin ligands and 15 glutathione S-transferase ligands into the apo-protein structures (reflecting the ligand-free binding site conformations), only 9 of the 42 thrombin ligands and 9 of the 15 GST ligands could be docked without modeling protein conformational change, even when the ligands were provided in their

protein-bound conformations3

Among the range of structural changes in the receptor protein in protein-ligand binding one of the crucial changes are the small side-chain rearrangements in the binding pocket residues, which maximize the interaction between the proteins receptor and the ligand. A study done by Najmanovich et al. showed that 90% of the proteins binding sites display a conformational change upon ligand binding.4 What accounts for the side-chain rearrangements to happen is the side chains flexibility level. Assessing amino acid side-chain flexibility has been a challenge because of the vast degrees of freedom that have to be consideredand the analysis of flexibility can give insights valuable to improve protein-ligand docking algorithms and can provide an index of amino‐acid side-chain flexibility potentially useful in molecular biology and protein engineering studies4. Here, the use of 3D Homology Mapping is proposed as an approach of understanding the environment surrounding amino acid rotamers as they are undergoing conformational changes and thereby quantifying the flexibility of their side-chains based on the interactions between a residue and its neighbor.

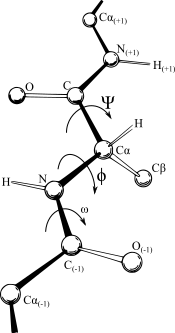
Homology Modeling

Homology modeling, also called comparative modeling or template-based modeling, is a method of creating 3D models of proteins using a known experimental structure of a homologous protein. The structure provided by homology modeling contains sufficient information about the spatial arrangement of important residues in the protein and may guide the design of new experiments.

Dr. Kellogg alongside with his team-members of The Kellogg Lab here at VCU Department of Medicinal Chemistry, developed their own approach to 3D Homology Mapping where the environments surrounding amino acid residues within proteins are described in a both qualitative and quantitative way. The maps are qualitative in that they indicate the character of interactions between the residue and its neighbor in terms or four classes: favorable polar, unfavorable polar, favorable hydrophobic and unfavorable hydrophobic. These four classes are named hydropathy.

The methods of this approach are found in Ahmed et al.5 which is the primary source that I am going to be using for this proposal. A computational tool called HINT (Hydropathic INTeractions), also developed by the Kellogg Lab, is used to score the hydropathy interactions from structural models. HINT is a scoring tool that exploits the free energy information from partition coefficients of solute transfer between water and 1-octanol as a force-field that recognizes hydropathic interactions, while inherently encoding entropy and solvation/desolvation.

Methods

HINT basis interaction mapping

First, a data set of 2703 randomly selected proteins from the RCSB Protein Data Bank6 was created. Each protein had no cofactor in their structure, ensuring that all possible 3D orientations and environments of Tyrosine are represented. 55 of these proteins did not contain tyrosine. Hydrogen atoms were added to the structures to eliminate steric clashes.

Then, an 8 by 8 matrix Ramachandran plot was established with each element being a 45 by 45 square in phi-psi space. Ramachandran Plot is a way to visualize dihedral angles φ (phi) against ψ (psi) of amino acid residues in protein structure. In proteins the two torsion angles φ and ψ describe the rotation of the polypeptide chain around the two bonds on both sides of the Cα (alpha-Carbon) atom. (fig. 2)

figure 2. Backbone dihedral

angles φ and ψ

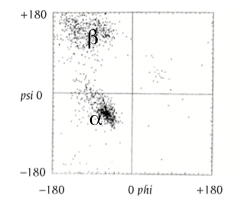
 Figure 3 shows an example of an Ramachandran plot. The horizontal axis shows φ values, while the vertical shows ψ values. Each dot on the plot shows the angles for an amino acid. The plot extends from -180 to +180 on both the vertical and horizontal axes. This is a convenient presentation and allows clear distinction of the characteristic regions of α-helices and β-sheets. The regions on the plot with the highest density of dots are the so-called “allowed” regions, also called low-energy regions. Some values of φ and ψ are not allowed since the involved atoms come too close together, resulting in a steric clash. For a high-quality and high resolution experimental structure these regions are usually empty or almost empty - very few amino acid residues in proteins have their torsion angles within these regions.7

Figure 3. Example of a Ramachandran plot

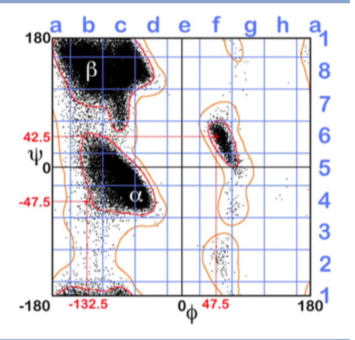
The phi-psi boundaries for the matrix used in this experiment were shifted by -20 and -25 respectively so the highest density regions are centered. Phi-psi angles for the tyrosine rotamers extracted from the protein dataset were calculated and each tyrosine was binned in its appropriate chess square based on the phi psi torsions. Then, a template model tyrosine residue was constructed at the center of each chess square with the phi and psi angles for that centroid.

Figure 4. Ramachandran plot from Ahmed et al.

The HINT force-field is based on atomistic parameters ai and Si, where ai represents the partial log Po/w, and Si represents solvent accessible surface area, for an atom i. ai is greater than 0 for hydrophobic atoms and less than 0 for polar atoms. Si is larger for atoms exposed to solvents, and near zero for atoms at the center of functional groups or fragments. The score for two atoms is given by:



Where r is the distance in A for two atoms, Tij is a function which accounts for the intrinsic acid, base properties of interacting atoms and Lij is an adaptation of the Levitt implementation of the Lennard-Jones potential function which is a function of the distance between the centers of two particles. bij > 0 represents favorable interactions, such as hydrophobic-hydrophobic or Lewis acid-Lewis base, and bij < 0 represents unfavorable interactions, such as hydrophobic-polar or Lewis acid-Lewis acid.

Then, a three dimensional box large enough to contain all possible rotamers of tyrosine was constructed. To represent 3D interaction environment of a residue, HINT basis interaction maps were calculated. The maps convert a HINT pairwise interaction into a 3D object that encodes interaction type, strength and spatial distribution of the interactions. The grid points were computed using the following equation:

Where ρ*xyz* represents the 3D map value at the given point (x, y, z), bij is the HINT interaction score between atoms i and j. xij, yij, and zij are coordinates of the midpoint of the vector between atoms i and j, and σ is a scaling factor that controls the width of the Gaussian map peak, which here was set to 0.5. Sums are computed over all pairs of interactions and separate maps were calculated for each of the four interaction classes (favorable polar, unfavorable polar, favorable hydrophobic and unfavorable hydrophobic).

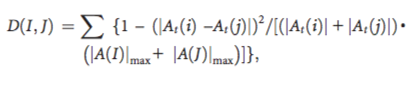
Quantitating similarity of maps and Map–map correlations

To quantitate similarity between the maps, a correlation coefficient-based metric is used. The metric compares two three-dimensional arrays of real valued points (G). Since high valued points would dominate in this calculation, the map data was transformed to log10 space:

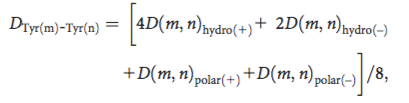
When Gt/F > 1.0, otherwise At =0

Where Gt is the value of the data-point in consideration and 1/F is a predefined floor value. Something that had to be addressed with this implementation is that since the map boxes are designed to hold all possible tyrosine confirmations, the maps under consideration will have a large proportion of zero-valued points. These would make all map pairs seem very similar. To address this issue, Boolean mask maps with the same coordinate systems were created for each map. The Boolean would evaluate TRUE when At(I) >= ( 8 \* Astddev ) or At(J) >= ( 8 \* Astddev ).

The scaled data-points that evaluated to TRUE were used in the map-map correlation similarity equation given by:



where A(i) and A(j) are point values in map I and corresponding point in map J and A(i)max and A(j)max are the maximum of the absolute values of grid points for the maps. Following this, map-map correlation scores were calculated for the individual maps for every class of interaction individually, and then combined using a weighted average in order to assess the overall map-map correlation between two structures and their environments. The weighted average is given by:



Average map, RMSD, and solvent-accessible surface area calculations

The working data for further calculations are average maps of the similarity sets**.** In order to assess side chain flexibility, the averaging formula outlined in Ahmed et al will be modified. The formula that will be used will create four subsets within two neighboring clusters. It will take into consideration the area of individual maps being averaged based on the distances between the two clusters and their centroids. The average map of each sub-region will be scored for similarity against all members of the region represented by their individual maps. Figure 5 shows two clusters A and B respectively, and r, the distance between them. r is divided into four equal parts, based on which the area of the subsets (AB1, AB2, AB3, AB4) will be determined. Tyrosines falling under each subset will be averaged individually, and each scored for similarity and clustered allowing for four new maps of tyrosine conformations. (fig. 5)

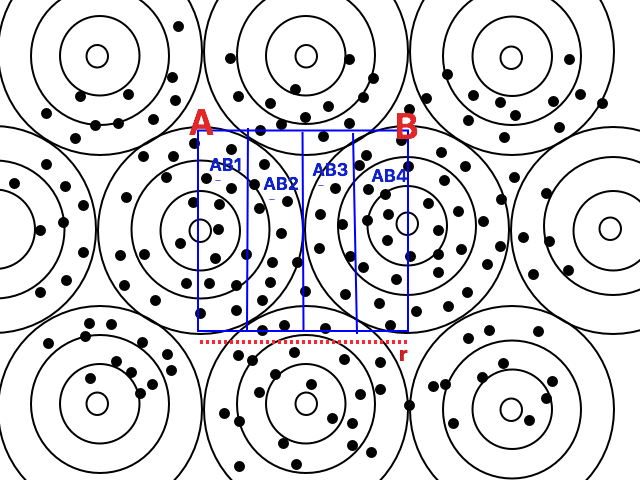
****

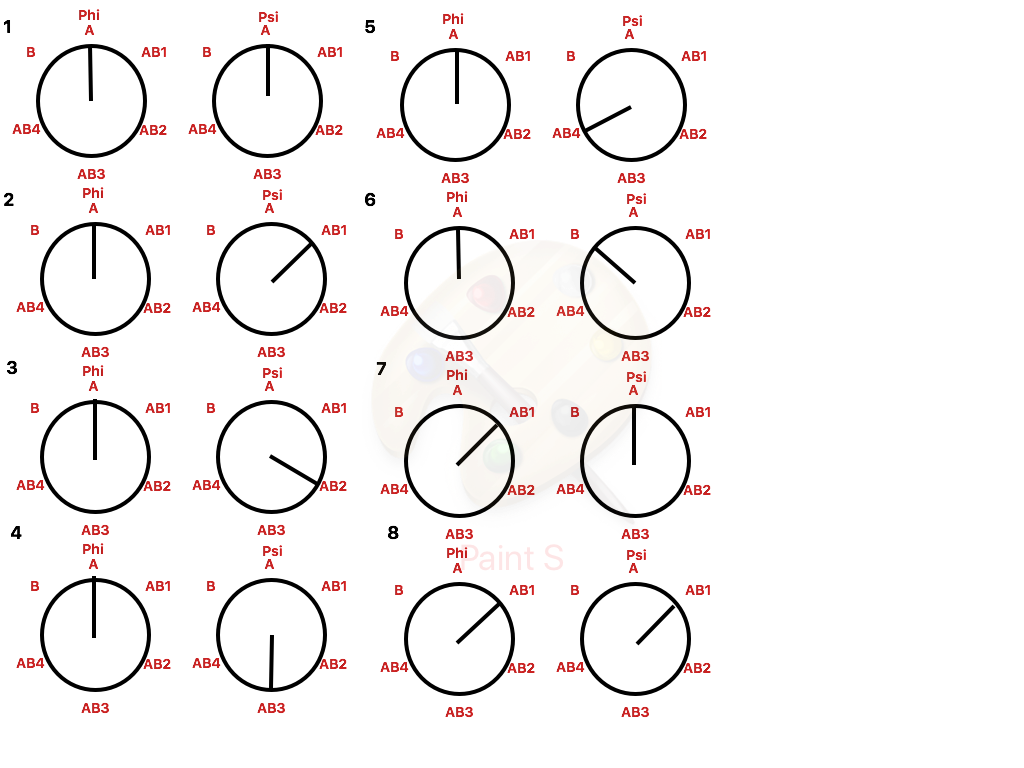
Figure 5. 2D cartoon representation of clusters with data points. Each data point  
represents a tyrosine residue and its distance from the cluster centroid.

RMSDs for the tyrosine structures are also calculated by first averaging the atomic positions from all tyrosines within that set to create an average tyrosine model for the set, followed by calculating the RMSD for each atom by name. Solvent-accessible surface areas (SASAs) were calculated for all tyrosine sidechains using the GETAREA algorithm8 with default settings.

Similarity scores will be clustered using the k-means clustering algorithm which is readily available in R (a free software environment for statistical computing and graphics) and requires a user-defined number of clusters and the data from the pairwise map correlation coefficient. The number of clusters will also be computed in R using the gap statistic which finds optimal clustering solutions based on proximity.

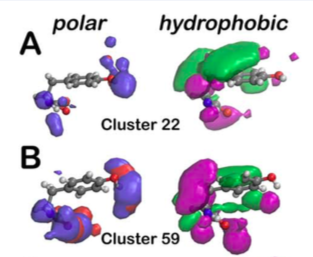
Conformational Analysis

After creating the maps of Tyrosine paths, atomic energy barriers will be evaluated by calculating the enthalpy of each path. Enthalpy is a measure of the total energy of a system, *i.e.*, the sum of the internal energies of the solute and solvent and the amount of energy required to make room for the system. For a binding process the binding enthalpy, reflects the energy change of the system when the ligand binds to the protein showing how likely the system is to bind. Systematic search, also called grid search procedure will be used on each conformational path. Systematic search procedures explore conformational space by varying torsion angles, the dihedral angles phi and psi, in one or more molecular structures, or by varying the relative displacements and orientations of several molecular structures. This is the most thorough method for searching conformational energy space. Energy is sampled over the entire range of bond rotations at regularly spaced intervals. The intervals, here, being the four maps generated above, the start point of the rotations is the average map of cluster A, the end point is the average map of cluster B. Rotations are carried out in a clockwise manner, with the phi angle rotating “slower”, and the psi angle rotating “faster”. When the psi angle finishes a cycle of its rotation the phi angle increments. This is illustrated in the following image:



Results and Discussion

The result of this study are 3D maps of tyrosine rotamer paths from one conformation to another along with the description of the hydropathic environment surrounding the residues\*. The path maps also allowed for conformational analysis to be performed. If only one angle had to be varied, meaning that the orientation of the other angle remained the same for both residue conformations, the result is going to look like a sin curve, but when two angles get varied the result is going to produce a high-dimensional surface implying how likely the rotamers are to assume each conformation. As mentioned 90% of amino acid residues undergo a change in side chain conformation when binding to a ligand. With the conformational analysis into consideration alongside the hydropathic maps, quantified paths of conformational change of an amino acid side-chain upon binding with an receptor can be taken into consideration in protein-ligand docking, allowing for more accurate results with prediction of ligand-receptor binding.   
.If methods described here were carried out on all 20 amino-acids and by result side-chain flexibility is quantified for each, the side chains of amino acids can efficiently be optimized for molecular docking at a low cost. Alternatively, if low energy paths between two residues don’t exist, this would show that the side-chains in question are not flexible for certain path and thereby those path not optimal for protein-ligand docking, since binding would not be ideal or most likely would not occur. In this case, the 3D maps of a residues conformational path can still be useful in further examination of amino acid side-chain behavior with the knowledge of favorable polar, unfavorable polar, favorable hydrophobic and unfavorable hydrophobic environments of the residues and the limitations the environments might present.



\*Since this is still a rough draft I did not get a chance to yet hypothesize and generate the four new maps of the paths resulting from this study. Here I have included an image of an example paths centroid Gaussian-weighted average three-dimensional contour maps of the start point (cluster A (22)) and end point (cluster B (59)) Left of each pair: polar maps (favorable—blue; unfavorable—red); right of each pair: hydrophobic maps (favorable—green; unfavorable—purple). For the finalized version of my proposal maps of the path will be added.

Bibliography

1.  http://www.intechopen.com/books/protein-engineering-technology-and-application/protein-protein-and-protein-ligand-docking

2. Du, X., Li, Y., Xia, Y. L., Ai, S. M., Liang, J., Sang, P., Ji, X. L., … Liu, S. Q. (2016). Insights into Protein-Ligand Interactions: Mechanisms, Models, and Methods. International journal of molecular sciences, 17(2), 144. doi:10.3390/ijms17020144

3. Kuhn, L. A. (no date) “Chapter 10. Strength in Flexibility: Modeling Side-Chain Conformational Change in Docking and Screening,” Computational and Structural Approaches to Drug Discovery RSC Biomolecular Sciences, pp. 181–191. doi: 10.1039/9781847557964-00181

4. Gaudreault, F., Chartier, M., & Najmanovich, R. (2012). Side-chain rotamer changes upon ligand binding: common, crucial, correlate with entropy and rearrange hydrogen bonding. Bioinformatics (Oxford, England), 28(18), i423-i430.

5. Ahmed, M. H., Koparde, V. N., Safo, M. K., Scarsdale, J. N. and Kellogg, G. E. (2015) “3d interaction homology: The structurally known rotamers of tyrosine derive from a surprisingly limited set of information-rich hydropathic interaction environments described by maps,” Proteins: Structure, Function, and Bioinformatics, 83(6), pp. 1118–1136. doi: 10.1002/prot.24813

6. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucl Acids  Res 2000;28:235–242.

7. The 20 Amino Acids: hydrophobic, hydrophilic, polar and charged amino acids. Available at: https://proteinstructures.com/Structure/Structure/Ramachandran-plot.html

8. Raczkiewicz R, Braun W. Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. J Comp Chem 1998;19:319–333.