

Automatic Prediction of Functional Site Regions in Low-Resolution Protein Structures

Jaspreet Singh Sodhi, Liam J. McGuffin, Kevin Bryson, Jonathan J. Ward,
Lorenz Wernisch and David T. Jones

Bioinformatics Group
Department of Computer Science
University College London
London United Kingdom
j.sodhi@cs.ucl.ac.uk
www.cs.ucl.ac.uk/staff/J.Sodhi

Abstract

World-wide structural genomics initiatives are rapidly accumulating structures for which limited functional information is available. Additionally, state-of-the art structural prediction programs are now capable of generating at least low resolution structural models of target proteins. Accurate detection and classification of functional sites within both solved and modelled protein structures therefore represents an important challenge.

We present a fully automatic site detection method, FuncSite, that uses neural network classifiers to predict the location and type of functionally important sites in protein structures. The method is designed primarily to require only backbone residue positions without the need for specific side-chain atoms to be present.

In order to highlight effective site detection in low resolution structural models FuncSite was used to screen model proteins generated using mGenTHREADER on a set of newly released structures. We found effective metal site detection even for moderate quality protein models illustrating the robustness of the method.

1. Introduction

The wealth of biological data being generated world-wide has resulted in major new challenges to allow in depth understanding of complex biological processes. In particular structural genomics initiatives, which aim to decipher the structure of target proteins on a genome-wide scale, are rapidly accumulating structures without clear functional assignments. Consequently there is a clear need for effective tools to improve the information content of genomic

databases as well as direct experimental biologist to fully exploit the vast diversity provided in nature.

Sequence searching tools have now become routine in initial investigations of new protein and DNA sequences. However, sequence based methods are incapable of directly encoding the three-dimensional spatial organization of functional residues and the atoms responsible for biochemical action in the folded protein.

Here we present FuncSite, a novel approach using artificial neural networks (ANN) to predict functional site regions in protein structure. The method is designed to identify a variety of site regions in super-families by combining PSI-BLAST sequence profile and structural information. We have applied the approach to a selection of metal binding sites, DNA binding interfaces and adenylate binding pockets. We avoid using side-chain atom information in order to allow effective site prediction in moderate quality structural models. We also show FuncSite predictions to be a useful distinguisher to aid in the effective ranking of model proteins.

2. Methods

2.1. Site Features and Definitions

Given a protein structure we retrieve many different attributes to encode sites. For each site, consisting of N residue (seed and $N-1$ neighbors), we define $20N$ PSI-BLAST profile scores, $3N$ secondary structure states, N solvent accessibility scores and an inter-atomic distance matrix between the N residues ($\frac{N(N-1)}{2}$). Residues within interacting range of a functionally relevant prosthetic group are labeled as site residues (Figure 1).

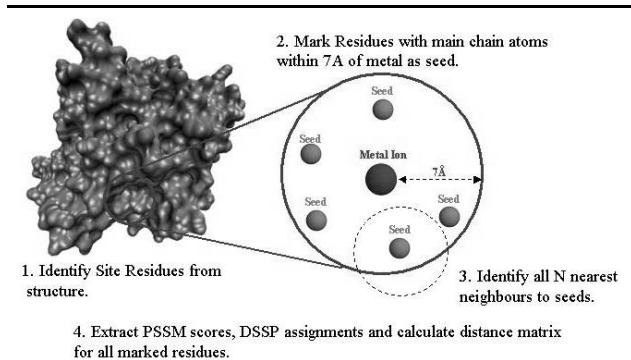


Figure 1. Encoding of Site Patterns.

2.2. Cross-Validation

The dataset was divided into groups for cross-validation such that proteins belonging to the same super-family were grouped together. This is a robust measure to ensure effective generalization across super-families.

2.3. Generating Structural Models

The mGenTHREADER[1, 2] fold recognition method was used to generate protein models. Model proteins were generated for a set of LiveBench [5] targets, known to contain metal binding sites.

3. Results

FuncSite was trained to identify commonly occurring metal binding sites occurring in nature. The top ranking FuncSite predictions, for super-family members containing functionally important metal ions, identified 85.9% of sites with a selectivity of 73.5%. We also screened newly released structures from the LiveBench structure prediction assessments. FuncSite was able to accurately detect the correct metal binding region in 19/24 (71.2%) of these targets. Figure 2 illustrates two example targets for which no metal binding information was obtainable using Interpro.

The method was also able to locate a Mn^{2+} binding site in the yeast POP2, as described in the literature, protein although the metal ion was absent in crystal structure. Interestingly, we also observed a Fe^{3+} strong hit in the structural genomics target HI0817 from *H. influenzae*.

Reasonable structural models were generated for 15 of the 24 metal containing targets. Of these modeled structures FuncSite correctly predicted 53% of metal binding sites.

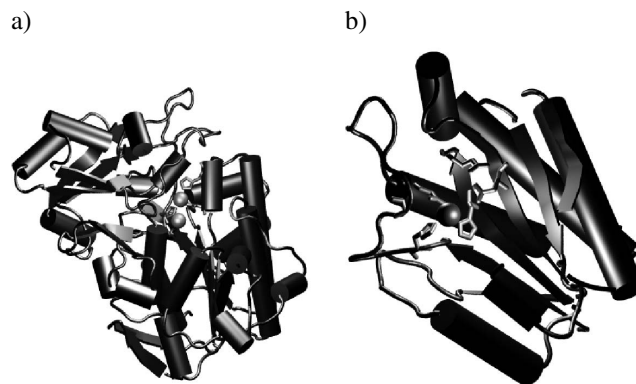


Figure 2. Metal site predictions of LiveBench targets a) N-acetylmuramoyl-L-alanine Amidase and b) Phosphonoacetate Hydrolase.

4. Conclusions

We have developed a functional site prediction tool capable of identifying sites in low resolution structural models. Functional site predictions in modeled structures are also likely to complement fold recognition methods, the correct spatial clustering of functionally important residues could be used as a measure of structure prediction quality, and efforts are underway to determine how FuncSite may be extended to improve structural predictions. Application of the method to identify DNA binding interfaces and adenylate binding pockets is promising as well as application of the method to structural models that have been generated across genomes for the Genomic Threading Database [3, 4]

References

- [1] D. T. Jones. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J Mol Biol*, 287(4):797–815, 1999.
- [2] L. J. McGuffin and D. T. Jones. Improvement of the GenTHREADER method for genomic fold recognition. *Bioinformatics*, 19(1):874–881, 2003.
- [3] L. J. McGuffin, S. Street, K. Bryson, S. Sorensen, and D. T. Jones. The Genomic Threading Database: a comprehensive resource for structural annotations of the genomes from key organisms. *Nucl. Acids Res. Database Issue*, 32:D196–D199, 2004.
- [4] L. J. McGuffin, S. Street, S. Sorensen, and D. T. Jones. The Genomic Threading Database. *Bioinformatics*, 20(1):131–132, 2004.
- [5] L. Rychlewski, D. Fischer, and A. Elofsson. LiveBench-6: Large-Scale Automated Evaluation of Protein Structure Prediction Servers. *Proteins (supplement)*, 53(6):542–547, 2003.