Automated Ligand-Based Active Site Alignment: A Freely Available Extension to PyMOL

University of Toronto
Abraham Heifets[1,2]
Ryan Lilien[1,2,3]

Department of Computer Science[1]
Centre for Cellular and Biomolecular Research[2]
Banting and Best Department of Medical Research[3]

Computational Biology Lab
department of computer science
university of toronto

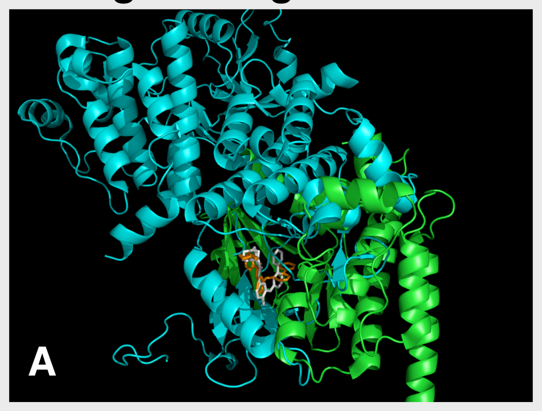
Overview

Abstract

The same ligand is likely to bind different proteins in similar, instructive ways. The goal of this project is to automate the comparison of active sites by developing freely available, easy to use visualization software [1]. We demonstrate a proof-of-concept, PyMol-based, structure visualization tool, which utilizes ligand-fragment based active site alignment.

Protein Alignment versus Ligand Alignment

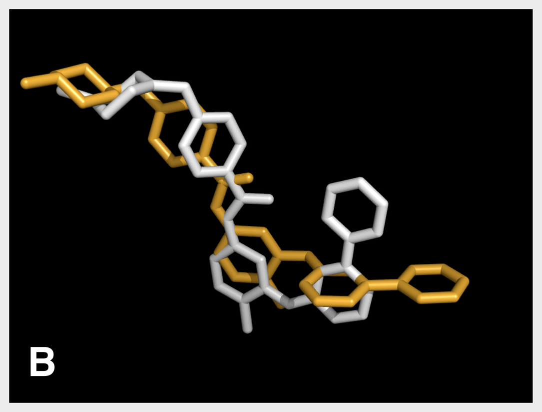
Although proteins which bind the same ligand may not exhibit a conserved global architecture, they are likely to possess a conserved local protein-ligand interface. We utilize bound ligand structures to compare the active sites of proteins which interact with a chosen ligand. Fig. A shows the Gleevec-based alignment of two proteins.



Gleevec alignment of 1XBB and 1OPJ

Ligand Flexibility

Given the structures of multiple protein-ligand complexes, we can align the active sites by properly aligning the ligands. The simplest method to align two ligands is a rigid transform but, as noted in [2], flexible molecules dock in a variety of conformations (Fig. B). This variability hinders comparison of active sites.



Gleevec conformational differences

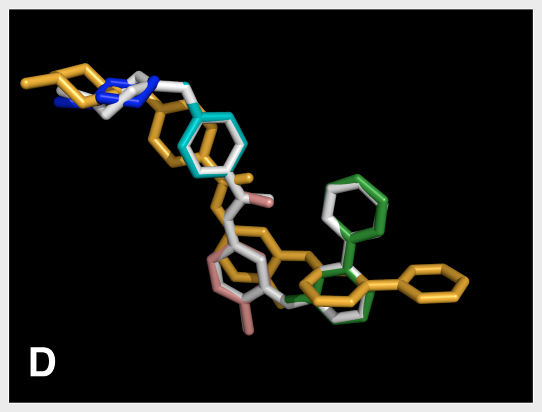
A Fragment-based Approach

We can enforce a correspondence across conformations by aligning ligand fragments. Given two ligands, one acting as a pivot (Fig. B, colored white) and one query to be fragmented (Fig. B, orange), we can automatically break the query into fragments (Fig. C) and independently align each.

Gleevec query ligand shown in original and fragmented forms

As we increase the number of fragments, the difference between the fragmented query and the pivot decreases (Fig. D).

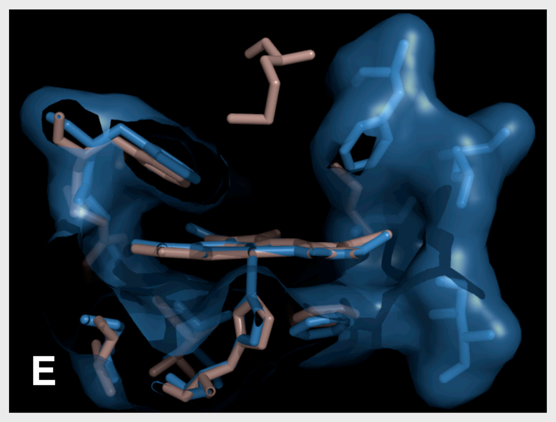
For each fragment, this generates an active site alignment based on the coordinate frame of only that fragment. This simplifies the comparison of local features within the active sites of multiple proteins.



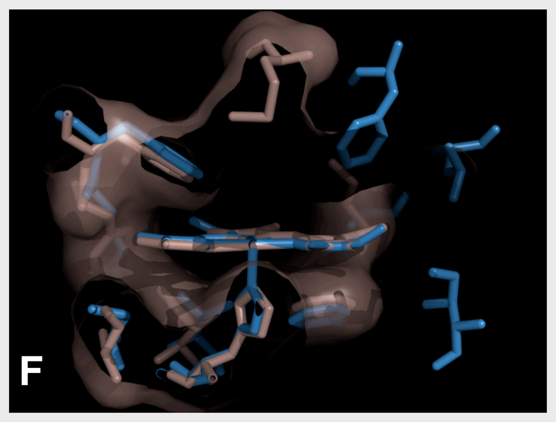
Query fragments aligned to white pivot

Active-site Variation and Conservation

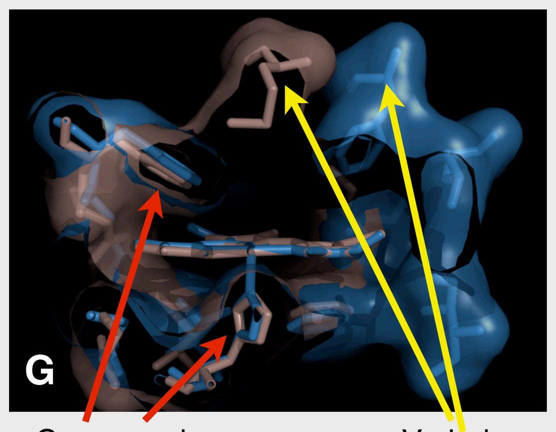
Rigid Alignment of C. lacteus mini-Hemoglobin and S. inaequivalvis Hemoglobin



Cut-away view of Hemoglobin for 1V07



Cut-away view of Hemoglobin for 1HBI

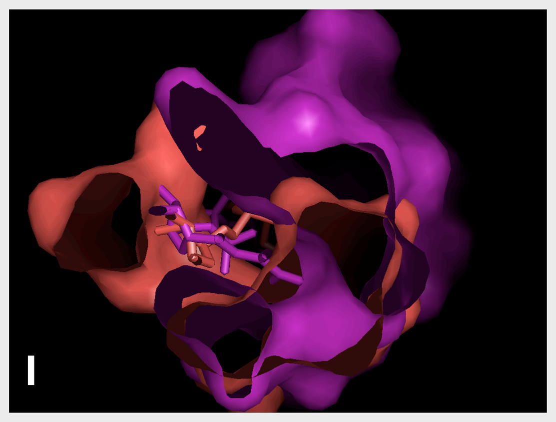


Conserved Varied

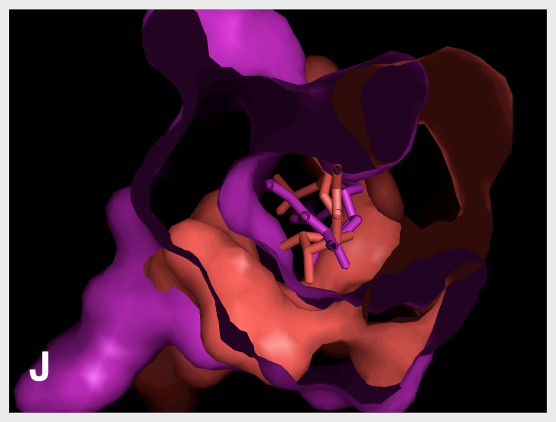
For ligands without internal degrees of freedom, our tool rigidly aligns the active sites of proteins which bind it. By constraining our visualization to the active site, we focus on the structural aspects contributing to protein-ligand binding. Our visualization allows the user to determine conserved and varying structural features, as well as to investigate ligand flexibility and conformational changes. Here, we may observe in the cross-section view that the region near the center of the porphyrin ring is conserved while regions more distant to the locus of activity vary.

Fragment-based Alignment of Diphtheria Toxin and Thermoautotrophicum NMN adenylyltransferase

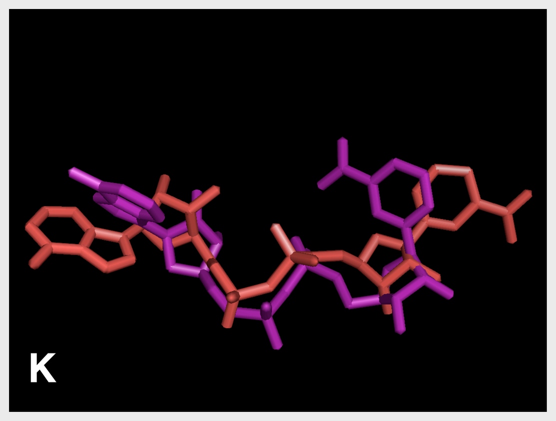
Ligand-based alignment 1TOX, 1EJ2



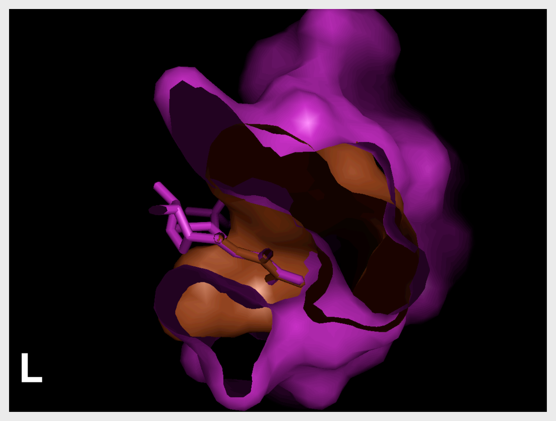
Rigid alignment, right-side of K



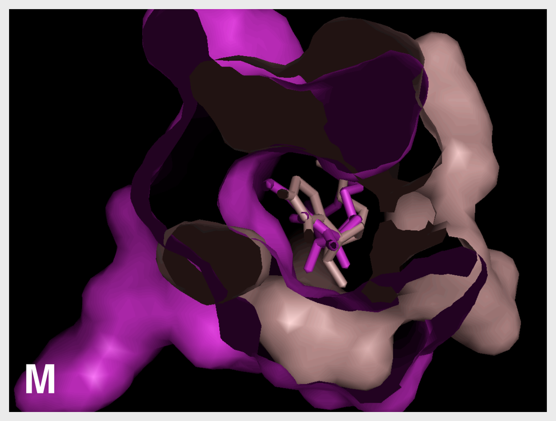
Rigid alignment, left-side of K



Rigid alignment of NAD ligand



Fragment alignment, right-side of K

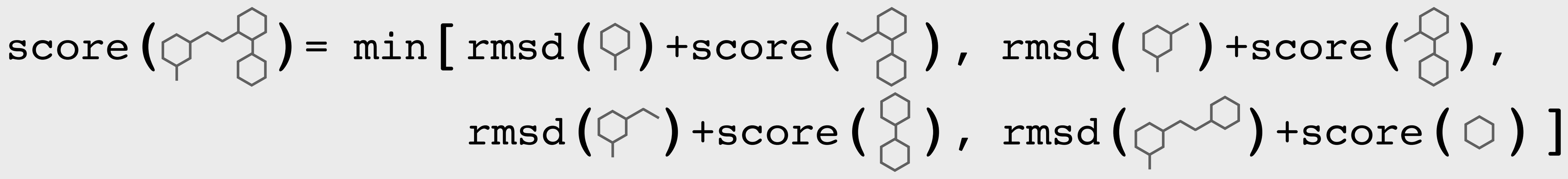


Fragment alignment, left-side of K

Both Diphtheria toxin and NMN adenylyltransferase bind to NAD. As can be seen in figure K, the minimal RMSD rigid alignment requires a skew in the alignment of the adenosine ring systems on the left of the molecule and a rotation in the relative positions of the nicotinamide rings on the right. Figures L and M show the benefit of aligning these regions separately. In figure L, we can see that the flexible alignment of the nicotinamide allows the local active site to be rotated to better align with the pivot active site, as compared to the rigid alignment in figure I. Similarly, the adenosine ring system on the left of NAD, when aligned independently as in figure M, allows the contours of the local protein surface to overlay more closely than with the rigid alignment in J. The structural conservation is better illustrated with fragment-based alignment.

Efficient Fragment Identification

Choose the fragmentation that yields the minimal score, where



Dynamic Programming

Finding the best set of fragments for a molecule can be solved in terms of the best fragmentations for pieces of the molecule. This recursive description is amenable to caching of the partial solutions, as well as necessary computations such as RMSD scoring. In the linear molecule case, this dynamic programming formulation reduces the runtime from exponential to polynomial. In branched molecules, however, alternative branches must still be checked.

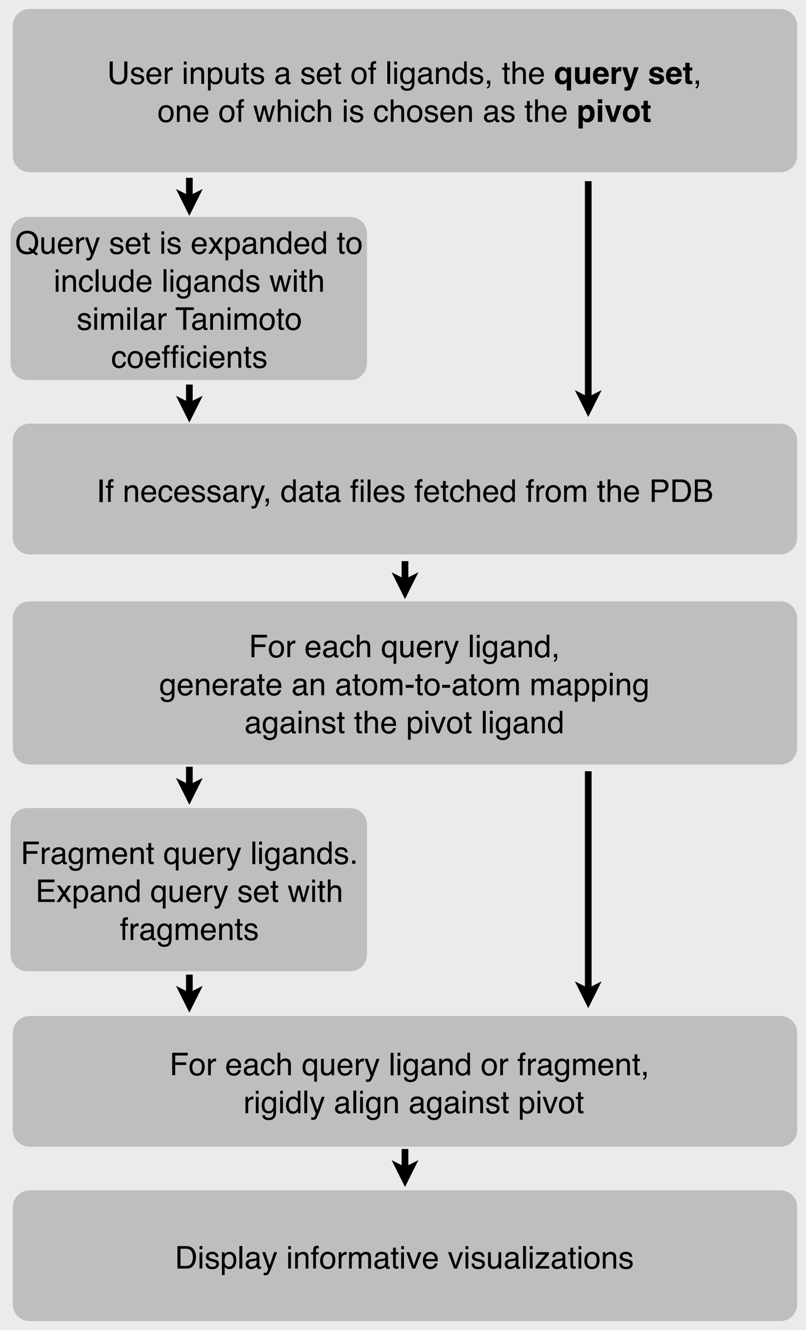
Branch and Bound

The score of a fragment is incrementally computed. Some pieces, such as the local RMSD score, are much cheaper to compute than others, such as the fragmentation of a substructure. This means we can use branch-and-bound techniques to stop the search quickly when we can prove that the cost of the current fragmentation will always return a higher score than a previously discovered solution.

Molecular Constraints

Because larger pieces of active sites have more area to contain interesting features, the sections of active sites around tiny fragments are usually not very informative. Therefore, it is often useful to restrict the minimum fragment size. Similarly, it is reasonable to require fragments to be connected components. Finally, rings may be considered rigid. These steps reduce the number of possible fragmentations and therefore improve the running time.

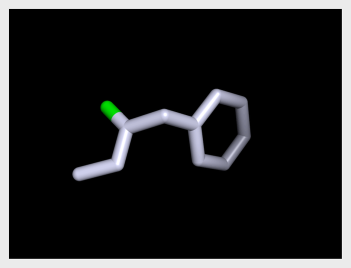
Process



Atom Mapping

Graph Fingerprints

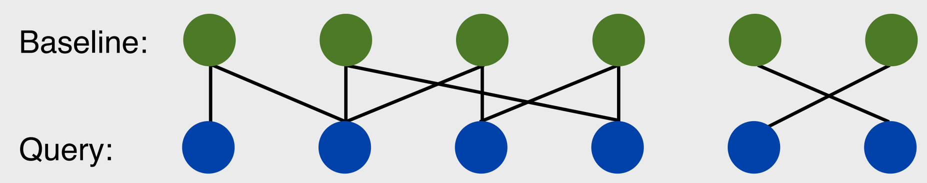
To score the alignment of two ligands, the system must have a mapping between the atoms in the molecules. We generate a graph-based fingerprint by computing the number of atoms at each distance from every atom. This can be efficiently computed using the Floyd-Warshall all-pairs distance algorithm.



A fingerprint for the green atom:

Distance	1	2	3	4
Counts	1	2	2	2

Given a fingerprint for every atom in the ligands of interest, we construct a correspondence that maximizes the similarity between the local neighborhoods around paired atoms. This can be efficiently computed as a bipartite matching problem. Care must be taken to permit inexact mappings, since we want to be able to compare proteins which are bound to similar but distinct ligands.



Runtime

$$\sum_{k=1}^{F} S(N,k) = \sum_{k=1}^{F} 1/k! \sum_{i=0}^{k} (-1)^i \binom{k}{i} (k-i)^N = \sum_{k=1}^{F}\sum_{i=0}^{k} \frac{(-1)^i (k-i)^n}{i!(k-i)!} \approx F^N, \text{ for } N \gg F$$

Possible partitions of a molecule with N atoms into F fragments grows as a summation of Stirling numbers of the second kind, up to the number of fragments. Fortunately, molecules tend to be sparse graphs.

References

1. Get it! http://compbio.cs.toronto.edu/ligalign
2. Kahraman A, Morris RJ, Laskowski RA, Thornton JM., "Shape variation in protein binding pockets and their ligands", J Mol Biol. 2007 Apr 20;368(1):283-301.
3. J.-C. Nebel, "Modelling of P450 active site based on consensus 3D structures", International Conference on Biomedical Engineering, BioMed 2005, Innsbruck, Austria, 16-18 Feb 2005.
4. DeLano, W.L. The PyMOL Molecular Graphics System, DeLano Scientific, Palo Alto, CA, USA, 2002
5. Weisstein, Eric W. "Stirling Number of the Second Kind." http://mathworld.wolfram.com/StirlingNumberoftheSecondKind.html