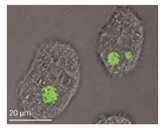
Computational Identification of Protein Catalytic Sites: Tests, Validation

Daniel Kirshner[1], Jerome Nilmeier[2], Felice Lightstone[2]

[1]Science Teacher and Researcher Program [2]Biosciences and Biotechnology Division, Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory

Abstract

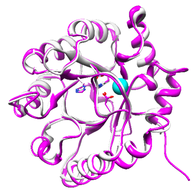
This project is one element of the analysis "pipeline" (illustrated, right) to characterize an organism that previously has not been well-studied. Once a protein of unknown structure has been computationally modeled (based on its sequence similarity to proteins with solved structures), then catalytic sites are identified on the model by comparison to a library of known sites. This work tested the identification algorithms with a set of proteins that have known structures and catalytic sites.



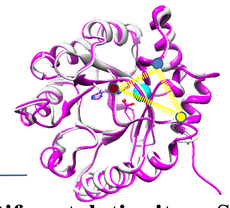
Organism of interest. In this case, *Francisella tularensis*

```
  1 msknylftse ...
 61 sawvdieelv ...
121 qglmfgfatn ...
181 fidtivlstq ...
241 cgltgrkiiv ...
301 ayaigvakpv ...
```

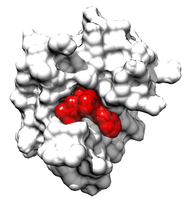
Sequencing: DNA → gene identification → amino acids



Homology model [1]: Create structure of target protein based on sequence similarity to proteins of known structure



Identify catalytic sites: Search target protein for sites similar to those in library of known site structures



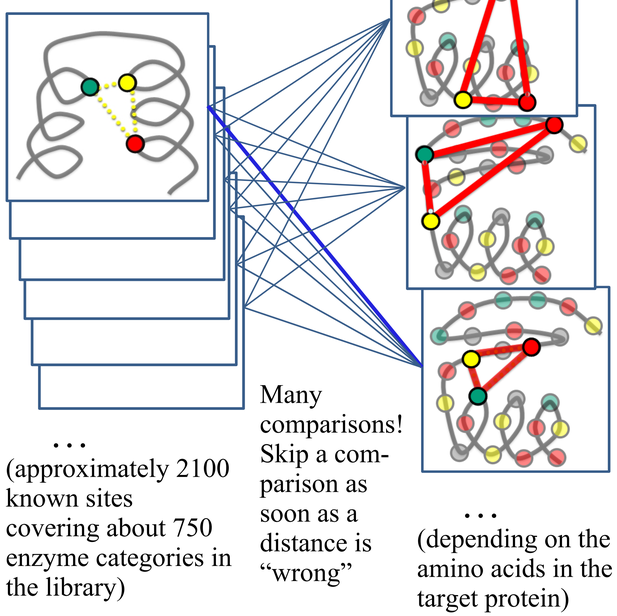
Docking: Search for molecules that can bind to site, blocking its function

Goal: Drug – prevent or cure disease

Identification screen 1: Compare target protein with library of known catalytic sites

For each site in the Catalytic Site Atlas [2], calculate the distances between critical amino acids ("catalytic residues"). (Distance measured from the principal backbone atom (C_{α}).

For each possible set of similar amino acids in the target protein, calculate the distances.

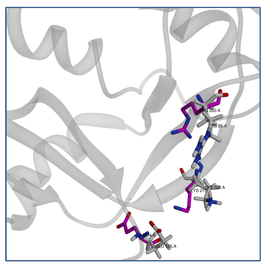


. . .
(approximately 2100 known sites covering about 750 enzyme categories in the library)

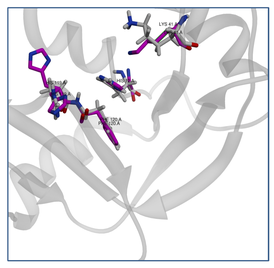
Many comparisons! Skip a comparison as soon as a distance is "wrong"

. . .
(depending on the amino acids in the target protein)

Identification screen 2: Compare alignments of critical amino acids of catalytic sites

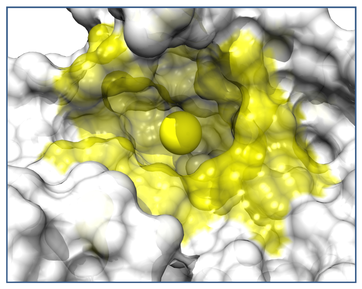


Example 1: different orientations. While the distances between the principal backbone atom (C_{α}) of each three amino acids in the target protein (white, gray) and the library catalytic site (magenta) are similar, the corresponding target and library amino acids are not oriented the same way. (Target: pdb 2k11; library protein: 1euy)

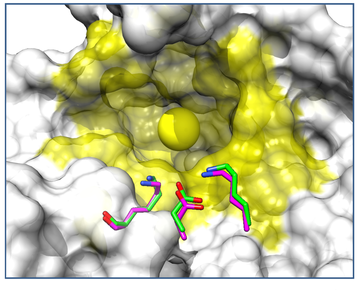


Example 2: similar orientations. This case shows a different set of four amino acids in the target protein (the same protein as in Example 1 – white, gray) and four corresponding amino acids in different library catalytic site (magenta). The corresponding target and library amino acids.have similar positions and orientations. (Target: 2k11; library: 1rbn)

Identification screen 3: Check candidate site critical amino acids against protein surface pockets



Identify surface binding sites/ pockets. SiteMap [3] finds pockets. (The yellow sphere indicates a binding site centroid for pdb 3ex4.)



Proximity to binding site. Amino acids identified as catalytic in steps 1 and 2 should be near a pocket.

Tests against known proteins/catalytic sites: initial results

The initial test was conducted using target proteins with known catalytic sites having four critical ("catalytic") amino acids. Target proteins were also restricted to high-resolution structures (<2A), having only one chain and only one binding site. These restrictions simplify the problem, with the aim of debugging/tuning the algorithm.

Target protein PDB ID	Target Enzyme Commission (EC) number	Catalytic Site Atlas top hit	Correct?	Notes
1e6u	1.1.1.271	1e7q	Yes	Screen 1 sufficient
1d3h	1.3.3.1	1d3g	Yes	Screen 1 sufficient
2k11	3.1.27.5	1euy	No	If screen 2 taken by itself, it would find correct hit
2zj3	2.6.1.16	1moq	Yes	Screen 1 sufficient
1a5p	3.1.27.5	1rbn	Yes	Screen 1 sufficient
3cuj	3.2.1.91	1tz3	No	In screen 1, the correct protein is hit #3
2kp1	5.3.4.1	1qz9	No	Screen 2, if adjusted for number of catalytic amino acids, would be correct

Recommendation: combine the scoring criteria (screens 1, 2, and 3) with a machine-training algorithm such as a logistic regression or a support-vector machine.

Acknowledgments and References

Test set derived from proteins with known catalytic sites clustered by sequence identity by Kristin Lennox.

Homology models calculated by Adam Zemla.

[1] Zemla, A., C. E. Zhou, et al. (2005). "AS2TS system for protein structure modeling and analysis." Nucleic acids research **33(suppl 2):** W111.

[2] Porter, C. T., G. J. Bartlett, et al. (2004). "The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data." Nucleic acids research **32(suppl 1):** D129.

[3] Halgren, T. (2007). "New Method for Fast and Accurate Binding-site Identification and Analysis." Chemical Biology & Drug Design **69(2)**: 146-148.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. IM number LLNL-POST-491696.