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


It is advisable to refer to the publisher's version if you intend to cite from the work.

To link to this article DOI: http://dx.doi.org/10.1093/bioinformatics/btq543

Publisher: Oxford University Press

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the End User Agreement.

www.reading.ac.uk/centaur
The Binding Site Distance Test Score: A robust method for the  
assessment of predicted protein binding sites

Daniel B. Roche¹, Stuart J. Tetchner¹ and Liam J. McGuffin¹,*  
¹School of Biological Sciences, University of Reading, Whiteknights, Reading RG6 6AS, UK

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

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 develop-  
ers of new methods and by the assessors for the community wide  
prediction experiment – CASP8. Whilst 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 af-  
fected 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, rela-  
ting 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 us-  
ing both the MCC and BDT methods and the scores were com-  
pared. The BDT was found to strongly correlate with the MCC  
scores whilst also being less susceptible to the subjectivity of defin-  
ing 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: lj.mcguffin@reading.ac.uk

1 INTRODUCTION
The prediction of a protein’s ligand binding site location and poten-  
tial 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 catego-  
ry 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 as-  
sessing these terms, the CASP7 (Lopez, et al., 2007) assessors  
decided that CASP was not the best place for this format of func-  
tion prediction. Thus, for CASP8, function prediction was included  
in a different format, with the assessment of observed ligand bind-  
ing 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 Mat-  
thews 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 data sets, such  
as binding versus non-binding residues (Lopez, et al., 2009).

In order to assess binding residue prediction accuracy, the ob-  
served 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 back-  
bones, side chains and many large ligands. The distances used to  
define residue-ligand contacts can be adjusted, nevertheless, once a  
cut-off has been set all “non-binding” residues are treated as incor-  
crect 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 Stern-  
berg 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 paper we are proposing a simple new metric, the Binding-  
site Distance Test (BDT) score, which addresses the problems  
associated with the MCC whilst maintaining the advantages. The  
score is highly correlated with the MCC, it appropriately penalizes  
both under and over predictions, whilst also considering the dis-  
tance 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 struc-  
ture (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 con-  
verted to an S-score using the standard equation:

© Oxford University Press 2005
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 \( d_0 = 3 \)).

\[ S_{ij} = \frac{1}{1 + \left( \frac{d_{ij}^2}{\pi} \right) ^2} \]

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 \( S_i \), score, \( \max(S_i) \), was then determined for each predicted residue. The final BDT score was simply the sum of the maximum \( S_i \) scores normalized by the greater value of the number of predicted residues (\( N_p \)) and the number of observed residues (\( N_o \)).

\[ BDT = \frac{\sum_{i=1}^{N_p} \max(S_i)}{\max(N_p,N_o)} \]

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 2B.

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 score alters the range of BDT scores; however BDT scores with different cut-offs are highly correlated with conserved ranking. The BDT scoring method maintains the penalty for over and under predictions, using the normalisation \( \max(N_p,N_o) \), it is appropriate for biased data sets 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 \( p \) and Kendall’s \( \tau \) 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 Figure 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 missing residues (i.e. disordered regions) (for this paper 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 Figure 2B).

<table>
<thead>
<tr>
<th>( d_0 ) value (Å)</th>
<th>Pearson’s ( r )</th>
<th>Spearman’s ( p )</th>
<th>Kendall’s ( \tau )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>0.966</td>
<td>0.928</td>
<td>0.764</td>
</tr>
<tr>
<td>2.0</td>
<td>0.963</td>
<td>0.914</td>
<td>0.745</td>
</tr>
<tr>
<td>3.0</td>
<td>0.955</td>
<td>0.892</td>
<td>0.717</td>
</tr>
<tr>
<td>5.0</td>
<td>0.922</td>
<td>0.840</td>
<td>0.663</td>
</tr>
<tr>
<td>7.0</td>
<td>0.882</td>
<td>0.810</td>
<td>0.619</td>
</tr>
<tr>
<td>9.0</td>
<td>0.839</td>
<td>0.778</td>
<td>0.583</td>
</tr>
</tbody>
</table>

ACKNOWLEDGEMENTS

Funding: An RCUK Academic Fellowship (LJM) and a University of Reading Faculty Studentship (DBR).

REFERENCES