Ideal Ligand Substituent Prediction Using HINT for the Purpose of De Novo Ligand Design

James A. Surface, Glen E. Kellogg

Institute of Structural Biology and Drug Discovery, Virginia Commonwealth University, Richmond, VA 23284

August 6th, 2008; andy@vanderwaal.net

An exploratory first step *de novo* ligand prediction program is made and tested from two different fragment-based ligand design methodologies using the HINT force field model and a statistically validated water relevancy algorithm based on HINT Scores and another geometrically-based protocol called RANK. One methodology used water molecules as starting points for fragment optimization and novel ligand assembly while the other method used a 3-D hydrophobic (or acid/base) map, using the extreme points on these maps as starting points for the molecules tagged as best for those specific interactions. Throughout this study a new fragment-based drug design protocol is tested and results are compared to existing ligand binding data, and it was found that while using water molecules as starting points don't correctly predict actual ligand substituent groups, it might be used as a proper protocol to predict optimizations to currently existing ligands by superimposing pre-ligand and post-ligand binding events. More development is needed for the second methodology, while the development of both developed novel database organizations and several algorithms and data-storing techniques for preserving the information found from these particular calculations.

Introduction

The design of new drugs is an extensive and ever-growing problem. As simple happenstance drug discoveries have waned and with the advent of more detailed information being learned about the particulars in certain disease mechanisms, the need has come about for designing specific drug molecules to address particular identified target sites.

The most common and easiest handles for drug targets are proteins. Proteins are excellent targets because they control nearly every mechanism in your body, are large and bulky, and contain numerous handles for binding drug molecules in ways that alter their own function and behavior. Additionally, more and more is being learned about proteins each day as more protein structures are elucidated and understood through X-Ray Crystallography, Electron Microscopy, and Nuclear Magnetic Spectroscopy. As these structures are elucidated and as the mechanisms they control are categorized and mapped, control targets are developed for which specific drug molecules are needed.

Many novel ways have been developed for understanding how ligands will bind to proteins and from thence control and effect the protein's behavior, controlling the mechanism it manages. These efforts have largely been put forth to create high throughput, inexpensive, computational assay calculations[1] that will enable chemists to narrow what compounds they need to synthesize from millions to a few hundred.

One method that has been used in the past for understanding these ligand-protein interactions has been specifically studying proteins before and after known binding ligands have been attached, usually through their crystal structure.[2] These studies specifically exploit actual ligand binding information as a predictive element for understanding how a new protein who's ligand information is unknown might bind and what it might bind to. As there are many more proteins of which little is known about its binding mechnism than there are proteins that detailed binding mechanism is understood, this method, although valuable in some areas, is inherently limited. Other methods must then be explored.

One such other method, known as PRO_LIGAND[3], uses a novel algorithm to conduct a study of known pharmacophores and make generalizations of the binding data which is then used as a predictive element for understanding other similar interactions. Essentially trained data is used to come up with a more applicable novel result. The method that it uses for ligand-protein interactions is called *docking*, and many such other docking studies have been done using this general concept.

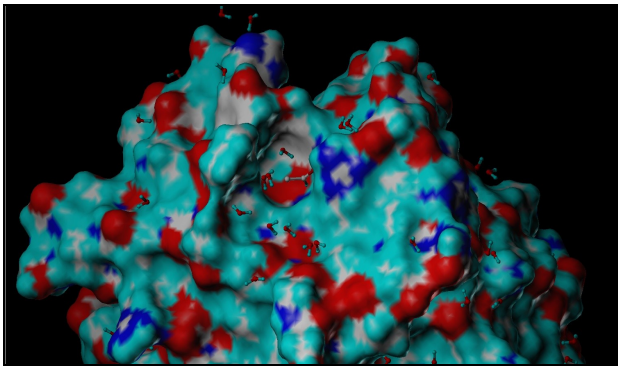

Fig. 1

Proteins are good drug targets because they are large, bulky, and possess many good binding sites. This protein, 2CTV, as most of the other proteins in the human body, is surrounded by water molecules. A drug molecule binds in the cavity and is just visible.

Essentially docking mechanisms make use of some kind of scoring technique based on some kind of factors that contribute to energy and they do so by computationally moving the ligand around in the target area of a protein until the most energetically favorable conformation is reached. Studies have been made evaluating different scoring functions[4] for these docking algorithms as well as conducting a general comparison the different docking algorithms themselves, which are built on differing assumptions and models.[5] Some of these automated docking models are better than others, and the ranking of these models are generally situationally specific.

To make some of these models better, some have focused on building better scoring mechanisms that are less situationally dependent. Others have focused on synthesizing the various docking methods together in a combinatorial way with varied successes.[6,7] Generally, many such specific studies have been conducted to evaluate the performance in a generalized situation with many different ligands to be docked in an attempt to gain more general knowledge about the docking procedure and the

specific system being studied. Some of these studies have even gone so far as to make predictions for ligands that would potentially bind the compounds.[8]

Alternatively, many have turned to mathematical systems that use graphs and various branched and genetic algorithms to solve the problem of compound selection, comformation, and idealization in the protein active site. Most of these studies are concerned with developing the algorithms efficient and robust enough to handle the stream of information that needs to be processed in reducing the natural complexity of ligand binding to large proteins. One such study[9] developed a unique graphed algorithm to more efficiently classify and scan potential compounds and protein active sites, while another developed the optimization mathematics necessary to solve a similar problem.[10]

Another area in which much recent work has been accomplished is that of the interaction of water and proteins. The human body is well known as an aqeuous environment as water abounds around nearly every chemical reaction occuring in the body. Water also affects nearly every protein's creation, conformation, and activity. Thus, studying the interaction of water with proteins could demonstrable valuable information about ligand-protein interaction, protein interaction, and other protein mechanisms.

Some of these water-protein studies have focused on the general water molecules found in the crystal structures, and how their interactions can be understood within the matrix of the protein.[11] These studies have demonstrated many times over that water is also an affector of ligand binding in active sites of target proteins.[12,13] Klebe notes that about two-thirds of all liganded proteins have at least one water molecule mediating the ligand binding phenomenon.[14]

As water is a demonstrated affector of ligand-protein interactions, many have tried various methods that attempted to classify water molecules and understand how they mediate and affect ligand-protein interactions. One such study, as conducted by Essex, attempted to use bayesian probability and extensive computational models that attempt to predict individual water molecules roles in the ligand-protein binding mechanism.[15] Other studies have attempted to use water molecules as further handles for predicting ligand optimizations by the addition of substituent groups to these ligands with the energy necessary to replace the water molecules.[16] These models assume that water molecules are in a particular position within a protein because they are "happy" there, presumably making some kind of polar bond. If polar side-groups are added to the known ligand structure that would be more optimal than the water molecule, presumably the ligand would have another handle with which it can bind to the protein. This additional handle would ideally boost the affinity of the drug molecule, making its interaction with the protein stronger and more ideal.

Nevertheless, another method in which many recent developments have been made is better understanding what the water-protein interactions say about protein active sites, particularly understanding the effect of the hydrophobic effect[17] on ligand-protein interactions. As recently demonstrated by Amadasi *et al.*[18] HINT, a computational protocol that scores hydropathic interactions proportionally to free energy[19] based on the LogP coefficient and other geometrical considerations,[20,21] can be used in combination with another geometrically validated protocol, RANK,[22] to accurately identify water molecules that will play a significant role in the binding of the ligand. This model, which will not be described in great detail here due to space limitations, attempts to use these very hydrophobic interactions to create relevant, simple, and intuitive predictive calculations that shed light on the nature of the protein and the ligand the interaction the two have together. These significant water molecules can be considered "relevant" in protein active sites and are important in docking and binding considerations.[23] Additionally, the "irrelevant" waters, molecules that are energetically replaced in protein active sites during a ligand binding event, are also supposed to be energetically indicative of the probabilistic ligand binding structure. Examination of crystalized structures (both uncomplexed and complexed) show replaced waters and indicate that ligand/protein interactions are more energetically favorable than the interactions the previous occupying water molecules had.[6] This indicates ligand binding information "stored" within the mapped energetics of water molecules currently occupying protein structures.

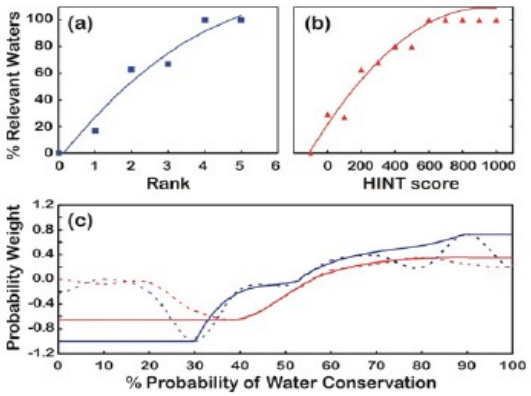

Fig. 2

HINT Score and Rank together provide an overall probability system that is predictive of water molecule relevancy.[18]

Given the demonstrated statistical relationship between Rank, HINT, and water molecule relevancy during a ligand binding event,[23,24] it is presumably safe to assume that Rank and HINT can provide a hydropathic description of the active site of a protein and give a three-dimensional predictive map of the active site that can be used to provide information about the sort of ligand needed to best fill the cavity and bind affecting the protein.

The experiments conducted in this paper expound and develop two methodologies that provide a proof-of-concept first step extraction of the information presumed to be contained within water molecules native to uncomplexed proteins.

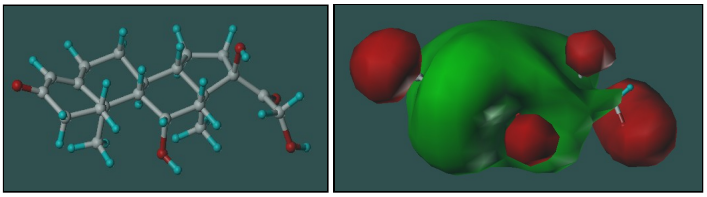

Fig. 3

The hormone and ligand cortisol, as portrayed (left) without the the HINT Map of the HINT force field and (right) with it. The red regions correspond to hydrophillic regions and the green regions correspond to hydrophobic regions.

Before introducing the two unique developmental methodologies discussed in this paper, it is important to note that this study of ligand-protein behavior does not attempt to incorporate protein flexibility and protein conformation into the results. The proteins specifically chosen in the data set were proteins that had little conformational change from their uncomplexed forms to their complexed forms. This system was chosen so as to minimize the error added to a system by protein conformational change, unlike other studies which have attempted to include protein conformation analysis into the ligand binding phenomenon.[25]

The first methodology uses water molecules as a starting points for building novel ligands by replacing the single water molecules with substituent groups carefully chosen and built into an assay database. Using the HINT force field, the substituent is optimized and scored to provide the strength of interactions it has with the protein in its current position. These results can then be compared to the water molecule it has superceded.

The second methodology involves a 3-D hydrophobic map (or alternatively an acid/base map) using the extreme points on these maps (corresponding to places of strong hydrophobic/ hydrophillic or strong acid/base interactions) as starting points for aligning the molecules, optimizing them, and building initial substituent locations for novel ligands.

Both methodologies involve starting with a small fragment typical to many kinds of ligand molecules, placing them and optimizing them in those specified locations to build a first-step probabilistic map of good substituent group locations for a novel ligand binding to that particular protein. Increased numbers of specific substituent groups to the database necessarily increases the accuracy and specifity of the database, and the results from these initial calculations are tabulated energy (HINT score) and position/conformation values (cartesian coordinates and polar coordinates) given the individual interaction of that substituent group with the localized portion of the protein nearest the substituent.

Ideally, this proof-of-concept methodology would create an "outline" of substituent groups that are energetically situated on specific handles located within the protein active site, handles that could be "joined" to create a novel ligand using a linker database of center groups of molecules on which substituents could be attached in various locations. Although this paper does not addres the linker problem, it does attempt to describe via proof-of-concept programming how the probabilistic substituent prediction methodoligies work to create a "substituent map" for most probable ligands.

Results and Discussion

The following experimental calculations were made on Linux, dual-core HP computers that were linked over a network. The version of the HINT software was 3.12, along with some initial experimental versions of 3.13. The molecular modeling program, Sybyl v7.3 and v8.0 were used for the modeling and calculations made. The substituent replacement protocol was made using the Sybyl front-end interface, and later automated using the SPL (Sybyl Programming Language) interface to increase the accuracy and speed of the calculations being made. All calculations were run on Sybyl v7.3. Most graphics for this paper were made in v8.0.

The database of substituents for both developed methdodologies was made using the Sybyl Database Framework, and Table 1 describes the substituents used in the database.

DB Code Name	Actual Group
ALDO	*Aldehyde*
AR	*Aromatic Ring*
CH3	*Methyl Group*
CONH2	*Acetamide*
COO	*Carboxylate*
COOH	*Carboxylic Acid*
NCH	*Imino*
NH	*Cyano*
NH2	*Amino*
NH3	*Ammonio*
OH	*Alcohol*
SH	*Thiol*

Table 1

The table of substituents used in the database for water replacement.

Two databases were made, one for the Water Target methdology and one for the HINT Map Target methodology. The Water Target database was designed so that the dummy carbon atom and the atom of the substituent attached to it would have certain id numbers so that the program could center the atom attached to the dummy carbon on the oxygen of the water molecule being targeted. The HINT Map Target database uses Sybyl Dummy atoms to deliniate localized acid/base and hydrophobic/hydrophillic targets. The following describes the methods used to write the two SPL programs.

Water Targets

The automated SPL program was written according to the following protocol:

1) Set up environmental variables for calculation
2) Identify target waters
3) Center of Gravity (COG) of selected waters calculated (Fig 4, attached)
4) Substituents from pre-defined database are assayed
5) Each substituent is then run through a subroutine (see Figs. 5-7, attached):
    1. Move substituent (center over target water)
    2. Optimize using HINT Force Field
    3. Bring Carbon Tail around to proper configuration
    4. Take HINT Intermolecular Score, record
    5. Calculate distance and three-dimensional angle of carbon tail.
6) Combine all intermolecular scores from assay of database together into one file.
7) Close program, database, and results file.

Every intermolecular score was stored in its own .tab file and the results file was stored into a .csv file that could be opened by a spreadsheet program like EXCEL to calculate results.

Each water molecule was evaluated and substituents that possessed better scores (higher scores) than the water molecule were reloaded into Sybyl and superimposed on top of the actual real ligand that bound to the protein to see how the substituent groups matched with the real ones on the actual ligand.

Only substituents with high scores were analyzed as the HINT score has an inverse relationship with ΔG. Lower ΔG values mean greater affinity, thus higher HINT score values mean greater affinity. The hope was that the method would correctly predict actual substituents. The reality was the ligands in the structures never exactly replaced a water molecule with its substituents so the testing system was directly limited in that the substituents would be "stuck" where the waters were positioned.

Interestingly, all of the best scoring substituents were correctly positioned (their carbon tails were pointed towards or outwards of the active site, in such a way that it could easily bind to a linker or another drug molecule. Some of the attached figures (Figs. 8-11) demonstrate the predictions made by the program.

HINT Map Targets

One of the continuing portions of this project is the planned calculations on the HINT Map Targets, as described earlier in this paper. A preliminary SPL program was written that correctly ran automated substituent translation and HINT optimization onto the HINT Map Targets. Preliminary results were inconclusive as the coordinate systems for the HINT Map and the HINT Partition did not match properly, causing errors. More HINT Maps need to be generated to be able to run the SPL program and do more analysis.

A novel database organization method was designed using the Sybyl Database Format as in the previous water target SPL program. This format, unlike the previous database, incorporated information that included acid/base and hydrophobic/hydrophillic targets within the substituent molecules so that their acid/base or hydrophobic/hydrophillic portions could be aligned directly onto the corresponding acid/base or hydrophobic/hydrophillic targets on the HINT Map to assist with HINT Optimization and idealize the interaction between the substituent and the protein. These targets within the substituents were made using Sybyl Dummy Atoms which the SPL program would delete after using their coordinates as reference points for the translation and optimization of the molecule.

Conclusions

While there are no conclusively ideal results, the results of the paper did reveal valuable information: namely that water molecules are not necessarily the best starting points for novel ligand prediction but that they might be good points for working predictive ligand optimizations.

Initial results are not yet complete for the mapped acid/base and hydrophobic/hydrophillic points. The program for this method has been debugged and tested on one protein system which just included the uncomplexed protein. Initial optimization results for this experiment do reveal that having multiple targets (acid/base and hydrophobic/hydrophillic targets) help the optimization algorithm by pre-starting it in its near ideal location with the protein. This method increases reproducibility and theoretically increases the accuracy of the calculations.

The results of this study also demonstrate successfully the eventual feasability of a fragment-based *de novo* drug design program based on the HINT Force Field. Initial results, while not ideal, were indicative enough to merit further research and programming.

Eventually, when the substituent mapping is made complete, a linking system will be programmed. It is suspected at this branched optimization problem will involve some kind of branched tree structure solution or the use of some kind of genetic algorithm. Many studies have already demonstrated how this particular problem might be eventually solved. One such study[26] uses a breadth-first, branched tree procedure algorithm for determining maximum substructures through a potential combinatorial means of different potential ligand attributes. Another has developed a novel moethod of aligning and matching drug ligand molecules using surface shape similarity.[27] This particular method may be valuable for its ability to "match" different attributes of molecules together, helping to find the correct linking compounds to fit in the proper orientation within the protein active site. A final proposed solution for the problem may be the use of genetic algorithms. Another particular study14 makes use of genetic algorithms and demonstrates how they are good for molecular recognition and design by solving combinatorial optimization problems using the genetic algorithm's mutations and crossover operators.

However the problem may be solved, the research discussed in this paper demonstrates how the research might be extended to the point of designing successful linker molecules to complete the *de novo* molecule prediction algorithm using the HINT Force Field model.

Acknowledgements

Bioinformatics and Bioengineering Summer Institute
National Institute of Health (NIH)
National Science Foundation (NSF)
The Institute of Structural Biology and Drug Development
Dr. Glen E. Kellogg
Ashutosh Tripathi
Dr. J. Neel Scarsdale

References

[1] Kuhn, P.; Wilson, K.; Path, M.G.; Stevens, R.C. (2002). The Genesis of High-throughput Structure-based Drug Discovery using Protein Crystallography. *Current Opinion in Chemical Biology.* **6,** 704-710.

[2] Feher, M.; Gao, Y.; Baber, J.C.; Shirley, W.A.; Saunders, J. (2008). The Use of Ligand-based de novo Design for Scaffold Hopping and Sidechain Optimization: Two Case Studies. *Bio. & Med. Chem.* **16,** 422-427.

[3] Waszkowycz, B.; Clark, D.E.; Frenkel, D.; Li, J.; Murray, C.W.; Robson, B.; Westhead, D.R. (1994). PRO_LIGAND: An Approach to *de Novo* Molecular Design. 2. Design of Novel Molecules from Molecular Field Analysis (MFA) Models and Pharmacophores. *J. Med. Chem.* **37,** 3994-4002.

[4] Xing, L.; Hodgkin, E.; Liu, Q.; Sedlock, D. (2004). Evaluation and Application of Multiple Scoring Functions for a Virtual Screening Experiment. *J. Comp.-Aided Molecular Design.* **18,** 333-344.

[5] de Graaf, C.; Pospisil, P.; Pos, W.; Folkers, G.; Vermeulen, N.P.E. (2005). Binding Mode Prediction of Cytochrome P450 and Thymidine Kinase Protein-Ligand Complexes by Consideration of Water and Rescoring in Automated Docking. *J. Med. Chem.* **48,** 2308-2318.

[6] Charifson, P.S.; Corkery, J.J.; Murcko, M.A.; Walters, W.P. (1999). Consensus Scoring: A Method for Obtaining Improved Hit Rates from Docking Databases of Three-Dimensional Structures into Proteins. *J. Med. Chem.* **42,** 5100-5109.

[7] Bissantz, C.; Folkers, G.; Rognan, D. (2000). Protein-Based Virtual Screening of Chemical Databases. 1. Evaluation of Different Docking/Scoring Combinations. *J. Med. Chem.* **43,** 4759-4767.

[8] Warren, G.L.; Andrews, C. W.; Capelli, A-M.; Clarke, B.; LaLonde, J.; Lambert, M.H.; Lindvall, M.; Nevins, N.; Semus, S.F.; Senger, S.; Tedesco, G.; Wall, I.D.; Woolven, J.M.; Peishoff, C.E.; Head, M.S. (2006). A Critical Assessment of Docking Programs and Scoring Functions. *J. Med. Chem.* **49,** 5912, 5931.

[9] Gardiner, E.J.; Artymiuk, P.J.; Willett, P. (1997). Matching Three-Dimensional Molecular Structures. *J. Mol. Graph. & Mod.* **15,** 245-253.

[10] Gardiner, E.J.; Willett, P. (2000). Graph-Theoretic Techniques for Macromolecular Docking. *J. Chem. Inf. Comput. Sci.* **40,** 273-279.

[11] Lu, Y.; Wang, R.; Yang, C-Y.; Wang, S. (2007). Analysis of Ligand-Bound Water Molecules in High-Resolution Crystal Structure of Protein-Ligand Complexes. *J. Chem. Inf. Model.* **47,** 668-675.

[12] Levitt, M.; Park, B.H. (1993). Water: Now You See it, Now You Don't. *Structure,* **1,** 223-226.

[13] Cozzini, P.; Fornabaio, M.; Marabotti, A.; Abraham, D.J.; Kellogg, G.E.; Mozzarelli, A. (2002). *J. Med. Chem.* **45,** 2469-2483.

[14] Klebe, G. (2003). Virtual Ligand Screening: Strategies, Perspectives, and Limitations. *Drug Discovery Today.* **11,** 580,594.

[15] Barillari, C.; Taylor, J.; Viner, R.; Essex, J.W. (2007). Classification of Water Molecules in Protein Binding Sites. *J. Am. Chem. Soc.* **129,** 2577-2587.

[16] Raymer, M.L.; Sanschagrin, P.C.; Punch, W.F.; Venkataraman, S.; Goodman, E.D.; Kuhn, L.A. (1997). Predicting Conserved Water-mediated and Polar Ligand Interactions in Proteins Using a K-nearest-neighbors Genetic Algorithm. *J. Mol. Bio.* **265,** 445-464.

[17] Doolittle, R.F.; Kyte, J. (1982). A Simple Method for Displaying the Hydropathic Character of a Protein. *J. Mol. Biol.***157,** 105-132.

[18] Amadasi, A.; Surface, J.A.; Spyrakis, F.; Cozzini, P.; Mozzarelli, A.; Kellogg, G.E. (2008). Robust Classification of "Relevant" Water Molecules in Putative Protein Binding Sites. *J. Med. Chem.* **51,** 1063-1067.

[19] Spyrakis, F.; Amadasi, A.; Fornabaio, M.; Abraham, D.J.; Mozzarelli, A.; Kellogg, G..E.; Cozzini, P. (2007). The Consequences of Scoring Docked Ligand Conformations Using Free Energy Correlations. *Eur. J. Med. Chem.* **42,** 921-933.

[20] Verdonk, M.L.; Chessari, G.; Cole, J.C.; Hartshorn, M.J.; Murray, C.W.; Nissink, J.W.M.; Taylor, R.D.; Taylor, R. (2005). Modeling Water Molecules in Protein-Ligand Docking Using GOLD. *J. Med. Chem.,* **48,** 6504-6515.

[21] Cozzini, P.; Fornabaio, M.; Marabotti, A.; Abraham, D.J.; Kellogg, G.E.; Mozzarelli, A. (2002). Simple, Intuitive

Calculations of Free Energy of Binding for Protein-Ligand Complexes. 1. Mmodels without Explicit Constrained Water. *J. Med. Chem.* **45,** 2469-2483.

[22] 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.

[23] Amadasi, A.; Surface, J.A.; Spyrakis, F.; Cozzini, P.; Mozzarelli, A.; Kellogg, G.E. (2008). Robust Classification of "Relevant" Water Molecules in Putative Protein Binding Sites. *J. Med. Chem.* **51,** 1063,1067.

[24] Amadasi, A.; Spyrakis, F.; Cozzini, P.; Abraham, D.J.; Kellogg, G.E.; Mozzarelli, A. (2006). Mapping the energetics of water-protein and water-ligand interactions with the "natural" HINT forcefield: precictive tools for characterizing the roles of water in biomolecules. *J. Mol. Biol.* **358***,* 289-309.

[25] Firth-Clark, S.; Kirton, S.B.; Willems, H.M.G.; Williams, A. (2008). De Novo Ligand Design to Partially Flexible Active Sites: Application of the ReFlex Algorithm to Carboxypeptidase A, Acetylcholinesterase, and the Estrogen Receptor. *J. Chem. Inf. Model.* **48,** 296-305.

[26] Brint, A.T.; Willett, P. (1987). Algorithms for the Identification of Three-Dimensional Maximal Common Substructures. *J. Chem. Inf. Comput. Sci.* **27,** 152-158.

[27] Cosgrove, D.A.; Bayada, D.M.; Johnson, A.P. (1999). A Novel Method of Aligning Molecules by Local Surface Shape Similarity. *J. Comp. Aid. Mol. Des.* **14,** 573-591.

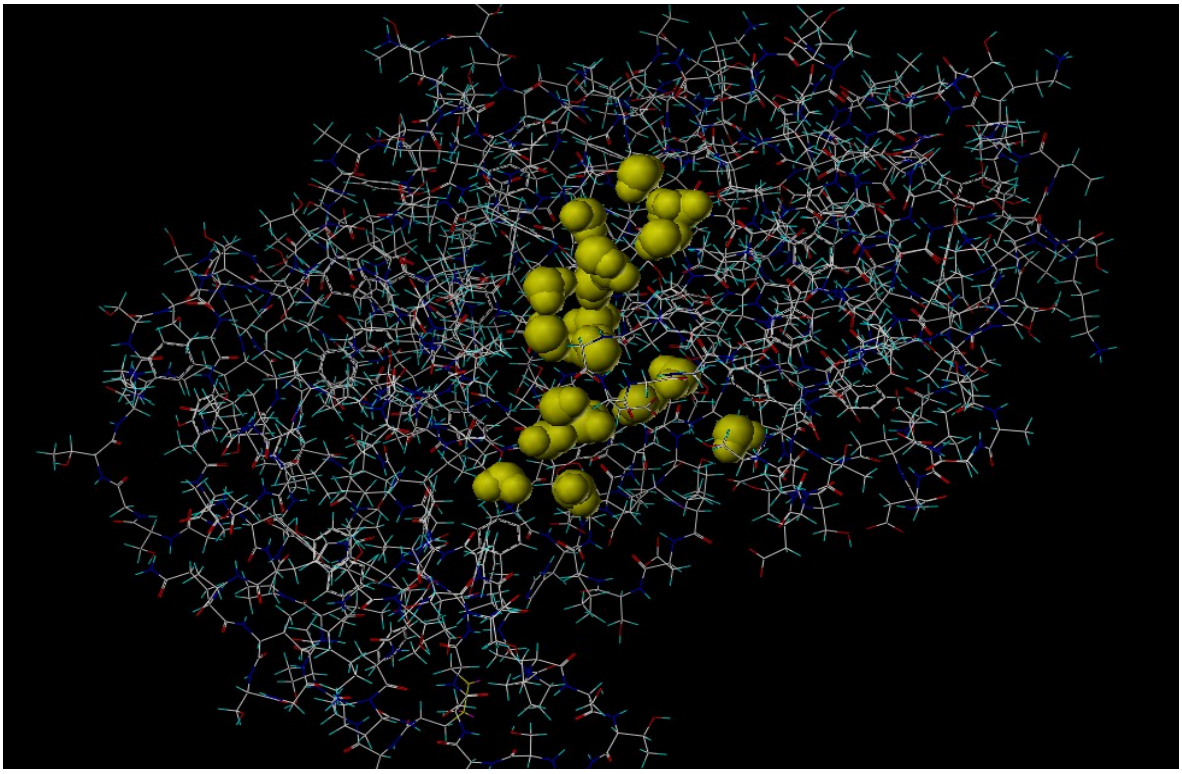

Fig 4.
Waters are chosen in the protein active site.

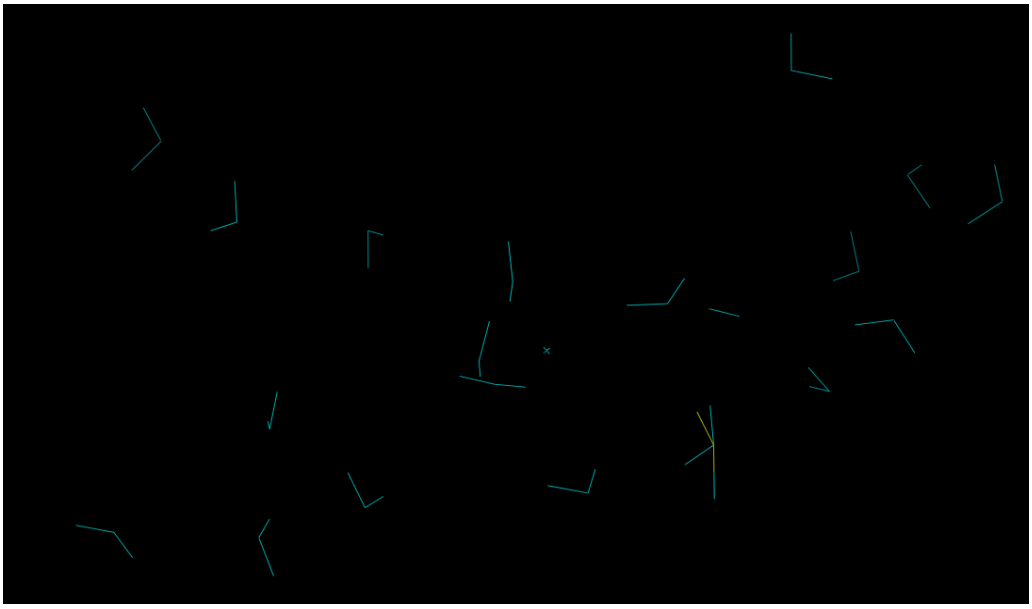

Fig. 5
The Center of Gravity is calculated from amongst the chosen water molecules . . .

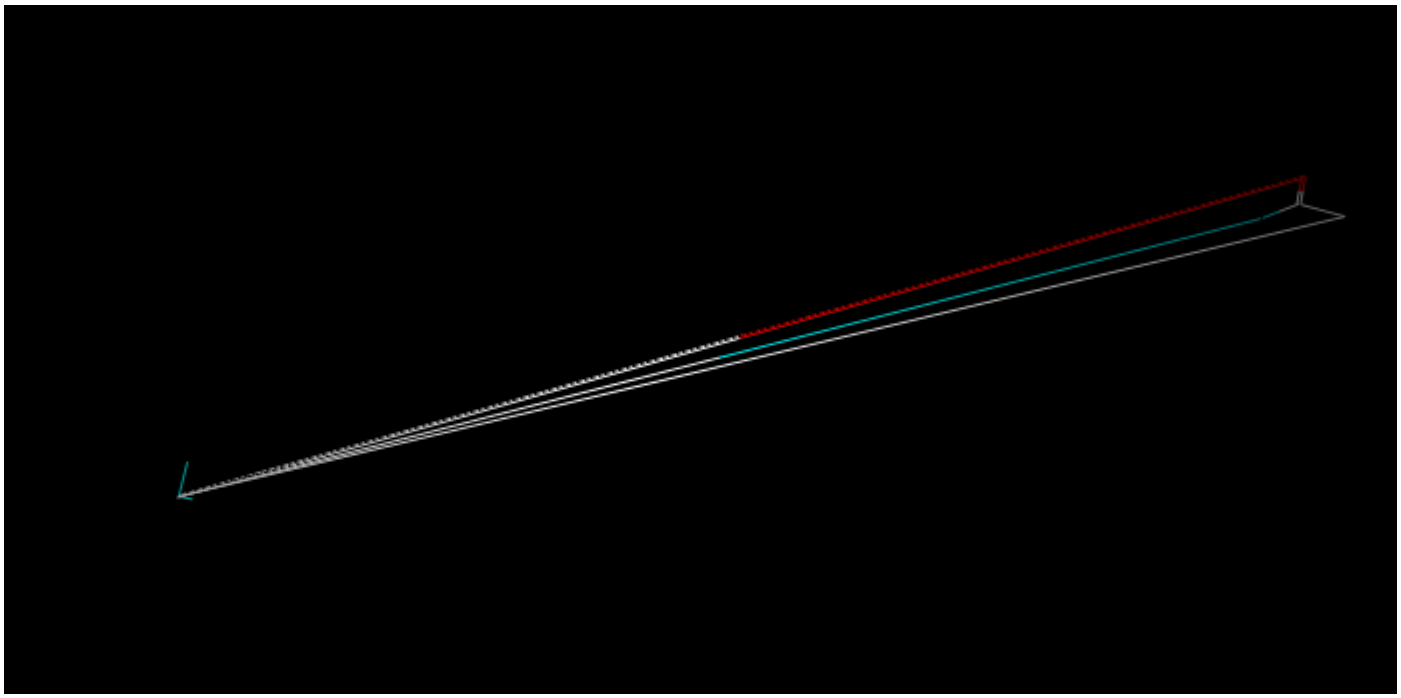

Fig. 6

The individual substituents are translated into the position of the oxygen on the target water molecule.

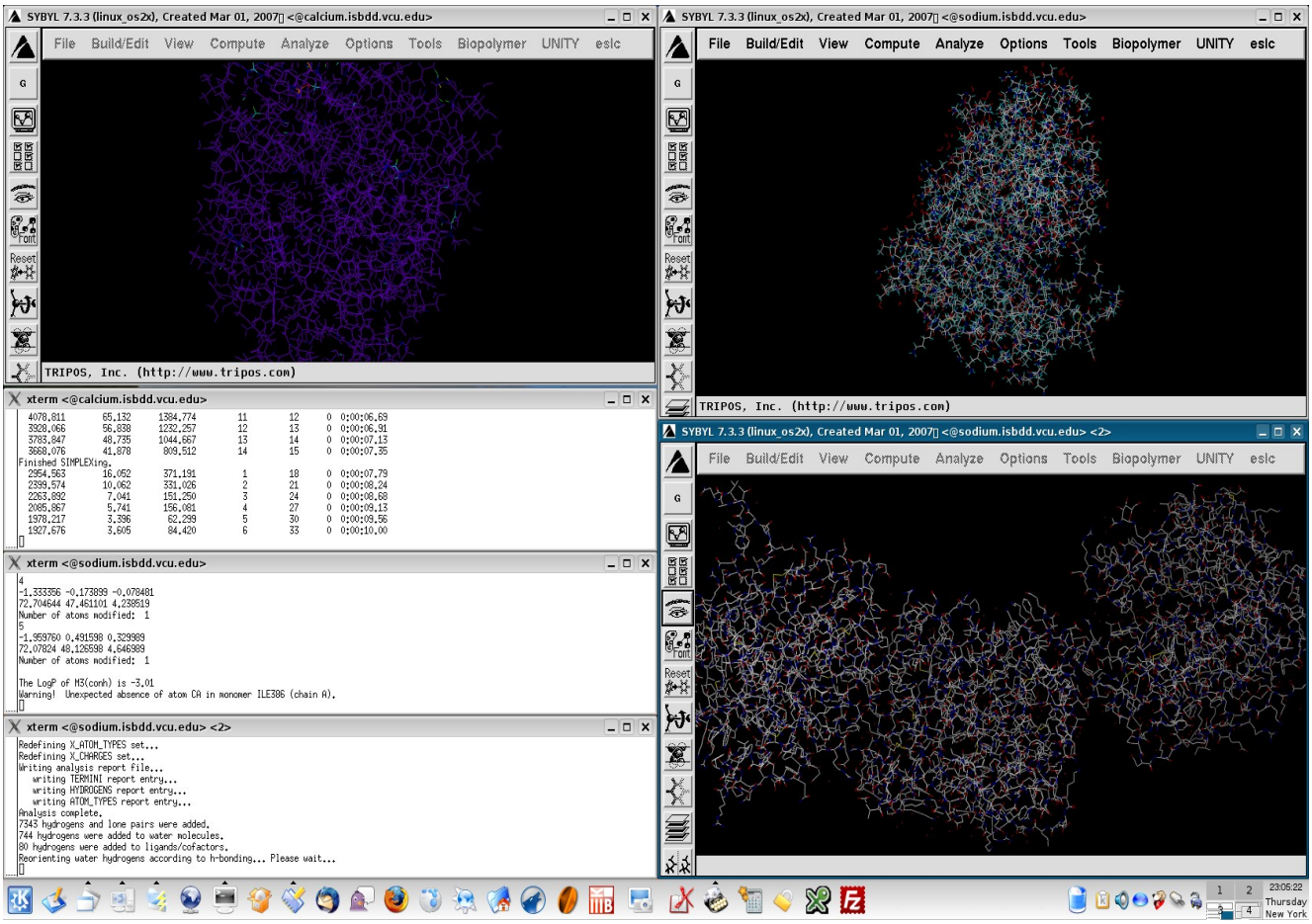

Fig. 7

The SPL Script running in three different session on three different proteins simultaneously.

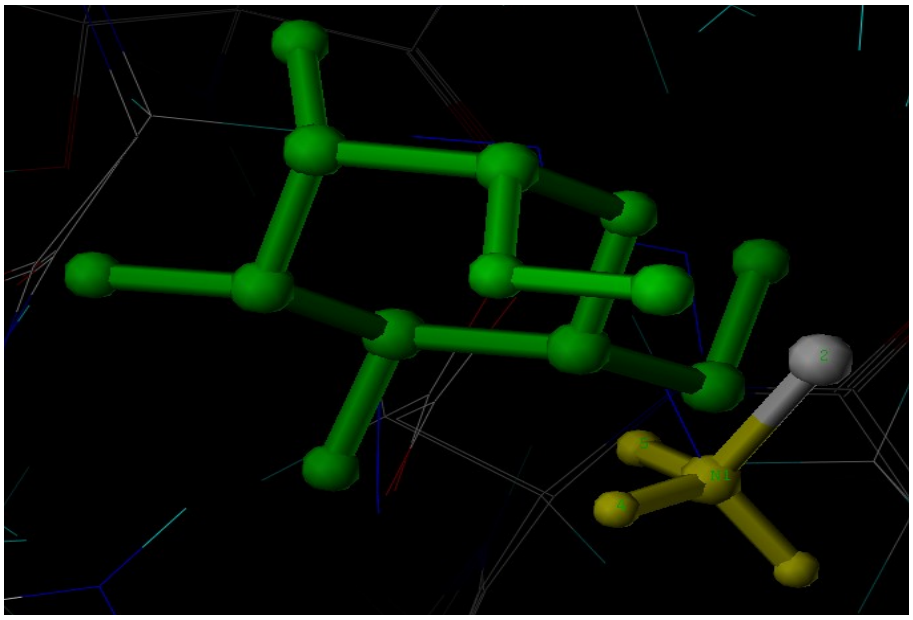

Fig. 8

A predicted NH3 substituent position, with dummy carbon aligned relative to the real ligand molecule (green). (2CTV)

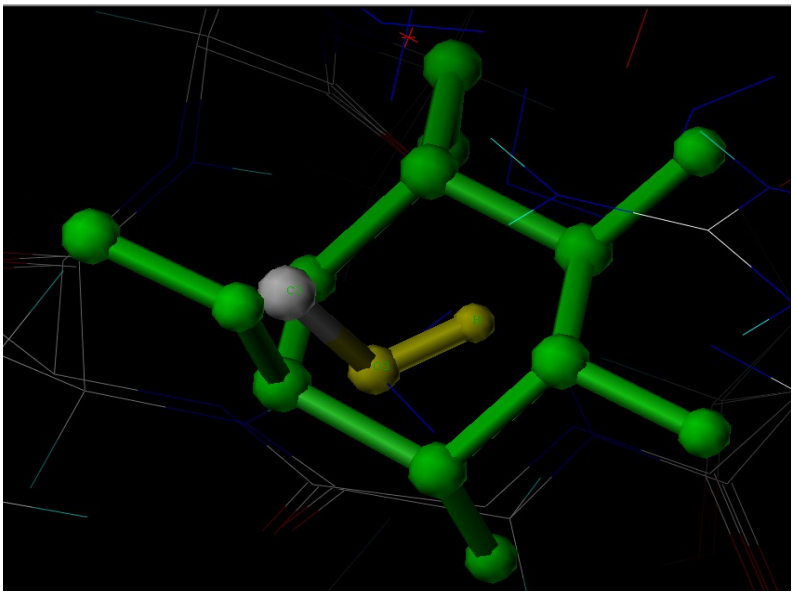

Fig. 9

Predicted Alcohol group, with dummy carbon aligned parallel to ligand arm. The water it replaces can be clearly seen.

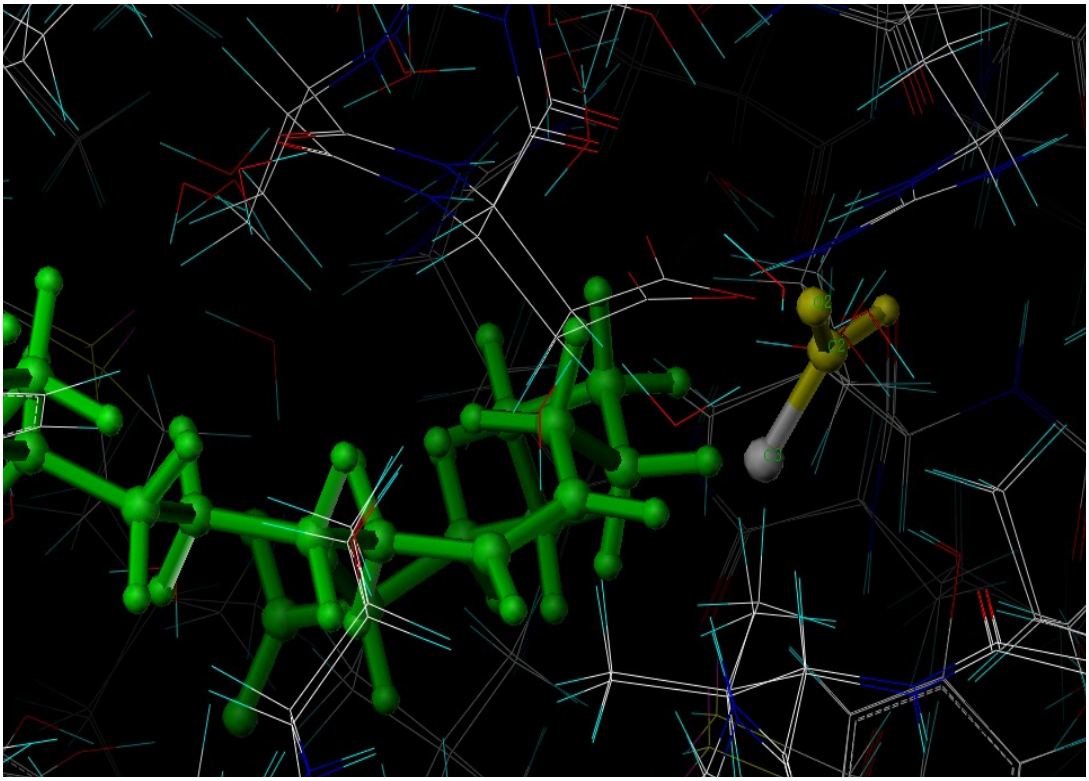

Fig. 10

A predicted aldehyde group on 1LID. This particular result shows how an arm might be added to the real ligand to make its interaction stronger with the protein by incorporating the aldehyde.

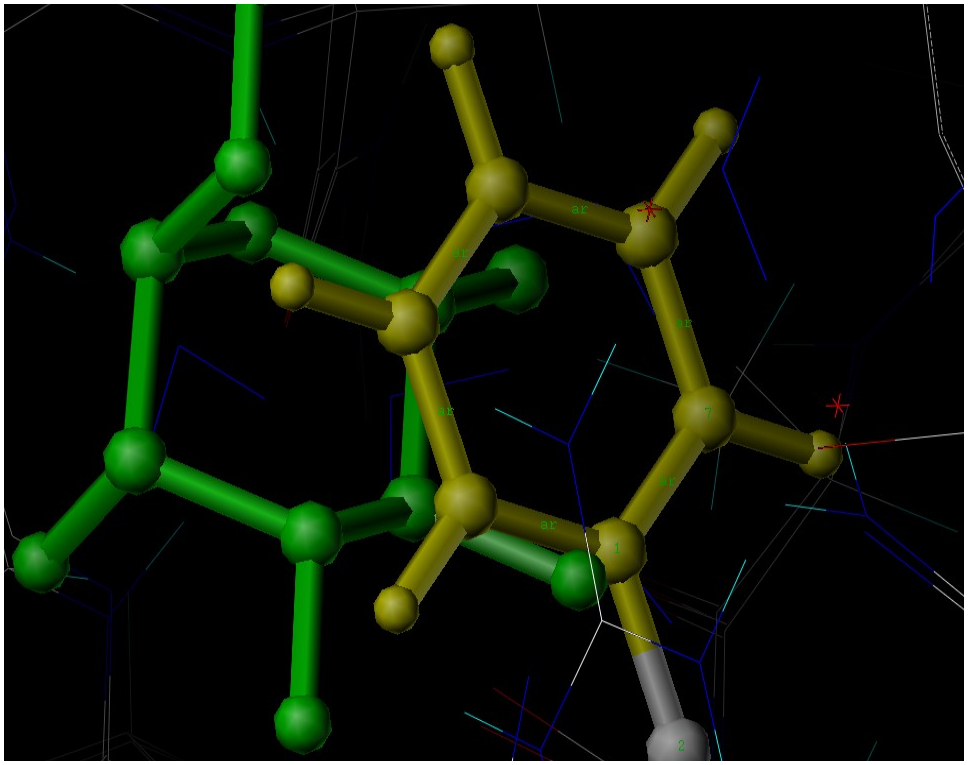

Fig. 11

An aromatic ring very closely aligns itself with the carbon (non-aromatic ring) for the ligand in 2CTV. This was not the most energetically favorable result, but if translation were allowed for the aromatic molecule it might be.