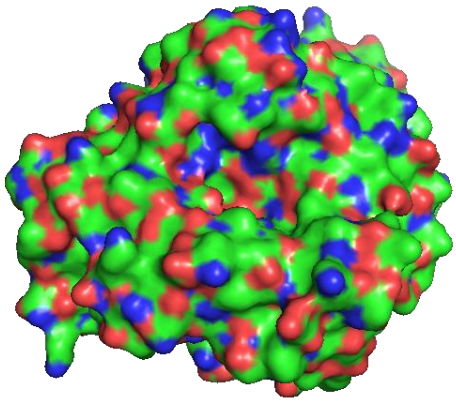
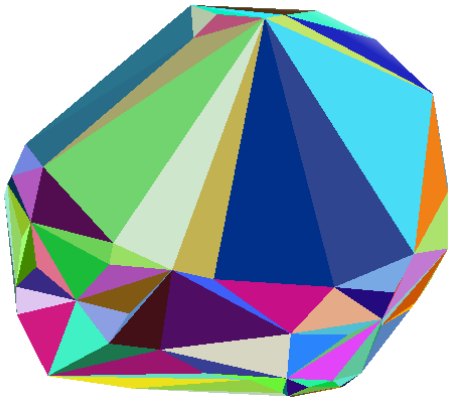




3DPocket - Computational Prediction of Ligand Binding Sites in Proteins

By Advait Maybhate

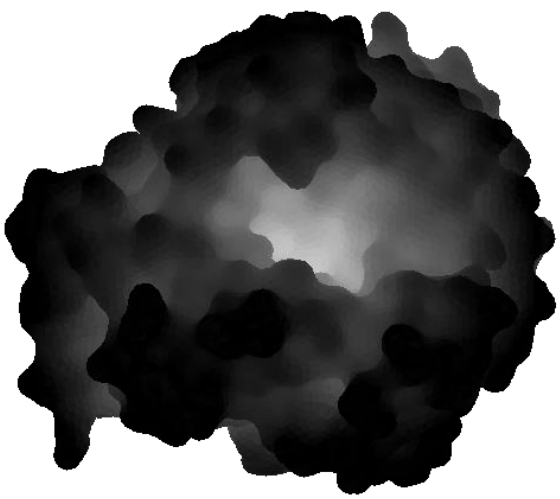
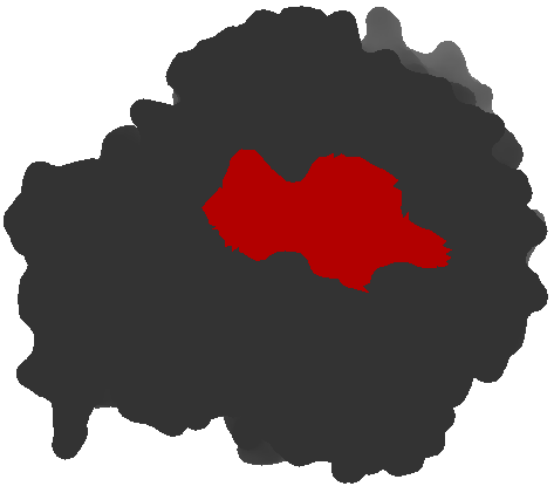
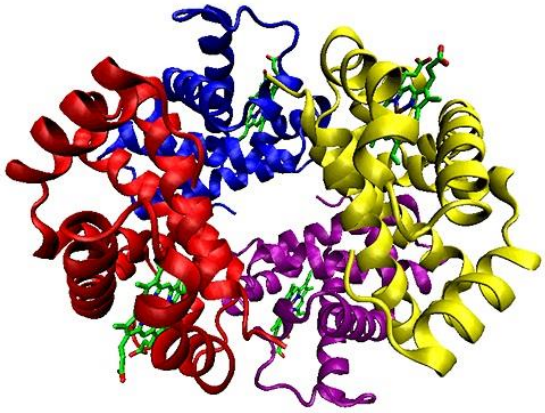



Table of Contents

Background Research

Proteins

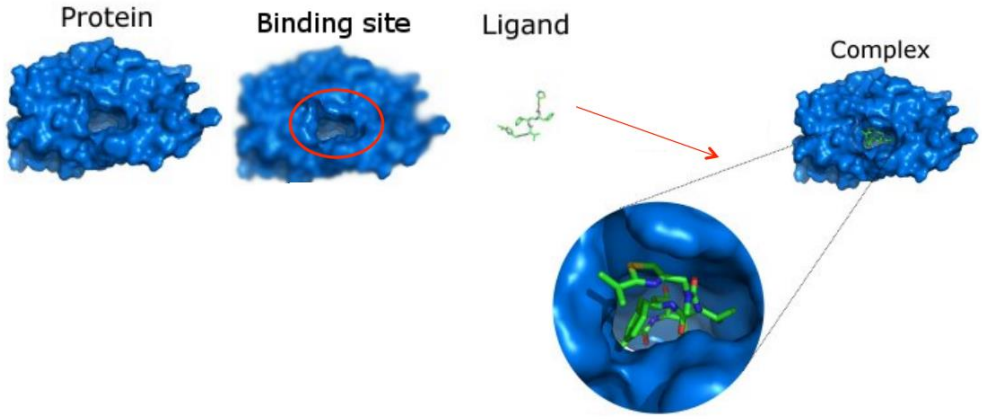
Proteins are nitrogenous organic compounds used throughout the body. They play an important role in many critical functions of living organisms. Specifically, in the human body, they carry out various essential functions from structural cell support to enzymatic activity. Proteins are made up of amino acids which are organic compounds containing amine and carboxyl functional groups. They are formed when DNA goes through the process of transcription to form a messenger ribonucleic acid (mRNA) and this mRNA is then translated to synthesize proteins.



Ribbon diagram of the structure of hemoglobin (a protein).

Ligands

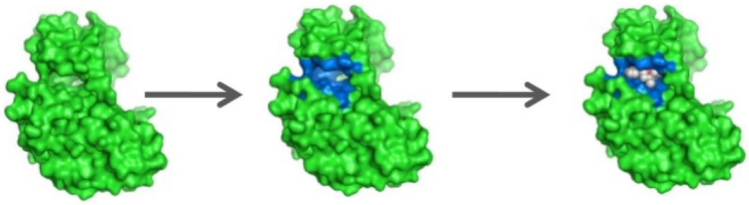
Ligands are substances that form complexes with biological molecules to serve a specific purpose. For example, heme, a ligand that contains an iron ion, binds to hemoglobin within red blood cells so they can successfully carry oxygen around the body. These ligands bind to protein at specific sites, allowing a biological complex to be created. In some cases, a slight mutation to a protein may cause its binding sites to be incorrectly configured, potentially resulting in important ligands not being able to bind to the protein.



Potential Factors Used to Identify Protein Binding Sites

Geometric Factors

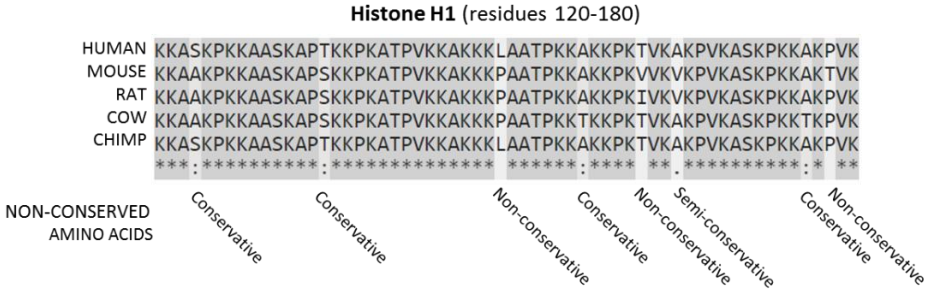
Concavities, such as pockets/cavities, most likely indicate a protein binding site. It is unlikely for a ligand to bind to the very exterior of a protein. On the other hand, ligands prefer to be surrounded by the protein, allowing a large surface of potential interaction between the protein and ligand. This also ensures that the complex formed by the protein and the ligand is relatively stable as opposed to the ligand breaking off from the complex.



An example of a pocket within a protein that can house a ligand.

Conservation Information

Usually, residues that bind ligands are conserved. This means that the same sequence (amino acids for a residue) can be found in multiple species of the protein that are distantly related (orthologous sequences) or within a genome (paralogous sequences). Conservation information of the protein sequence can be extracted from multiple sequence alignments (MSAs). Then, it can be used to improve the prediction of structure-based approaches or used on its own to predict protein binding sites.

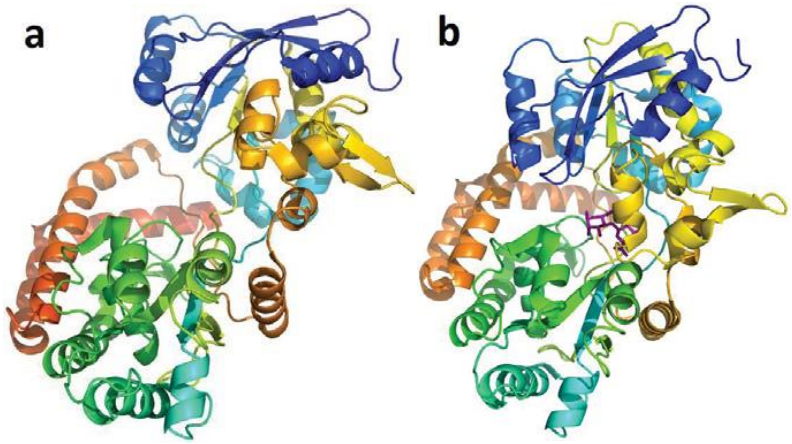


A multiple sequence alignment of five mammalian histone H1 proteins is shown above. Residues that are conserved across all sequences are highlighted in grey. Below each position of the protein sequence alignment is a key:

- conserved sites (*) – same amino acid
- sites with conservative replacements (:) – very similar biochemical properties
- sites with semi-conservative replacements (.) – relatively similar
- sites with non-conservative replacements ( ) – radically different

Holo-structures versus Apo-structures

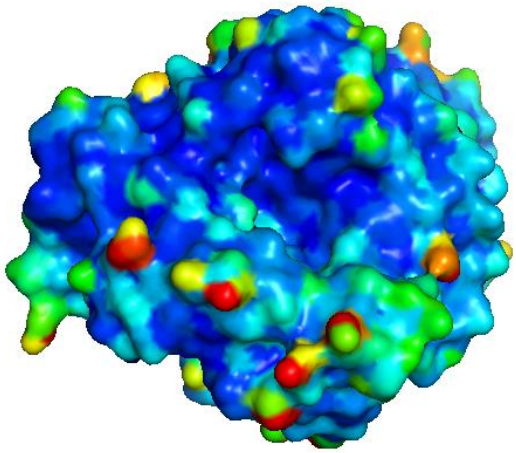
Having the holo-structure of a protein is usually preferred in drug design, when predicting protein binding sites and finding their shape/size. The holo-structure of a protein (conjugated protein) is the apo-protein combined with its prosthetic group (a ligand). Many enzymes may require cofactors or coenzymes in order to function fully. Such cofactors or coenzymes would be present within the holo-structure of the protein. Due to the induced fit model of enzymes, the enzyme's inherent structure itself may also change when it is bound to a ligand. Thus, it is more beneficial to analyze the holo-structure of a protein as opposed to its apo-structure. However, the holo-structure may not always be available.



A protein's (a) apo-structure (PDB: 1Y3Q) and (b) holo-structure (PDB: 1Y3N). It can be seen from the images that, in the holo-structure, the binding ligand (in purple stick representation) caused the protein to collapse more upon itself. After such a collapse, an existing cavity may be more "well defined" than it was in the apo-structure, which implies that detecting the cavity from the holo-structure would be easier than from the apo-structure.

B-Factor

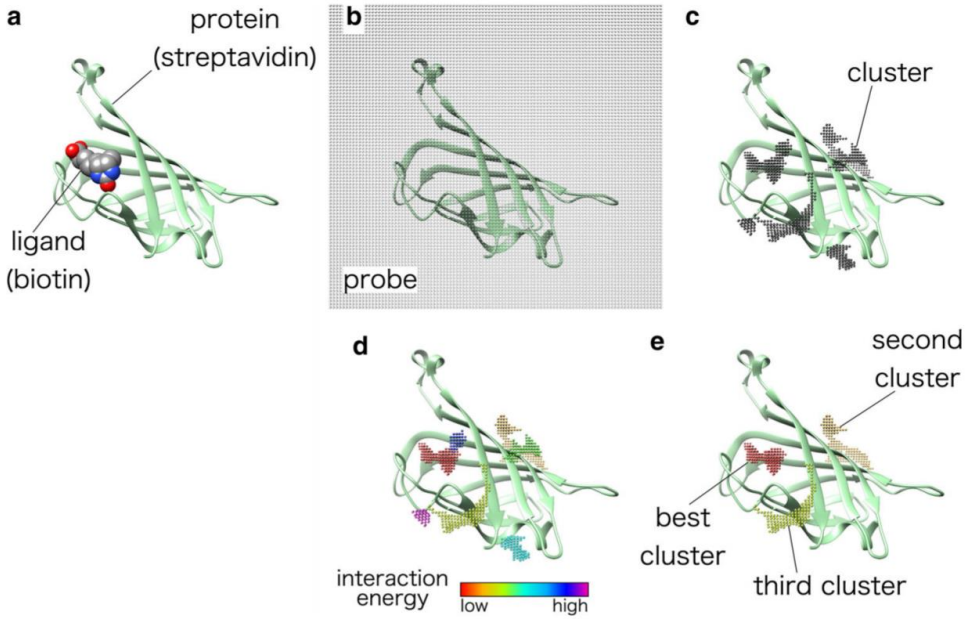
The B-factor or the temperature factor is considered as an indication of the fluctuation of an atom in a protein. Atoms with low B-factors belong to the well-ordered part of the structure, whereas a large B-factor suggests high mobility of an atom. For interacting proteins, the interface residues are less flexible than the rest of the protein surface and, therefore, are often associated with small B-factors. Deeply buried atoms in the core of the protein are usually rigid with a low B factor, and interfacial residues in protein binding complexes also have lower B-factors in comparison to the rest of the tertiary structural surface. The B-factor for each atom of a protein is available as part of the PDB (Protein Data Bank) file format.



Salmonella typhimurium LT2 neuraminidase or STNA (2SIL) coloured by B-factor indicating vibrational movement (blue-red corresponds to low-high vibrational movement).

Van der Waals Interaction Energies

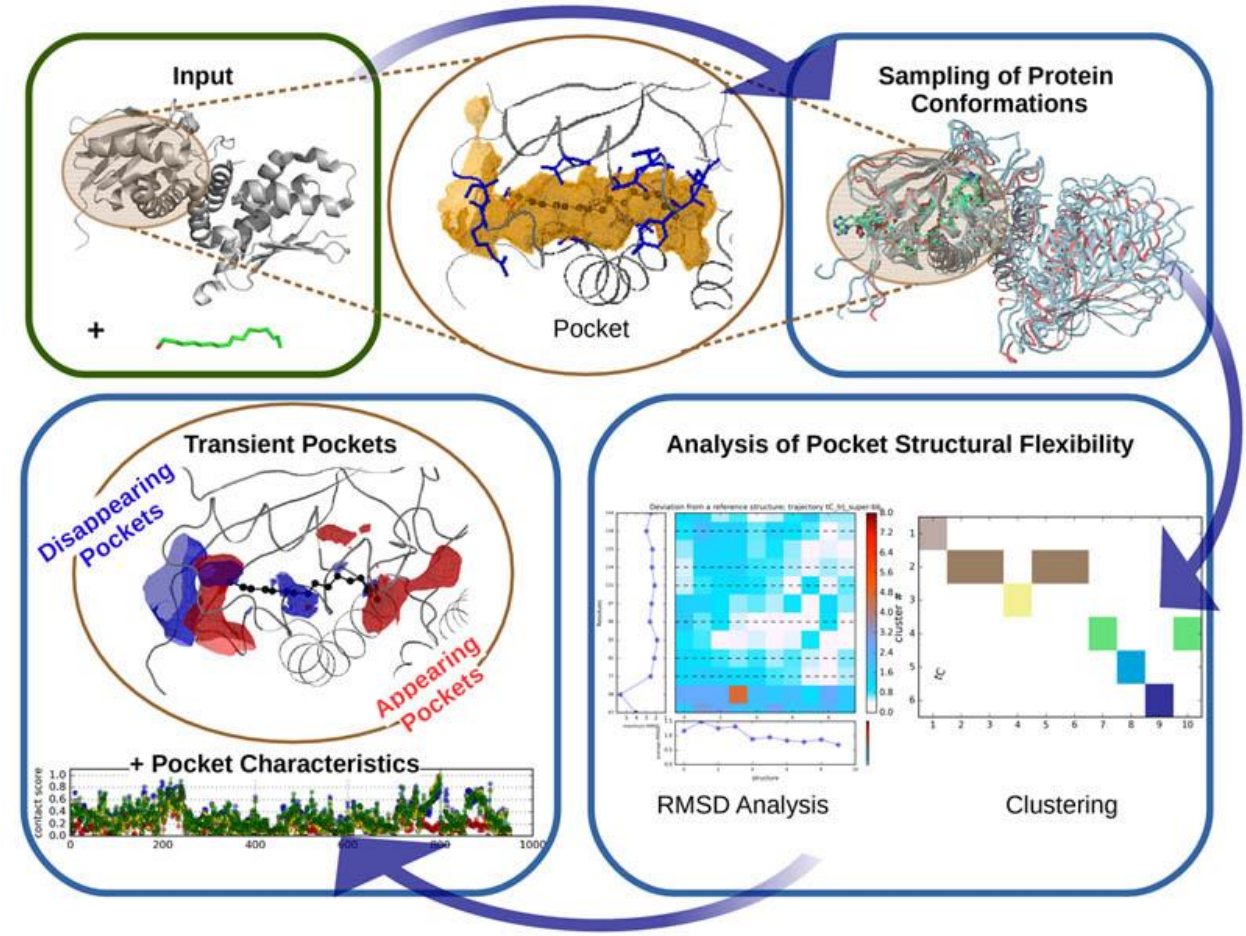
The interaction energy between a protein and a van der Waals probe can be used to locate energetically favourable binding sites. For example, in Q-SiteFinder, energetically favourable probe sites are clustered according to their spatial proximity and clusters are then ranked according to the sum of interaction energies for sites within each cluster. Specifically, it uses a methyl probe (−CH3) to calculate these interaction energies. Ligands have been found to bind to sites where the interaction energy within the protein is minimal, representing a stable binding site.



An example showing an energy-based approach to predicting protein binding sites.

Protein Dynamics

Proteins are flexible structures that can have more than one possible conformation, and this is why one static structure is not always enough to predict the binding sites. It is possible to produce ensembles of conformations using well-known molecular dynamics tools (e.g. GROMACS—GROningen MAchine for Chemical Simulations). Having such an ensemble of conformations, a binding site prediction program can then be used to predict “transient” and “conservative” pockets e.g. TRAPP (TRAnsient Pockets in Proteins).



TRAPP workflow (note: RMSD analysis stands for root-mean-square deviation analysis of atomic positions which is the average distance between atoms of superimposed proteins).

Charge of Residue

The charge of a specific residue may affect its binding affinity, depending on the charge of the ligand it is binding to. For example, positively charged residues, such as arginine, are more likely to interact with the negatively charged backbone of DNA.

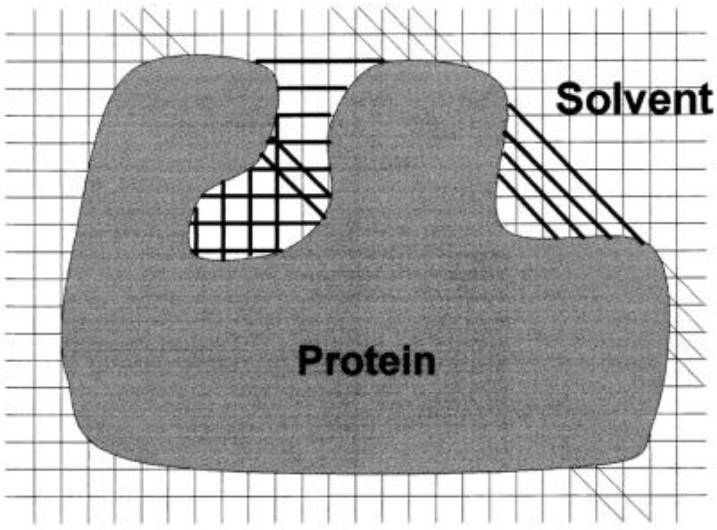
Purpose

All biological functions are based on protein interactions. Predicting the binding sites of proteins, where these interactions take place, is a crucial task. In modern times, computers are increasingly being used to reduce costs and time spent finding such binding sites. The purpose of this project is to create a novel geometry-based algorithm that tackles the challenge of predicting ligand binding sites within proteins.

Previous Algorithms

LIGSITE

In LIGSITE, pockets are identified by scanning along the x, y, and z axes and the cubic diagonals (only four diagonals are shown) for areas that are enclosed on both sides by protein (indicated by boldface lines):

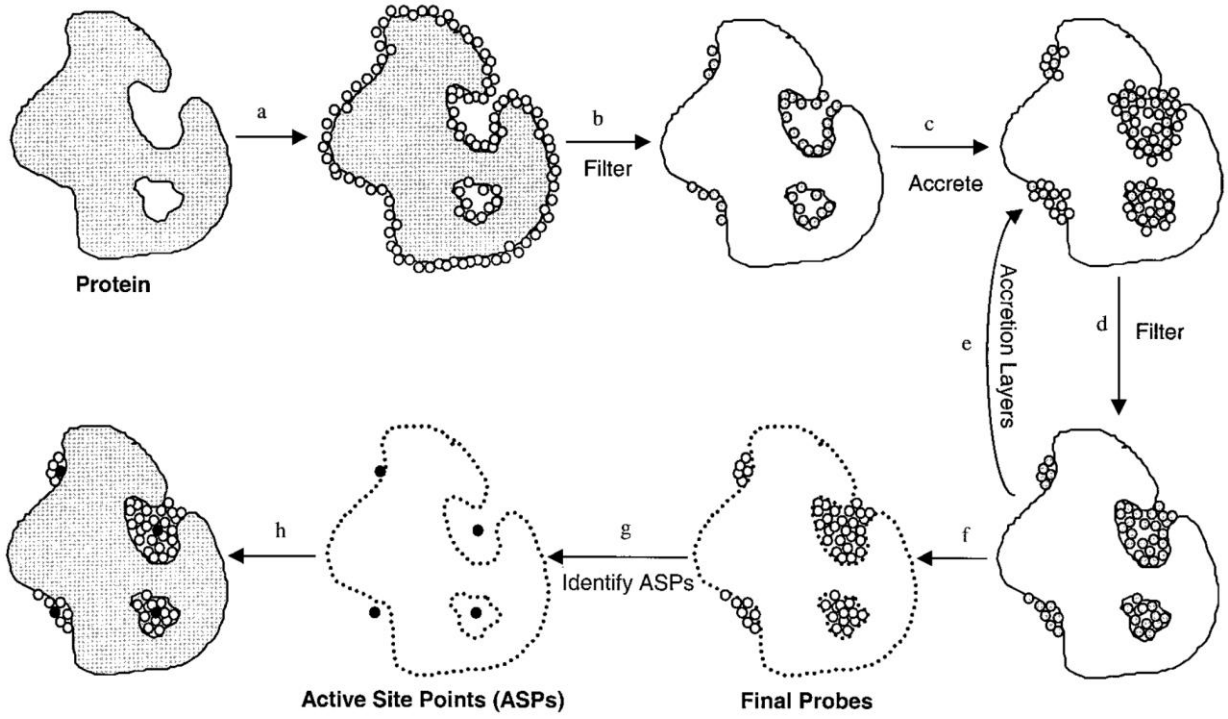


Essentially, it identifies solvent-protein-solvent events to predict pockets and cavities that may be binding sites. The accuracy with which the surface of pockets and cavities is determined depends on the step size that is used. At an infinitely small step size, the surface determined by this algorithm would be equivalent to the contact surface. A large grid step size, e.g. 2.0 Å, decreases calculation time but results in a very angular surface.

PASS

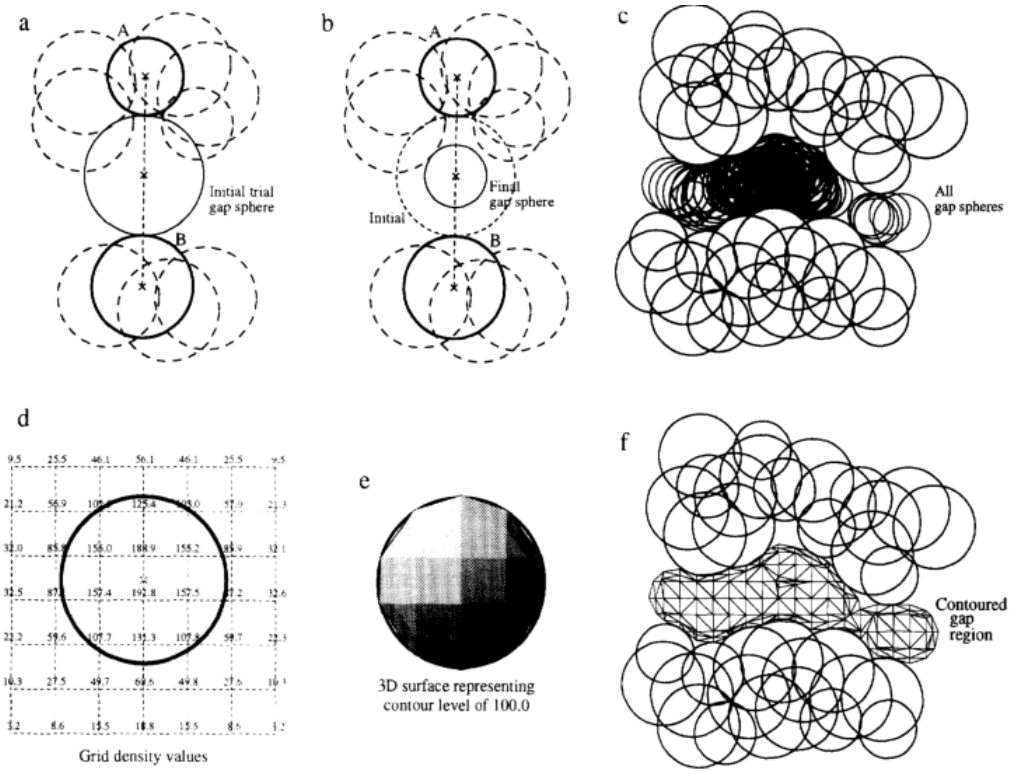
PASS (Putative Active Sites with Spheres) uses the concept of probe spheres and "burial scores" to identify binding sites:

a. Coat the protein with an initial layer of spherical probes
b. These probes are filtered to eliminate those that:
   1) Clash with the protein
   2) Are not sufficiently buried
   3) Lie within 1 Å of a more buried probe.
c. A new layer of spheres is accreted on top of the previous layer.
d. These probes are filtered as described in step b.
e. A new layer of spheres is accreted onto the existing probes, as in step c.
f. Accretion and filtering (steps e and d) are repeated until a layer occurs in which no newly found probes survive the filters. This leaves the final set of probe spheres.
g. Probe weights (PW) are computed for each sphere and active site points (ASPs) are then identified from amongst the final probes.
h. The final PASS prediction is produced. The final probe spheres are first smoothed, leaving only clusters of four or more.



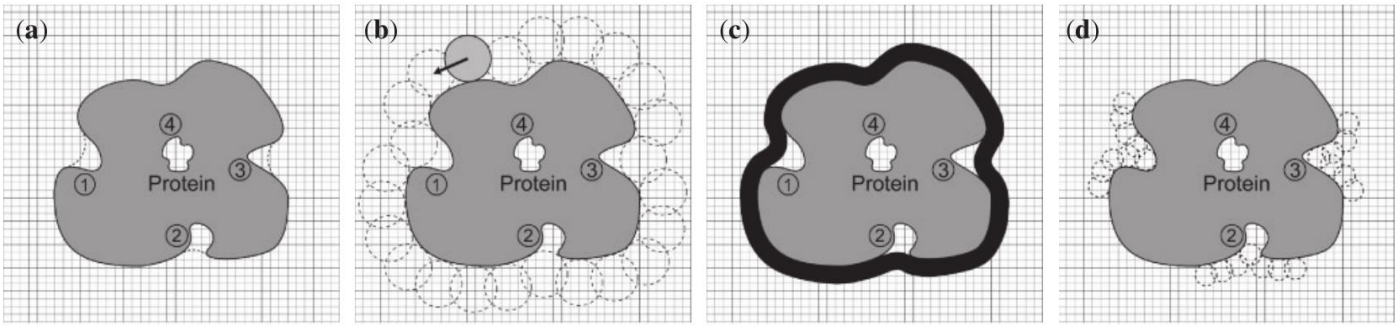
SURFNET

SURFNET uses “gap spheres” to fill regions between two molecules in order to identify pockets/cavities. The gap regions are filled with gap spheres that are defined by the surface of two specific atoms. Then, the gap spheres are reduced in size until they no longer intersect the volumes of other atoms in the area. This process is repeated for all combinations of two atoms within the gap region. Then, a polyhedron approximating the sphere can be used to generate a larger polyhedron representing the gap region as a whole:



Roll

The key concept of Roll is to generate a crust-like surface called the probe surface enveloping protein. This allows Roll to identify the region between the probe surface and protein surface as a ‘pocket’ and the region surrounded by protein surface as a ‘cavity’. Such a probe surface is created by “rolling” a probe sphere around the surface of the protein. The size of the probe sphere can determine the size of the binding sites to be identified (a smaller probe sphere may cause some binding sites not to be identified).



a) 2D Slice of the grid system. The dashed and solid lines show the probe surface and the protein surface, respectively. The grey region is the protein. Regions 1–3 are defined as pockets and region 4 is defined as a cavity.
b) The rolling process. The light grey ball indicates the starting position and the dashed balls show the trace of rolling.
c) The black area is the probe surface. Pockets 1-3 have been identified (between probe surface and protein surface) in addition to cavity 4 (surrounded by protein surface).
d) The smaller probe sphere causes pocket 1 to disappear and pockets 2–3 to become smaller, while it does not affect cavity 4.

Design Criteria

- Accuracy—Accurately predicts experimentally-known binding sites within enzymes and DNA-binding proteins.
- Speed—Aim is to create an algorithm which is faster than existing algorithms e.g. LIGSITE and SURFNET.

3DPocket Algorithm

The steps within the 3DPocket algorithm are as follows:

1. Obtain the protein's structural file (Crystallographic Information File or CIF) from the RCSB (Research Collaboratory for Structural Bioinformatics) PDB (Protein Data Bank) or elsewhere.
2. Use PyMol, a software used to view/manipulate protein structures, to compute the Connolly surface of the protein.
3. Export the Connolly surface computed by PyMol to a WRL file written in the Virtual Reality Modeling Language (VRML). Parse the coordinates of all the triangular faces that create the Connolly surface from the WRL file to Python objects (arrays).
4. Use the Quickhull algorithm to compute the 3D convex hull surrounding the protein. Store the coordinates of the triangular faces making up this convex hull.
5. Calculate the centroid within each triangular face of the Connolly surface.
6. Find the coefficients of the scalar equations representing the planes created by each triangular face within the convex hull (3 points define a unique plane).
7. Calculate the distance from each centroid to every plane created by the triangular faces within the convex hull (point-to-plane distance).
8. Identify the minimum distance from each centroid to the convex hull. Assign these distances as "scores" to each triangular face within the Connolly surface.
9. Normalize the set of scores from 0-1 (1 being the largest distance).
10. Use the normalized scores to colourize the protein (white representing likely binding sites and black representing unlikely binding sites). This representation of the protein can show a gradient-based view of the binding sites.
11. Apply a threshold to distinguish a binding site and a non-binding site. A threshold of scores above 0.4 (when normalized) was primarily used in 3DPocket.

Once the predicted binding sites have been computed, it is possible to compare these binding sites to the actual binding sites using the following procedure:

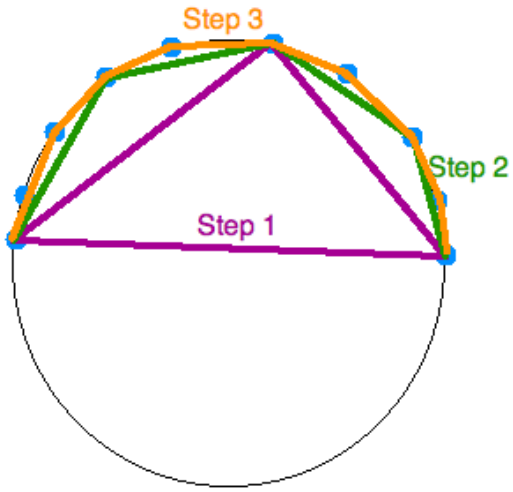
1. Obtain the actual, experimentally-determined binding sites from the LIGASITE database. They will be provided as a set of binding residues.
2. Convert the set of binding triangular faces identified by applying the threshold to a set of predicted binding residues:
    a. For every binding triangular face within the Connolly surface representation of the protein, identify all atoms within 1.7 Å.
    b. Mark these set of atoms as "binding atoms". Any residue that contains any of these atoms will be considered a "binding residue".

3. Create a confusion matrix by comparing the predicted binding residues to the actual binding residues.
4. Calculate the accuracy and Matthews correlation coefficient (MCC) values using the confusion matrix (allows for the comparison of 3DPocket to previously created algorithms).

Computing the Convex Hull (Quickhull)

In 2 dimensions, the Quickhull algorithm involves the creation a triangles bounding the interior points recursively. It can be broken down to the following steps:

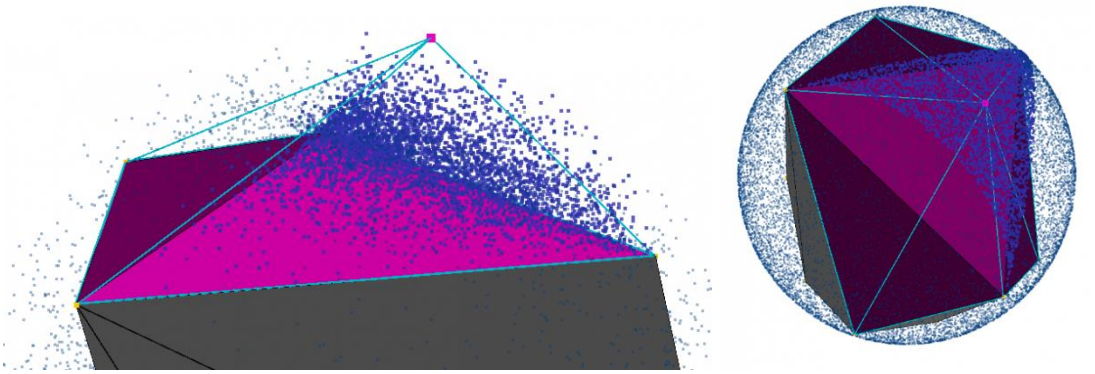
1. Find the points with minimum and maximum x coordinates —these will always be part of the convex hull.
2. Use the line formed by the two points to divide the set in two subsets of points. These subsets will be processed recursively.
3. Determine the point, on one side of the line, that is the maximum distance away from the line. The two points found initially, in addition to this one, form a triangle.
4. The points contained within the triangle cannot be part of the convex hull and can therefore be ignored in the next steps.
5. Repeat the previous two steps on the two lines formed by the triangle (not the initial line).
6. Keep on doing so on until no more points are left. Then, the recursion has come to an end and the points selected constitute the convex hull.



When extending Quickhull to 3 dimensions, tetrahedrons can be used, analogous to triangles in 2 dimensions:

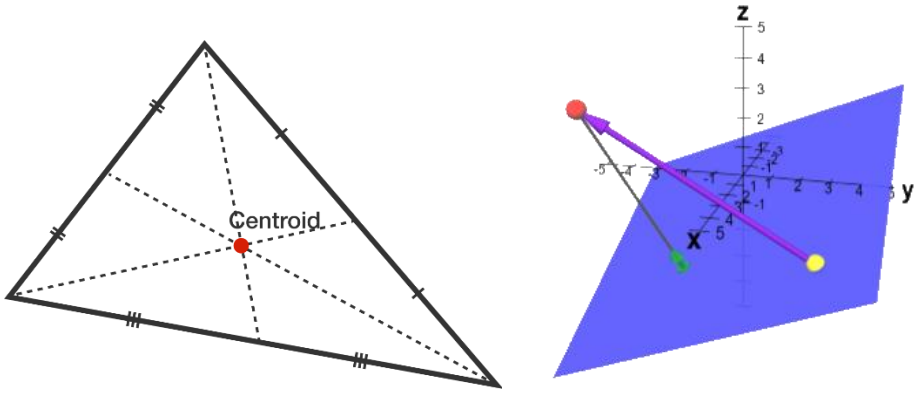
1. Initially, the first tetrahedron can be created by choosing 4 points from the 6 extreme points (extreme positive/negative points in the three dimensions—x, y, and z).

2. Then, the algorithm can be applied recursively to each of the 4 faces within the tetrahedron (the furthest point from each of these faces can be found to create another tetrahedron).
3. This process is continued until no more points are left. Any points contained within the tetrahedrons generated are not part of the convex hull, whereas points selected as vertices of tetrahedrons constitute the convex hull.



Calculating Minimum Distances

To determine the distance from the Connolly surface to the convex hull, the Connolly surface is first converted into a polyhedron with triangular faces, and the centroid of each face is determined. Then, a plane is constructed from each triangular face of the convex hull, and the point-to-plane distance is calculated between each centroid and every plane (an arbitrary vector connecting the point to the plane can be projected onto the plane's normal). Next, the minimum distance calculated for each centroid is used to assign a score for the triangular face.



Applying Confusion Matrices

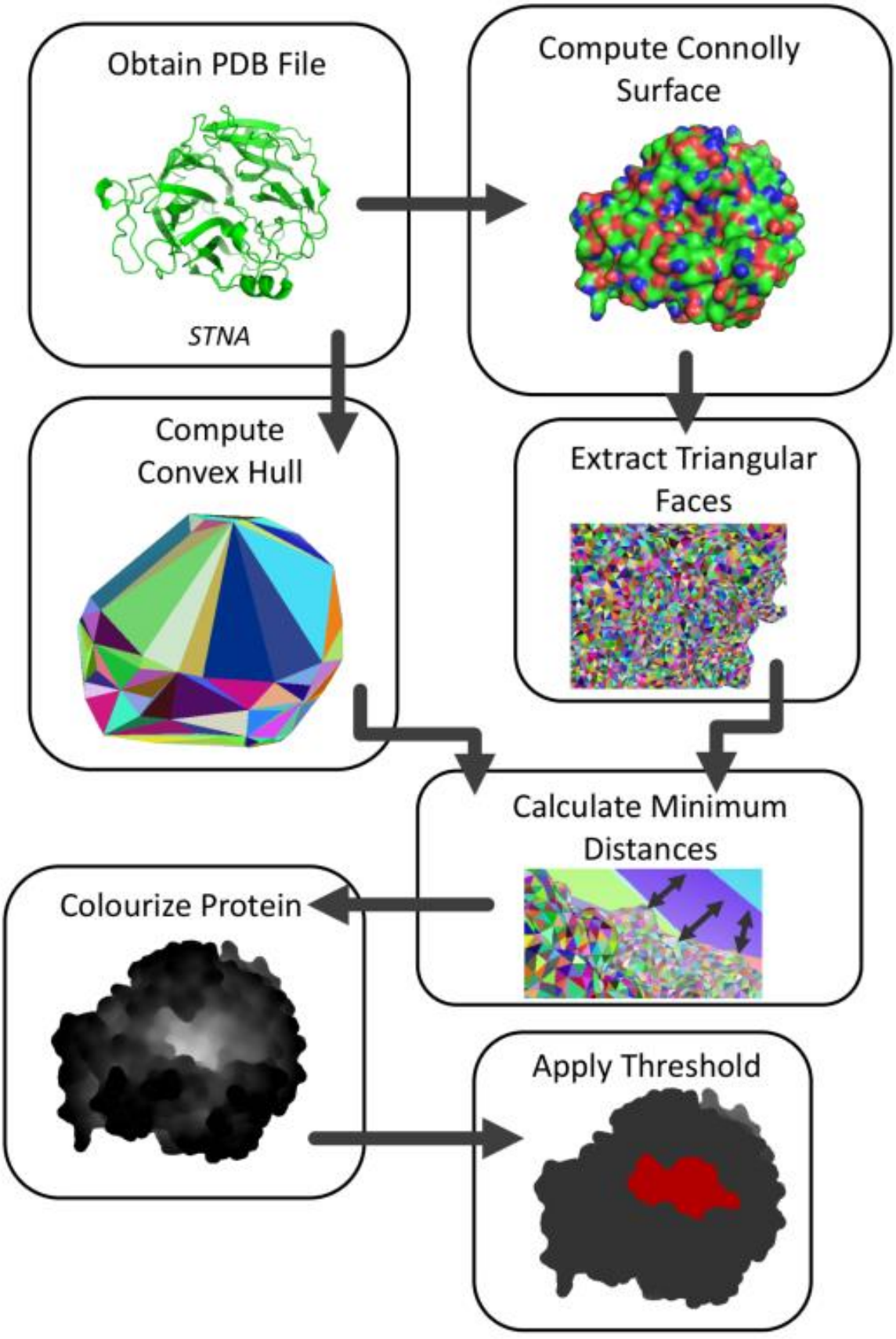
To determine the performance of 3DPocket, a confusion matrix is generated. Such a matrix compares predicted binding sites to those determined experimentally:

Predicted Value (predicted by the test)		Actual Value (as confirmed by experiment): positives	Actual Value (as confirmed by experiment): negatives
	positives	**TP** True Positive	**FP** False Positive
	negatives	**FN** False Negative	**TN** True Negative

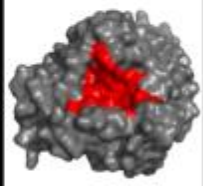
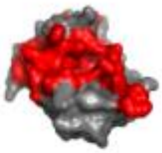
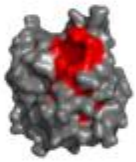
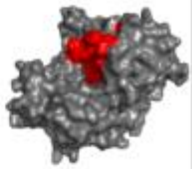
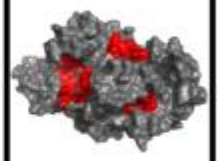
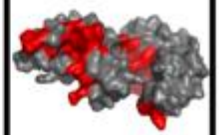
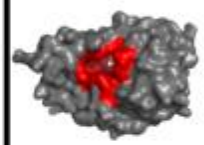
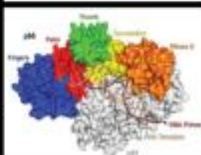
Then, the accuracy (ACC) and Matthews Correlation Coefficient (MCC) are used to gauge performance against previous algorithms:

$$ACC = \frac{TP+TN}{P+N} \qquad MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

3DPocket Flow Chart



Results

Protein Name	Protein Function	Ribbon Diagram	Raw Colourized Protein	Predicted Binding Sites	Actual Binding Sites	Accuracy	Matthews Correlation Coefficient
Salmonella typhimurium LT2 neuraminidase (STNA)	• Catalyzes breakdown of glycosides carrying neuraminic acid • Causes typhoid in mice • Causes Salmonella infections, commonly through food poisoning					76.56%	0.3255
Bothropstoxin-I (BthTX-I)	• Myotoxin • Isolated from *Bothrops jararacussu* snake venom • Causes severe muscle necrosis					71.13%	0.3006
Cellular Retinoic Acid Binding Protein Type II	• Cytoplasmic binding protein coded by the CRABP2 gene • Retinoic acid-mediated regulation of human skin growth					86.11%	0.4264
LipA from *Xanthomonas oryzae* (Xoo)	• Xoo causes a serious disease within rice • A secretory virulence factor responsible for eliciting innate immune responses • Binds to glycoside ligand					90.17%	0.3571
Delta1-piperideine-2-carboxylate reductase	• Catalyzes metabolism reaction within saprotrophic soil bacterium, *Pseudomonas syringae*					84.57%	0.3945
Schizosaccharomyces Pombe Alkyltransferase-like proteins (ATL)	• Recognizes DNA damage caused by methylating agents • Recruits the help of a UvrA protein to repair the damage					77.63%	0.3124
HIV-1 Protease	• Essential for the life-cycle of HIV, the retrovirus that causes AIDS • Cleaves newly synthesized polyproteins					81.03%	0.5229
HIV-1 Reverse Transcriptase	• Transcription of the RNA genome of the HIV virus into DNA					—	—

Matthews Correlation Coefficient (MCC) Comparison

Name	*3DPocket*	*SURFNET*	*DEPTH*	*LIGSITE*	*GHECOM*	*SiteHound*	*Fpocket*
MCC score	0.38	0.31	0.23	0.38	0.33	0.35	0.33

Source: "Comparison of structure-based tools for the prediction of ligand binding site residues in apo-structures" (Ezzat and Kwoh, 2012)

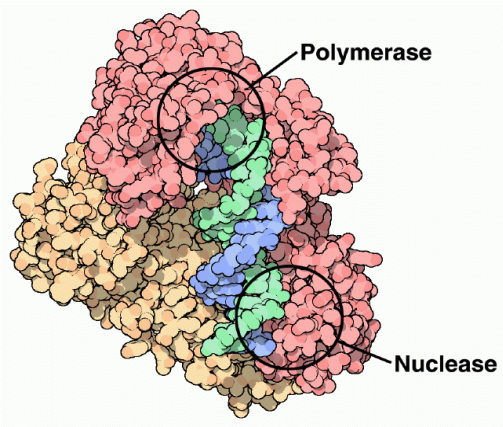
Thus, 3DPocket was able to tie LIGSITE for the top place in terms of MCC (a statistic that combines the rate of true/false positives and true/false negatives) among other algorithms. This shows that the 3DPocket algorithm is very accurate and robust. However, a larger sample of proteins must be tested in order to definitively prove that 3DPocket is reliable and accurate. Additionally, testing it on a larger variety of proteins including DNA-binding proteins and enzymes would be beneficial, in order to reinforce its reliability.

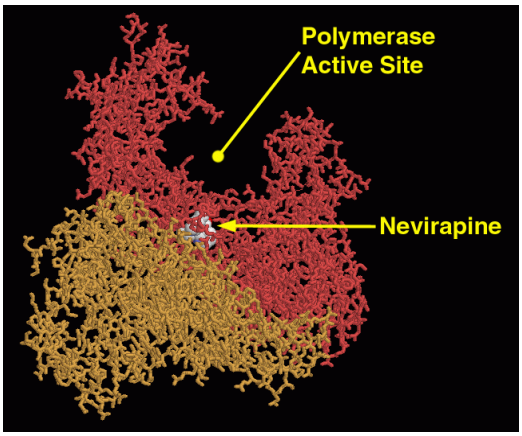
Conclusion

In conclusion, a new algorithm, 3DPocket, was created to predict protein-ligand binding sites. The algorithm is relatively fast (~10 minutes compared to 30+ minutes with PocketPicker). When tested on a variety of previously analyzed proteins, its results matched experimental findings. 3DPocket performs better than comparable algorithms (MCC = 0.38).

Applications

- Determining shape and size of binding sites
- Identifying mutations causing a change in binding site shape
- Aid in designing effective drugs to inhibit proteins e.g. designing Nevirapine to obstruct the polymerase active site within HIV-1 reverse transcriptase (integral in HIV)





- Identifying protein-ligand binding mechanisms and protein functions

Future Directions

- Employing the use of biochemical properties e.g. b-factor (describes the movement of atoms—low movement corresponds to likely binding sites deep within the protein), hydrophobicity (small molecules prefer hydrophobic protein pockets) and conservation information.
- Running X-ray crystallography in the lab in order to experimentally determine the structure of a protein and then run 3DPocket to predict its binding sites.
- Use nuclear magnetic resonance (NMR) spectroscopy in order to experimentally determine binding sites of proteins and compare these results to the predicted binding sites by 3DPocket.
- Test algorithm on a wider variety of proteins (both DNA-binding proteins and enzymes) from the Protein Data Bank (PDB) in order to ensure reliability.
- Optimize the number of calculations done by 3DPocket by employing heuristics to reduce the number of combinations that need to be tested when calculating minimum distances from the Connolly surface to the convex hull.

Acknowledgements

I would like to thank Dr. Andrew Doxey, at the University of Waterloo, for being my mentor and helping me understand biological terms and concepts, in addition to guiding me in finding the PDB files of various proteins.

Bibliography

Cintra, A. C. O., et al. “Bothropstoxin-I: Amino Acid Sequence and Function.” *Journal of Protein Chemistry*, vol. 12, no. 1, 1993, pp. 57–64., doi:10.1007/bf01024915.

Connolly, M. “Solvent-Accessible Surfaces of Proteins and Nucleic Acids.” *Science*, vol. 221, no. 4612, 1983, pp. 709–713., doi:10.1126/science.6879170.

Ezzat, Ali, and Chee Keong Kwoh. “Comparison of Structure-Based Tools for the Prediction of Ligand Binding Site Residues in Apo-Structures.” *Procedia Computer Science*, vol. 11, 2012, pp. 115–126., doi:10.1016/j.procs.2012.09.013.

Fathi, Seyed Saberi, and Jack A Tuszynski. “A Simple Method for Finding a Protein’s Ligand-Binding Pockets.” *BMC Structural Biology*, vol. 14, no. 1, 2014, p. 18., doi:10.1186/1472-6807-14-18.

Guilloux, Vincent Le, et al. “Fpocket: An Open Source Platform for Ligand Pocket Detection.” *BMC Bioinformatics*, vol. 10, no. 1, 2009, p. 168., doi:10.1186/1471-2105-10-168.

Hendlich, Manfred, et al. “LIGSITE: Automatic and Efficient Detection of Potential Small Molecule-Binding Sites in Proteins.” *Journal of Molecular Graphics and Modelling*, vol. 15, no. 6, 1997, pp. 359–363., doi:10.1016/s1093-3263(98)00002-3.

H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne. (2000) The Protein Data Bank Nucleic Acids Research, 28: 235-242.

Huang, Bingding, and Michael Schroeder. “LIGSITEcsc: Predicting Ligand Binding Sites Using the Connolly Surface and Degree of Conservation.” *BMC Structural Biology*, vol. 6, no. 1, 2006, p. 19., doi:10.1186/1472-6807-6-19.

Jr., G. Patrick Brady, and Pieter F.w. Stouten. “Fast Prediction and Visualization of Protein Binding Pockets with PASS.” *Journal of Computer-Aided Molecular Design*, vol. 14, no. 4, May 2000, pp. 383–401., doi:10.1023/a:1008124202956.

Kalidas, Yeturu, and Nagasuma Chandra. “PocketDepth: A New Depth Based Algorithm for Identification of Ligand Binding Sites in Proteins.” *Journal of Structural Biology*, vol. 161, no. 1, 2008, pp. 31–42., doi:10.1016/j.jsb.2007.09.005.

Krivák, Radoslav, and David Hoksza. “P2RANK: Knowledge-Based Ligand Binding Site Prediction Using Aggregated Local Features.” *Algorithms for Computational Biology Lecture Notes in Computer Science*, 2015, pp. 41–52., doi:10.1007/978-3-319-21233-3_4.

Krivák, Radoslav, and David Hoksza. “Improving Protein-Ligand Binding Site Prediction Accuracy by Classification of Inner Pocket Points Using Local Features.” *Journal of Cheminformatics*, vol. 7, no. 1, Jan. 2015, doi:10.1186/s13321-015-0059-5.

Laskowski, Roman A. “SURFNET: A Program for Visualizing Molecular Surfaces, Cavities, and Intermolecular Interactions.” *Journal of Molecular Graphics*, vol. 13, no. 5, 1995, pp. 323–330., doi:10.1016/0263-7855(95)00073-9.

Levitt, David G., and Leonard J. Banaszak. “POCKET: A Computer Graphies Method for Identifying and Displaying Protein Cavities and Their Surrounding Amino Acids.” *Journal of Molecular Graphics*, vol. 10, no. 4, 1992, pp. 229–234., doi:10.1016/0263-7855(92)80074-n.

Pan, Xiaoyong, and Hong-Bin Shen. “RNA-Protein Binding Motifs Mining with a New Hybrid Deep Learning Based Cross-Domain Knowledge Integration Approach.” *BMC Bioinformatics*, vol. 18, no. 1, 2017, doi:10.1186/s12859-017-1561-8.

Sarafianos, Stefan G., et al. “Structure and Function of HIV-1 Reverse Transcriptase: Molecular Mechanisms of Polymerization and Inhibition.” *Journal of Molecular Biology*, vol. 385, no. 3, 2009, pp. 693–713., doi:10.1016/j.jmb.2008.10.071.

Tubbs, Julie L., et al. “Flipping of Alkylated DNA Damage Bridges Base and Nucleotide Excision Repair.” *Nature*, vol. 459, no. 7248, 2009, pp. 808–813., doi:10.1038/nature08076.

Weisel, Martin, et al. “PocketPicker: Analysis of Ligand Binding-Sites with Shape Descriptors.” *Chemistry Central Journal*, vol. 1, no. 1, 2007, p. 7., doi:10.1186/1752-153x-1-7.

Yu, Jian, et al. “Roll: a New Algorithm for the Detection of Protein Pockets and Cavities with a Rolling Probe Sphere.” *Bioinformatics*, vol. 26, no. 1, 2009, pp. 46–52., doi:10.1093/bioinformatics/btp599.

Zhu, Hongbo, and M. Teresa Pisabarro. “MSPocket: an Orientation-Independent Algorithm for the Detection of Ligand Binding Pockets.” *Bioinformatics*, vol. 27, no. 3, June 2010, pp. 351–358., doi:10.1093/bioinformatics/btq672.