BIOINFORMATICS APPLICATIONS NOTE

Vol. 26 no. 22 2010, pages 2920–2921
doi:10.1093/bioinformatics/btq543

Structural bioinformatics

Advance Access publication September 22, 2010

The binding site distance test score: a robust method for the assessment of predicted protein binding sites

Daniel B. Roche, Stuart J. Tetchner and Liam J. McGuffin*

School of Biological Sciences, University of Reading, Whiteknights, Reading RG6 6AS, UK

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: We propose a novel method for scoring the accuracy of protein binding site predictions—the Binding-site Distance Test (BDT) score. Recently, the Matthews Correlation Coefficient (MCC) has been used to evaluate binding site predictions, both by developers of new methods and by the assessors for the community-wide prediction experiment—CASP8. While being a rigorous scoring method, the MCC does not take into account the actual 3D location of the predicted residues from the observed binding site. Thus, an incorrectly predicted site that is nevertheless close to the observed binding site will obtain an identical score to the same number of non-binding residues predicted at random. The MCC is somewhat affected by the subjectivity of determining observed binding residues and the ambiguity of choosing distance cutoffs. By contrast the BDT method produces continuous scores ranging between 0 and 1, relating to the distance between the predicted and observed residues. Residues predicted close to the binding site will score higher than those more distant, providing a better reflection of the true accuracy of predictions. The CASP8 function predictions were evaluated using both the MCC and BDT methods and the scores were compared. The BDT was found to strongly correlate with the MCC scores while also being less susceptible to the subjectivity of defining binding residues. We therefore suggest that this new simple score is a potentially more robust method for future evaluations of protein–ligand binding site predictions.

Availability: http://www.reading.ac.uk/bioinf/downloads/

Contact: l.j.mcguffin@reading.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 5, 2010; revised on September 9, 2010; accepted on September 20, 2010

1 INTRODUCTION

The prediction of a protein's ligand binding site location and potential interacting residues is important in the elucidation of protein function, *de novo* drug design, mutagenesis studies and ligand binding specificity (Lopez *et al.*, 2009; Sankararaman *et al.*, 2010). The CASP experiment included a function prediction category for the first time in CASP6 (Soro and Tramontano, 2005), where the aim was to predict the Enzyme Commission number (EC) and Gene Ontology (GO) terms. Due to the difficulty in assessing these terms, the CASP7 (Lopez *et al.*, 2007) assessors decided that CASP was not the best place for this format of function prediction. Thus, for CASP8, function prediction was included in a different format, with the assessment of observed ligand binding site residues, as many CASP targets were shown to crystallize with biologically interesting ligands (Lopez *et al.*, 2009).

In CASP8, function predictions were assessed using the Matthews correlation coefficient (MCC) (Matthews, 1975). The MCC is a statistical metric that utilizes the number of true positive, false positive, false negative and true negative residues, giving a score between 1 and −1. A score of 1 indicates a prefect prediction and a score close to 0 indicates a random prediction. The MCC provides a good assessment statistic, because it heavily penalizes both over and under predictions and is appropriate for biased datasets, such as binding versus non-binding residues (Lopez *et al.*, 2009).

In order to assess binding residue prediction accuracy, the observed binding site residues must be defined. However, defining which residues are in contact with a ligand can often be subjective, particularly if we consider the inherent flexibility of protein backbones, side chains and many large ligands. The distances used to define residue–ligand contacts can be adjusted; nevertheless, once a cutoff has been set all 'non-binding' residues are treated as incorrect by the MCC score, regardless of their distance from the site.

The top methods in the function prediction category of CASP8 were methods by the Lee group (Oh *et al.*, 2009) and the Sternberg group (Wass and Sternberg, 2009). Both groups assessed their own predictions by two additional metrics: accuracy and coverage. However, these metrics also penalize close predictions to a similar extent as the MCC statistic (Oh *et al.*, 2009; Wass and Sternberg, 2009).

In this article we are proposing a simple new metric, the Binding-site Distance Test (BDT) score, which addresses the problems associated with the MCC while maintaining the advantages. The score is highly correlated with the MCC, and it appropriately penalizes both under and over predictions, while also considering the distance of predicted residues from the observed binding site.

2 METHODS

The BDT score was calculated by considering: the list of residue numbers in the protein predicted to be binding to a ligand, the list of residue numbers observed to be binding to a ligand, the PDB file of the observed structure (with residue numbering matching that of the sequence) and a distance threshold.

The Euclidean distance was calculated between each residue in the predicted set and each residue in the observed set. The distance was then converted to an *S*-score using the standard equation:

$$S_{ij} = \frac{1}{1+\left(\frac{d_{ij}}{d_0}\right)^2}$$

*To whom correspondence should be addressed.

© The Author 2010. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

Downloaded from http://bioinformatics.oxfordjournals.org/ at National Sun Yat-sen University on November 1, 2011

Table 1. Correlation coefficients for BDT scores versus MCC scores using the CASP8 data for binding site prediction

$d_0$value (Å)	Pearson's r	Spearman's ρ	Kendall's τ
1.0	0.966	0.928	0.764
2.0	0.963	0.914	0.745
3.0	0.955	0.892	0.717
5.0	0.922	0.848	0.663
7.0	0.882	0.810	0.619
9.0	0.839	0.778	0.583

Results for all predictions and all targets have been pooled. The BDT score has been calculated using different values for d_0, the optimal values are from 1 to 3 Å.

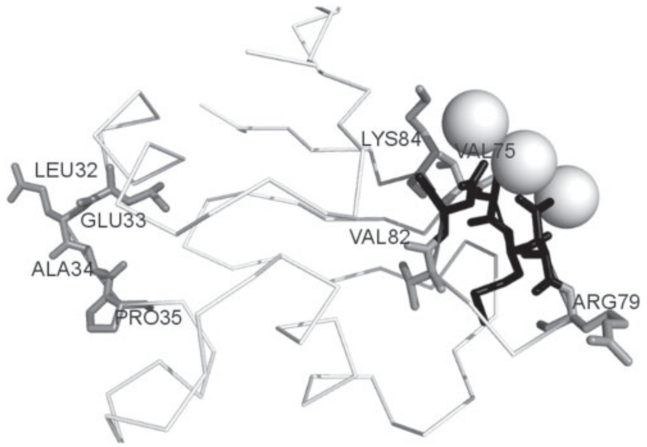

Fig. 1. Ribbon diagram of CASP8 target T0453. Hypothetical predicted residues are shown as grey sticks and the observed binding residues (76, 77, 78, 83) are shown as black sticks. For the first prediction on the left (32, 33, 34, 35) the MCC score is −0.046 and the BDT score is 0.017 (with $d_0 = 3$). For the second prediction on the right (75, 79, 82, 84), again the MCC score is −0.046; however, the BDT score is 0.384 (with $d_0 = 3$).

Where S_{ij} was the S-score between a predicted residue i and an observed residue j, d_{ij} was the Euclidean distance between the C-alpha coordinates of residues i and j and d_0 was a distance threshold (values between 1 and 3 Å are recommended, see Table 1). The maximum S_{ij} score, max(S_{ij}), was then determined for each predicted residue. The final BDT score was simply the sum of the maximum S_{ij} scores normalized by the greater value of the number of predicted residues (N_p) and the number of observed residues (N_o):

$$\text{BDT}=\frac{\sum_{i=1}^{N_p}\max(S_{ij})}{\max(N_p,N_o)}$$

3 RESULTS AND DISCUSSION

A potential problem with relying on the MCC is illustrated in Figure 1, where two hypothetical binding site predictions are shown for CASP8 target T0453. The prediction on the right hand side of the figure (75, 79, 82, 84) is closer to the observed binding site than the prediction shown on the left hand side of the image (32, 33, 34, 35); however, both predictions are assigned identical MCC scores (−0.046). Conversely using the BDT score with $d_0 = 3$, the prediction close to the site on the right is assigned a higher score (0.384) compared with that of the more distant prediction on the left (0.017). Using the MCC, all 'non-binding' residues in a prediction are considered equal, no matter how close they are to the actual site. Thus, small changes to the list of observed binding site residues can greatly affect the MCC score of close predictions. Further examples using real CASP8 predictions are shown in Supplementary Figure 2.

The BDT score ranges between 0 and 1, where perfect predictions achieve scores of 1 and distant predictions are assigned scores closer to 0. If we consider the flexibility of both ligands and proteins as well as the possibility of alternative ligands binding to the same site, the BDT score is a more appropriate score than the MCC. The BDT score takes into account the actual structure and distances between predicted and observed binding residues. Residues deemed false positives that are nevertheless close to the binding site score higher than distant predictions using the BDT score.

The distance threshold d_0 in the S_i score alters the range of BDT scores; however, BDT scores with different cutoffs are highly correlated with conserved ranking. The BDT scoring method maintains the penalty for over and under predictions, using the normalization max(N_p, N_o), it is appropriate for biased datasets and the scores are highly correlated with the MCC scores (Table 1, Supplementary Table 1), even though the metrics are conceptually different. There is an approximately linear dependence between the BDT scores at each cutoff and the MCC scores; however, the Spearman's ρ and Kendall's τ also show that the ranking of predictions is also maintained. The value for d_0 may be adjusted to vary the stringency of the score (Table 1). Outliers in plots of MCC scores versus BDT scores (Supplementary Fig. 1) are illustrated by the example in Figure 1.

Finally, the BDT score is relatively easy to calculate and because the actual PDB file is required for calculation there is no ambiguity concerning the missing residues (i.e. disordered regions) (for this article, all missing residues were also excluded from the calculation of MCC scores). Furthermore, the BDT score minimizes the penalty for ambiguous predicted residues that might be considered to be in the active site, or are considered to be in contact with an alternative ligand, but are nevertheless excluded from the observed subset (Supplementary Fig. 2B).

Funding: RCUK Academic Fellowship (to L.J.M.); University of Reading Faculty Studentship (to D.B.R.).

Conflict of Interest: none declared.

REFERENCES

Lopez,G. *et al.* (2007) Assessment of predictions submitted for the CASP7 function prediction category. *Proteins*, **69** (Suppl. 8), 165–174.

Lopez,G. *et al.* (2009) Assessment of ligand binding residue predictions in CASP8. *Proteins*, **77** (Suppl. 9), 138–146.

Matthews,B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.

Oh,M. *et al.* (2009) Protein-binding site prediction based on three-dimensional protein modeling. *Proteins*, **77** (Suppl. 9), 152–156.

Sankararaman,S. *et al.* (2010) Active site prediction using evolutionary and structural information. *Bioinformatics*, **26**, 617–624.

Soro,S. and Tramontano,A. (2005) The prediction of protein function at CASP6. *Proteins*, **61** (Suppl. 7), 201–213.

Wass,M.N. and Sternberg,M.J. (2009) Prediction of ligand binding sites using homologous structures and conservation at CASP8. *Proteins*, **77** (Suppl. 9), 147–151.

Downloaded from http://bioinformatics.oxfordjournals.org/ at National Sun Yat-sen University on November 1, 2011