Title

Using structural templates to recognise metal binding sites

Authors

James Torrance, Malcolm McArthur, Janet Thornton

Abstract

Structural templates representing metal binding sites have the potential to be a useful tool for identifying unrecognised metal binding sites and spotting similarities between known binding sites. We created a set of structural templates and analysed how well these discriminate matches in related structures from random matches. We found that templates based on C_α and C_β atom positions discriminate relatives from random matches with high sensitivity and predictive accuracy. By contrast, templates based on the atoms that are directly involved in binding metal were good at detecting metal binding sites in non-relatives.

Introduction

Small groups of residues in protein structures can be described using structural templates. These templates can then be used to search other protein structures for structurally similar groups of residues, which may have quite different sequence positions. We have previously used structural templates to characterise enzyme catalytic sites.[1] This paper describes the extension of that work to cover metal binding sites.

Proteins that contain metals are essential for the basic processes of life. Protein-bound metals play diverse roles including participating in enzymatic catalysis, facilitating electron transfer, regulating protein function, and maintaining protein structure. The most commonly occurring biological metals are Ca, Mg, Mn, Fe, Cu, Na and K.[3] Using templates to characterise metal binding sites can be useful, for the following reasons:

- Proteins which bind metals *in vivo* may lack metals in a crystal structure. If a group of residues matches a structural template describing a metal binding site, this may indicate that a metal is present *in vivo*.
- Recognising structural similarities between different metal binding sites can aid the study of metal binding and assist in understanding protein function. Structural templates make it easy to search the Protein Databank (PDB)[2] for metal binding sites that resemble one another.

We have created structural templates for a set of metal binding sites and investigated their ability to discriminate between related metal binding sites and random matches. Various analyses of metal binding site structure have previously been carried out.[3,4] However, we believe this is the first time structural templates have been used in the study of metal binding sites.

Methods

The methods used were based on those described in Torrance *et al.* (2005).[1] A set of 28 metal binding sites was taken from a set of protein structures all with resolution better than 1.6 Å, which were non-redundant at the level of CATH[5] and SCOP[6] structural superfamily. For each of these binding sites, a structural template was constructed. These structural templates described those residues that directly bound the metal ion. Two different types of template were created for each site. The first type of template represented residues in terms of the positions of their C_α and C_β atoms, and was therefore a reflection of backbone orientation. The second type represented each residue using only those atoms directly involved in binding the metal. This allowed us to investigate which atom subset resulted in the most effective templates. Note that the structural templates did *not* include a description of the metal ions themselves; this allowed the templates to match to metal binding sites where the metal does not feature in the crystal structure.

Relatives of these proteins were identified using PSI-BLAST[7] with an E-value threshold of 0.0005 and a maximum of 20 iterations. For the purposes of this paper, the set of relatives for a given template is referred to as its "family". The structural templates were used to compare the metal binding sites in the original proteins with the equivalent binding sites in these relatives. The comparison was carried out using the structural template matching program Jess.[8] Structural differences were quantified in terms of RMSD.

The structural templates were then used to search for all random hits in a non-redundant subset of the Protein Databank. The purpose of this was to establish the distribution of RMSDs for random template matches, so that this could be compared with the distribution for meaningful matches. The non-redundant subset of the PDB was based on the non-redundant chain set provided by the NCBI (www.ncbi.nlm.nih.gov/Structure/VAST/nrpdb.html); structures were regarded as redundant with one another if there was a BLAST hit between them with an E-value of 10^{-80} or less. The non-redundant subset searched by each template was filtered so that it contained no structures from the same SCOP or CATH superfamily as the template protein.

We examined how well the structural templates could discriminate between random matches and metal binding sites from family members.

Results

The distribution of family and random matches for templates of different sizes is shown in Figure 1, which combines the data for all families. Almost all family matches had RMSDs below 0.6 Å. The great majority of random matches had RMSDs above 0.6 Å. C_α/C_β templates tended to have lower RMSD matches than metal binding templates. The separation of family matches from random ones was consistently better for C_α/C_β templates than for metal binding atom templates. For some metal binding atom templates, there was considerable overlap between family and random matches. This overlap took the form of a group of random matches that did not conform to the overall distribution of the random matches, implying that they were not, in fact, random. These can be seen as the small peak in random matches that occurs at 0.2 Å in Figure 1. An equivalent phenomenon could not be observed for the C_α/C_β templates. This point will be returned to below.

In order to use templates to discriminate between family and random matches, it was necessary to set a threshold level of RMSD: below this threshold, matches were predicted as family members; above it, they were predicted as random. The optimal threshold level was defined as the one that maximised Matthews Correlation Coefficient[9] (MCC, an overall measure of the separation of family matches from random matches). A threshold was assigned individually for each template.

Table 1 shows performance levels for different types of template. The templates separated family matches from random ones very well: C_α/C_β templates had a MCC of 0.92, and similar scores for sensitivity and predictive accuracy. The C_α/C_β templates outperformed the metal binding atom templates on every measure.

Table 1: Ability of structural templates to discriminate matches to relatives from random matches All "mean" values are means of all values for the 28 individual templates. Values in brackets are standard deviations.

Template type	Mean Matthews Correlation Coefficient	Mean sensitivity	Mean predictive accuracy
Metal binding atoms	0.74 (0.28)	0.87 (0.25)	0.73 (0.34)
C_α/C_β	0.92 (0.17)	0.92 (0.20)	0.94 (0.14)

The lower performance of metal binding atom templates compared to C_α/C_β templates is largely due to a lower predictive accuracy: many structures have RMSDs better than the optimal threshold, but are not family members. Figure 1 suggests that this is not because the random distribution of matches overlaps with the distribution of family members, but at is at least partly due to a distinct population of matches. This suggests the possibility that these "false positives" may in fact be similar metal binding sites that have evolved independently (convergence).

To investigate this possibility, we analysed all apparent false positive hits on ligand atom templates to discover whether a metal ion lay within 3 Å of all residues matched. Where this was the case, we checked that the false positive had a SCOP code assigned, in order to ensure that it came from a different superfamily to the template protein. Out of 602 apparent false positive matches to ligand atom templates, 239 were metal binding sites whose SCOP codes showed them to be unrelated to the template protein.

The same analysis was repeated for C_α/C_β templates. There were only 14 apparent false positive matches to any of the 28 templates; only two of these were metal binding sites whose SCOP codes showed them to be unrelated to the template protein.

Discussion

Structural templates that describe metal binding residues in terms of their C_α and C_β atom positions can discriminate related metal binding sites from random matches very reliably. Structural templates

that describe metal binding sites in terms of the atoms directly involved in binding metal perform less well. This may be because the positions of C_α and C_β atoms seem to be better structurally conserved between related structures than the positions of the atoms that bind metal directly. This better structural conservation may seem surprising, given that the metal binding atoms are constrained in position by the need to bind the ion. It may be that the C_α/C_β atoms are generally better defined crystallographically and less flexible than the metal binding atoms, which usually lie in residue sidechains.

The lower performance of metal binding atom templates is also partly due to the fact that these templates pick up apparent false positives that are actually meaningful matches with unrelated metal binding sites. C_α/C_β templates pick up almost no such meaningful matches to non-relatives. This suggests that structural templates based on metal binding atoms may be more useful in comparing metal binding sites between non-relatives than structural templates that are based on C_α/C_β atoms.

References

[1] Torrance, J., Bartlett, G., Porter, C. & Thornton, J. (2005). Using a Library of Structural Templates to Recognise Catalytic Sites and Explore their Evolution in Homologous Families. *J Mol Biol* **347**, 565–81.

[2] Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I. & Bourne, P. (2000). The Protein Data Bank. *Nucleic Acids Res* **28**, 235–42.

[3] Harding, M. (2004). The architecture of metal coordination groups in proteins. *Acta Crystallogr D Biol Crystallogr* **60**, 849–59.

[4] Glusker, J.P.; Katz, A. B. (1999). Metal ions in biological systems. *Rigaku J* **16**, 8–16.

[5] Pearl, F., Todd, A., Sillitoe, I., Dibley, M., Redfern, O., Lewis, T., Bennett, C., Marsden, R., Grant, A., Lee, D., Akpor, A., Maibaum, M., Harrison, A., Dallman, T., Reeves, G., Diboun, I., Addou, S., Lise, S., Johnston, C., Sillero, A., Thornton, J. & Orengo, C. (2005). The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucleic Acids Res* **33 Database Issue**, D247–51.

[6] Murzin, A., Brenner, S., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* **247**, 536–40.

[7] Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–402.

[8] Barker, J. & Thornton, J. (2003). An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. *Bioinformatics* **19**, 1644–9.

[9] Matthews, B. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta* **405**, 442–51.