Prediction of Functional Sites in SCOP Domains using Dynamics Perturbation Analysis

Judith D. Cohn[1], Dengming Ming[1,4], Michael E. Wall*[1,2,3]
[1]Computer, Computational, and Statistical Sciences Division, Los Alamos National Laboratory, Los Alamos, NM, 87545 USA
[2]Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM, 87545 USA
[3]Center for Non-Linear Studies, Los Alamos National Laboratory, Los Alamos, NM, 87545, USA
[4]School of Life Sciences, Nanjing University, Nanjing, Jiangsu, 210093, People's Republic of China
*To whom correspondence should be addressed: mewall@lanl.gov

1. INTRODUCTION

Dynamics perturbation analysis (DPA) (1-3) finds regions in a protein structure where proteins are "ticklish," *i.e.*, where interactions cause a large change in protein dynamics. Such regions were shown to predict the locations of native binding sites in a docking test set (3), but the more general applicability of DPA to prediction of functional sites in proteins was not shown. Here we describe the results of applying an accelerated algorithm, called Fast DPA (4), to predict functional sites in over 50,000 SCOP domains.

2. METHODS

Fast DPA was performed as described in (4) on crystallographically determined protein structures consisting of a single chain or subset of a chain in SCOP (5) version 1.65. To validate DPA-predicted functional sites, we looked at two different types of annotations: a) catalytic residues in the Catalytic Site Atlas (CSA) (6); and b) protein residues close to a small molecule in the PDB. Sequence conservation of predicted sites (7) was used to filter false positives, and multiple sequence alignments (MSAs) (8) were used to assess the significance of predictions. Selected cases were the subject of deeper investigation in the literature.

3. RESULTS AND DISCUSSION

Table 1. Summary of the matches of DPA predictions to annotated sites in SCOP domains.

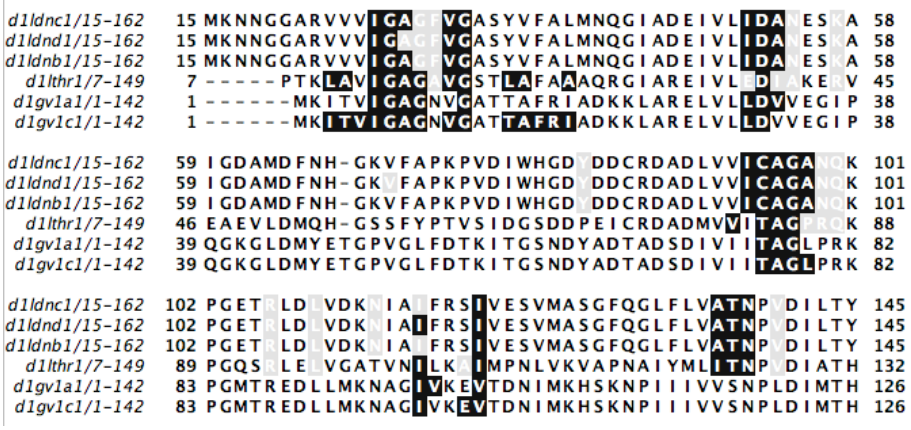
		Sequence Conservation			
	All	**No Info**	**Low**	**Medium**	**High**
Total Sites	63,787	3,321	50,555	4,905	5,006
Match Found	28,899	1,342	18,773	4,196	4,588
Binding Site	17,995	969	13,997	1,774	1,255
Catalytic Site	3,751	145	1,875	907	824
Both	7,153	228	2,901	1,515	2,509
No Match Found	34,888	1,979	31,782	709	418

Fast DPA predicted 63,787 functional sites on 49,245 SCOP domains; yielding O(1) predictions per domain, which is commensurate with the number of functional sites that we naively expect proteins to have. In these domains, CSA identified 49,834 catalytic residues, of which DPA correctly predicted 22,296 (44.6%). In addition, we inferred 86,305 binding sites based on the PDB. DPA predicted at least one binding residue in 38,853 of these sites (42.7%). In 43.3% of these matching sites, at least 50% of the binding residues were predicted, while in 69.6% of the sites, at least 25% were predicted. Table 1 summarizes the matches of DPA predictions to annotated sites in SCOP domains, organized according to degree of sequence conservation. Because these annotations incompletely characterize information about protein functional sites, we expect the match statistics to represent a lower limit on future performance. Based on the table, we estimate that at least 45.3% of DPA sites will match a binding site or a catalytic residue. We also estimate that at least 91.6% of highly-conserved DPA sites, and at least 88.6% of medium- or highly-conserved sites will match a binding site or catalytic residue. Further, we observe that sequence conservation provides greater enrichment for detection of catalytic sites than binding sites: 9.4%

of low-conservation sites overlap a catalytic site, compared to 68.2% of high-conservation sites; however, 33.4% of low-conservation sites overlap a binding site, compared to 75% of high-conservation sites. *The results therefore recapitulate much of the known information about functional sites in SCOP domains, and validate the general use of DPA to predict functional sites in proteins.*

We also used MSAs to assess the variation in DPA predictions for similar protein structures. An example is illustrated in Fig. 1 for SCOP family C.2.1.5, where DPA residues are highlighted in black. The first three sequences are equivalent chains from PDB entry 1LDN, and the last two are equivalent chains from 1GV1. These cases illustrate that DPA results are insensitive to crystal packing differences. Predictions are similarly robust across sequences from different proteins, enabling annotation transfer. Although there is no functional site annotation for PDB entry 1GV1, a similar site in 1LDN and 1THR has an NAD bound. The annotated domains are lactate dehydrogenases while the unannotated domains are malate dehydrogenases, which also use NAD as a coenzyme. We therefore predict that the DPA site in 1GV1 is an NAD functional site. We saw similar alignments in 44 of 50 families that we have examined in this way, potentially enabling transitive annotations for 157 predicted sites. *MSAs can therefore be used in combination with DPA to predict and annotate functional sites in proteins.*

Figure 1. Alignment of DPA sites with high sequence conservation from the SCOP family C.2.1.5. The display was generated using Jalview (9).



4. REFERENCES

1. Ming D. and Wall M.E. 2005. Quantifying allosteric effects in proteins. *Proteins* 59:697-707.

2. Ming D. and Wall M.E. 2005. Allostery in a coarse-grained model of protein dynamics. *Phys Rev Lett* 95:198301.

3. Ming D. and Wall M.E. 2006. Interactions in native binding sites cause a large change in protein dynamics. *J. Mol. Biol.* 358:213-223.

4. Ming D., Cohn J.D. and Wall M.E. 2008. Fast dynamics perturbation analysis for prediction of functional sites. *BMC Structural Biology* 8:5.

5. Murzin A.G., Brenner S.E., Hubbard T and Chothia C. 1995. SCOP: a structural classification of protein database for the investigation of sequences and structures. *J. Mol. Biol.* 247:536-540.

6. Porter, C. T., Bartlett, G. J. and Thornton, J. M. 2004. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res* 32:D129-33.

7. Sander, C. and Schneider, R. 1993. The HSSP data base of protein structure-sequence alignments. *Nucleic Acids Res* 21:3105-9.

8. Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J. and Higgins, D. G. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* 23:2947-8.

9. Clamp, M., Cuff J., Searle S.M. and Barton G.J. 2004. The Jalview Java Alignment Editor. *Bioinformatics* 20:426-427.