Rapid Catalytic Template Searching as an Enzyme Identification Procedure

Jerome P. Nilmeier, Daniel A. Kirshner, Felice C. Lightstone

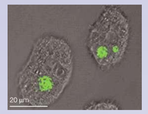
Biosciences and Biotechnology Division, Physical and Life Sciences Directorate

Lawrence Livermore National Laboratory

Abstract:

As a component of the Protein Function Prediction Platform, we present a catalytic site identification procedure. It uses a template-matching algorithm and a scoring procedure that allows for rapid, scalable protein-to-template matching for catalytic sites from a catalog of binding sites. We develop the procedure using the Catalytic Site Atlas (CSA) of Thornton. The procedure is able to process cofactors, ions, nonstandard residues, and point substitutions for both residues and ions. Sites with two critical residues are challenging cases, resulting in AUCs of 0.9411 and 0.5413 for the training and test sets, respectively. The remaining sites show excellent performance with AUCs greater than 0.90 for both the training and test data on templates of size greater than two critical residues.

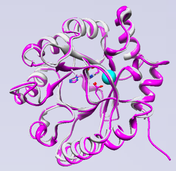
Protein Function Prediction (PFP) Platform: Sequence to Structure



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

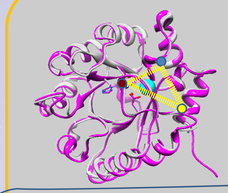
```
  1 msknylftse ...
 61 sawvdieelv ...
121 qglmfgfatn ...
181 fidtivlstq ...
241 cgltgrkiiv ...
301 ayaigvakpv ...
```

Sequencing
DNA → gene identification
→ amino acid sequence

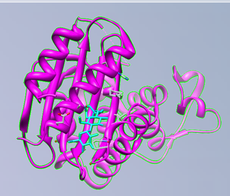


Homology models[1]
amino acid sequence → protein structure

Protein Function Prediction (PFP) Platform: Mechanistic Modeling



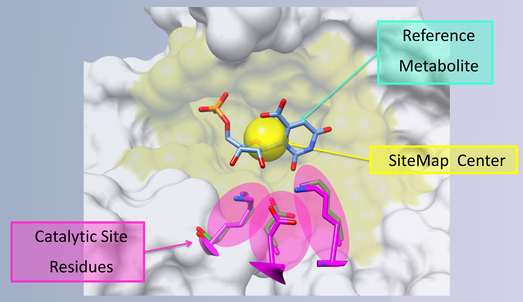
Identify catalytic sites[2]
Search target protein for sites similar to those in library of known site structures



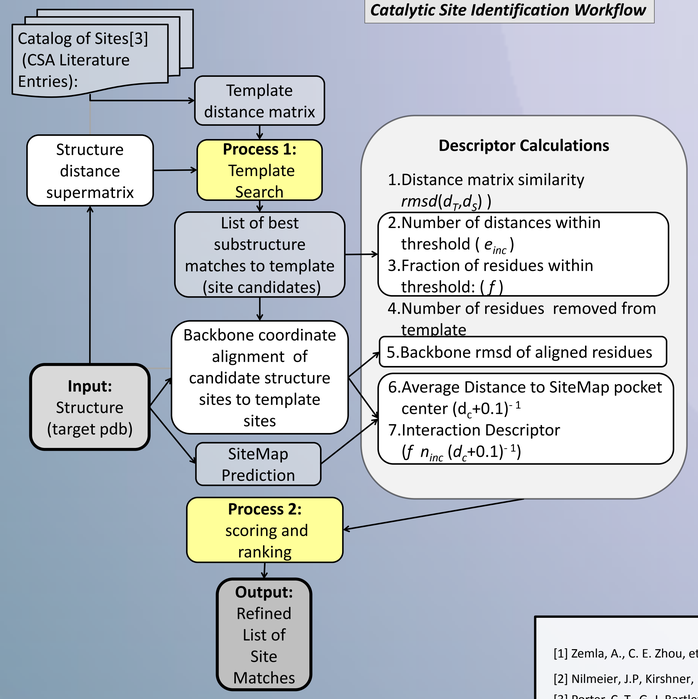
Docking
Search for molecules that can bind to site, blocking its function

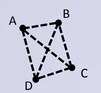
Goal
Drug – prevent or cure disease

The Challenge of Catalytic Site Identification:
Given a target structure, can we find a catalytic template that best matches the structure? If so, we can assign function.



Catalytic Site Identification Workflow

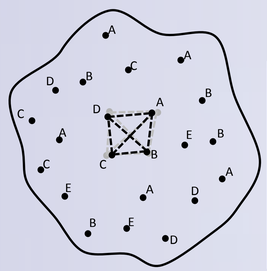




Template Distance Matrix (T)

The core search algorithm: Subgraph Isomorphism and Similarity Search

This is an NP hard problem, but approximations and pruning allow for linear scaling:

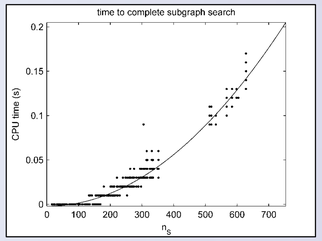


Structure (target pdb) (S)
Best Fitting subgraph superimposed

Scaling is linear in template size and quadratic in target size.

$$P_{full} = \prod_{i=1}^{n_T} \Omega_S(r_i) \cong \overline{\Omega}_S^{\,n_T}$$

$$P_{fast} \cong \overline{\Omega}_S\left[\overline{\Omega}_S + P_{max}\overline{\Omega}_S \cdot (n_T - 2)\right]$$



Incorporation of descriptor based scoring procedure (logistic regression) Improves performance.

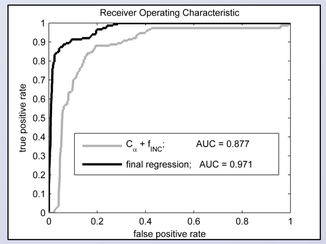
$$f(z) = (1+e^{-z})^{-1}$$

Logistic Function

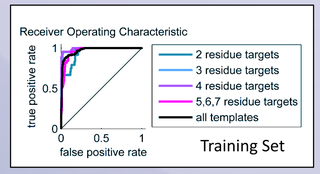
$$z = \beta_0 + \sum_i \beta_i(n_T)\cdot x_i$$

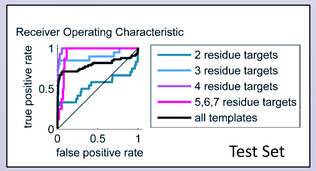
Descriptor

Template size specific parameter



Larger binding sites show excellent performance, and regression improves 3 and 4 residue cases substantially

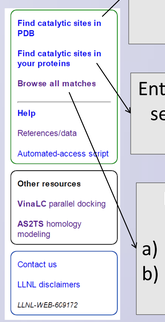




	AUC(train)	AUC(test)
n_T=2	0.9411	0.5413
n_T=3	0.9821	0.9040
n_T=4	0.9932	0.9935
n_T=5,6,7	0.9622	0.9369
All	0.9714	0.7989

The Next Phase…searching the Protein Datatbank:
http://catsid.llnl.gov
Administered by Dan Kirshner

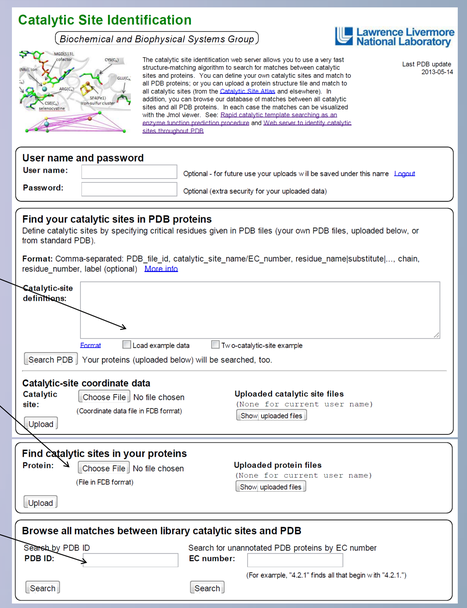
Server Details:
- 8 compute nodes (2x6 cores each)
- Intel Xeon X5690/3.46 GHz
- 96 GB memory per node.
- Parallelization: openMP
- Full PDB (>80k structures) search in <1min (on one node).



Enter user defined catalytic template(s) and search the protein databank .

Enter user defined protein search the catalytic site atlas.

Browse precomputed searches:
a) by pdb id
b) By Enzyme Comission Number



References

[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] Nilmeier, J.P, Kirshner, D.K. Lightstone, F.C., T. (2012). "Rapid Catalytic Template Searching as an Enzyme Identification Procedure" PLoS Comp Bio *In preparation*

[3] 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.

Nilmeier, J. P., Kirshner, D.A., Wong, S.E., Lightstone, F. C. (2013). "Rapid Catalytic Template Searching as an Enzyme Function Prediction Procedure." PLoS ONE, **8**(5): p.e62535

Nilmeier, J. P., Kirshner, D.A., Lightstone, F. C. (2013). "Catalytic site identification - a web server to identify catalytic site structural matches throughout PDB" (2013) Nucleic Acids Research.

url: http://catsid.llnl.gov/

Acknowledgements and References

The authors gratefully acknowledge the Defense Threat Reduction Agency (DTRA), for supporting this work, (grant B094679I), under the guidance of Carol Zhou. We would like to thank Carol Zhou and Jenn Sirp for helpful discussions in developing the automated procedures. We also thank Kristin Lennox and Eithon Cadag for helpful advice on the statistical modeling procedures. Sergio Wong and Kristin Lennox helped to provide initial datasets during development. We thank Adam Zemla, Sergio Wong, and Sebnem Essiz Ghokan for providing helpful discussions. This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. Release: LLNL-POST-637693