Using Bayesian Statistics to Predict Water Affinity and Behavior in Protein Binding Sites

J. Andrew Surface

Hampden-Sydney College / Virginia Commonwealth University

Introduction

In the past several decades drug development has come a long way in testing assays of medicinal compounds for drug activity potential. Particularly, the areas of development have been in computerized molecular modeling and *in silico* assays that enable scientists to test thousands of drug compounds simultaneously for drug potential on various biological molecules, including proteins. Proteins have been of particular interest as drug targets because proteins control body function and have many specific drug targets that are highly specific to certain biological systems within the body. Highly specific targets are ideal for medicinal research because they have the lowest potential for negative side-effects. Simulated docking procedures have been developed to fully understand how ligands bind to proteins and these systems have enabled medicinal chemists to build ligands that have high biological activity on proteins. Understanding docking sites and docking mechanisms have given chemists even more information on the process of making ligand-protein complexes with desirable properties.

One aspect of ligand-protein binding that has not been fully understood to date is the role that water plays in both the docking process of ligands and the stasis energetics of proteins. Biological environments are aqueous, and proteins are not only full of crevices and caverns, ideal for water molecules to "slip" into, but also are covered with both hydrophobic and hydrophilic regions. These regions attract or inhibit water from interacting with the protein's different surfaces, and the multiple hydrophilic regions produce multiple H-bonding sites, filled with both H-bond acceptors and donors. The different regions on a protein that are filled with waters can be strengthened or weakened depending upon the arrangement of water molecules within the protein. They can furthermore inhibit or aid (by building H-bonding bridges between ligand and protein) by slowing the conformational change of a protein needed during a docking procedure of a ligand or by being in the ligand's way within the docking site on a protein. Obviously, then, water is a large contributor to protein conformational change and complexing abilities, both promoting and inhibiting it in various ways.

Understanding how water interacts with the protein is therefore necessary to fully understand how the protein and different ligands interact in different biological environments. Further understanding the water/protein relationship and incorporating that knowledge into a molecular modeling program would greatly increase the accuracy of *in silico* assays in medicinal chemistry applications, aiding structure-based drug design.

To date, much work has been done by Dr. Glen Kellogg at Virginia Commonwealth University, as well as many others. The primary methodology developed to date has been a system known as HINT scoring (Hydropathic INTeractions). HINT is a developed method of scoring water molecules by their hydropathic interactions, giving a number that shows the strength of the interactions that the water molecule has with the protein at a given site. The HINT score is a number developed from a non-Newtonian forcefield that is based on the partition constant between 1-octanol and water (log $P_{o/w}$) [1, 2]. It works as follows:

$$\sum_i \sum_j \mathbf{b}_{ij} = \sum_i \sum_j (\mathbf{a}_i \mathbf{S}_i \mathbf{a}_j \mathbf{S}_j \mathbf{T}_{ij} \mathbf{R}_{ij} + \mathbf{r}_{ij})$$

Figure 1 [1]

The HINT equation is a simple representation of the interaction between two atoms [1]. *i* and *j* are the two atoms represented in the above equation. $\mathbf{b}_{ij}$ is the interaction score between the two atoms, **a** is the hydrophobic atom constant, **S** is the atomic solvent-accessible surface area, $\mathbf{T_{ij}}$ is a logic function that assumes the polar nature of interacting atoms (-1 or +1), and $\mathbf{R_{ij}}$ is an exponential function that relates the distance between the two atoms *i* and *j*. $\mathbf{r_{ij}}$ tells the distance between *i* and *j* but is related to the Lennard-Jones function [1].

Inherent within the HINT model is an incorporation of enthalpy, entropy, salvation and other energies, although they are not fully quantifiable [4]. As such, HINT score is a valuable number that represents a combinatorial data set that aids in estimating water behavior within a protein.

Another factor that can be taken into account when trying to understand th ebehavior of water molecules in proteins is called the Rank. The Rank is essentially a method that measures the number of H-bonds that a water molecule has the ability to make in a given location [3]. The Rank algorithm works as follows:

$$\text{Rank} = \sum_n \left\{ (2.80\ \text{Å}/r_n) + \left[\sum_m \cos(\theta_{Td} - \theta_{nm}) \right] \Big/ 6 \right\}$$

Figure 2 [1]

Where 2.80 A is the length of an ideal hydrogen bond, i.e., between the water and an external donor or acceptor, and r_n is the actual distance between the water molecule and the target hydrogen binding site on the protein. *n* is maximally 4, as that is the maximum number of H-bonds for a single water molecule. θ_{Td} is the ideal tetrahedral angle between all of the hydrogen bonds and θ_{nm} is the actual angle. This value is divided by 6 because there is a maximum number of 6 angles.

Experimentally, Rank and HINT have both been used in various applications to aid medicinal chemists in classifying water behavior in certain proteins. What has not been done, to date, is relating the Rank and the HINT score together with the goal of being able to predict the probability of a particular water molecule's behavior in any region of the protein when the protein is complexed. It is believed that using the Rank values and the HINT scores of various waters in various proteins would lead to several equations whose results could be weighted and then run through a statistical filter, out of which would come the percent-probability of water being retained in a protein. Since HINT score and Rank both incorporate the necessary data to calculate such problems, it is believed that this is the only data needed to compute the probability.

Methods

The proposed method of developing the algorithm to predict the water molecule behavior is as follows: to first prepare a training set of data of proteins that contain water molecules of known HINT score and Rank. This training set will then be used to develop the functions that will relate HINT score and Rank to percent-conserved waters. Using these two functions, a proper mathematical weight will be established for each one before using the results of each function in a Bayesian statistical function that will take the two percent-conserved values predicted from both Rank and HINT score functions, and output one total percent-conserved value that will establish the likelihood of the individual water being conserved.

Training sets have already been established to develop the HINT score function and the Rank function. They can be seen below:

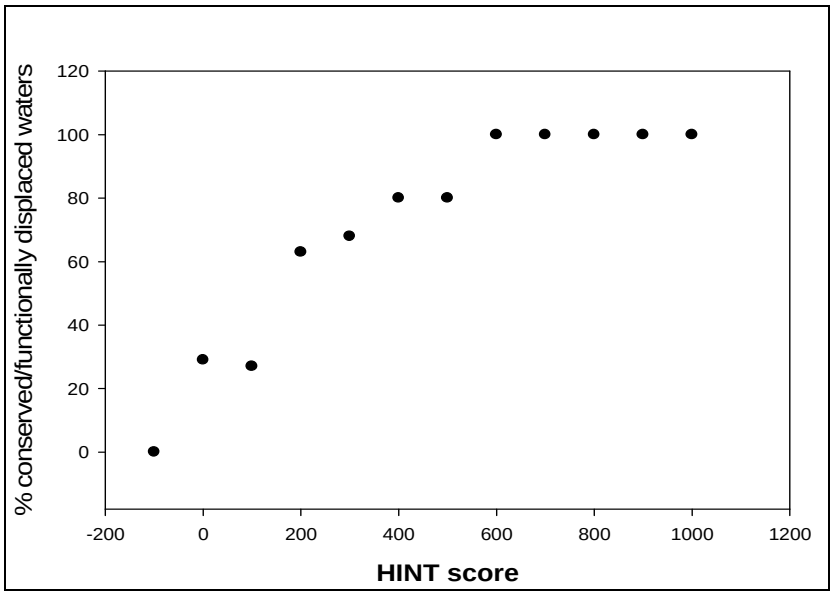

Figure 3

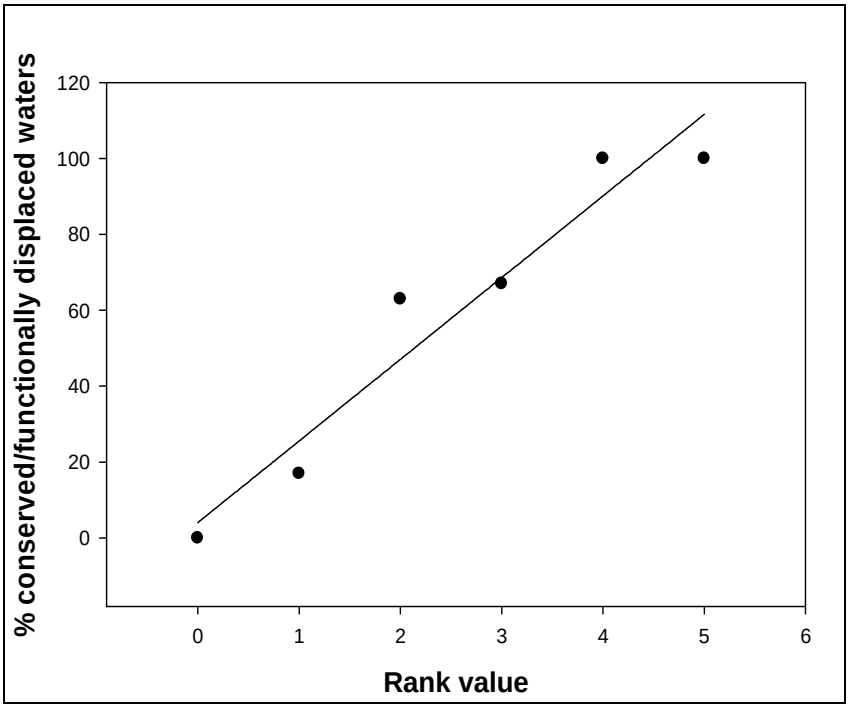

Figure 4

The Bayesian statistical analysis will also incorporate a training set of waters with known Ranks and HINT scores. This data set will use to further "train" the Bayesian algorithm.

Once the Bayesian algorithm has been designed based on the training set, another set of selected waters in proteins with known affinities will be tested against the algorithm by comparing the predicted percent-conserved water prediction of the algorithm from the realistic values. This will test the accuracy of the function.

Potential Results

Using Bayesian statistics to analyze the percent-conservation of waters has the potential of giving molecular modeling programs the ability to not only know the properties of water molecules in a particular protein, but also *how* they will affect the ligand-protein binding process by determining if they will remain (be conserved) or not after the protein has been complexed. The goal is to get the predictive Bayesian algorithm within 90% or greater accuracy. This will be an excellent and very useful addition to the HINT tool in molecular modeling programs such as Sybyl.

References:

[1] Amadasi, A., Spyrakis, F., Cozzini, P/, Abraham, D/J., Kellogg, G.E., and Mozzarelli, A. (2006). Mapping the Energetics of Water-Protein and Water-Ligand Interactions with the "Natural" HINT Forcefiled: Predictive Tools for Characterizing the Roles of Water in Biomolecules.*J. Mol. Bio.* **358**, 289-309.

[2] Kellogg, G. E., Semus, S.F. & Abraham, D.J. (1991). HINT: a new method of empirical hydrophobic field calculation for CoMFA. *J. Comput. Aided Mol. Des.* **5,** 545-552.

[3] Chen, D/L. & Kellogg, G. E. (2005). A computational tool to optimize ligand selectivity between two similar biomacromolecular targets. *J. Comput. Aided Mol. Des.* **19,** 69-82.

[4] Kellogg, G.E., Abraham, D. J. (2000). Hydrophobicity: is Log$\mathbf{P}_{o/w}$ more than the sum of its parts? *Eur. J. Med. Chem.* **35,** 651-661.