HTHquery: a method for detecting DNA-binding proteins with a helix-turn-helix structural motif

C. Ferrer-Costa¹, H. P. Shanahan²,*, S. Jones³ and J. M. Thornton⁴

¹Molecular Modelling and Bioinformatics, IRBB-Parc Cientific de Barcelona, UB, Josep Samitier, 1-5 08028 Barcelona, Catalonia, Spain, ²Department of Computer Science, Royal Holloway, University of London, Egham, Surrey, TW20 0EX, UK, ³Department of Biochemistry, School of Life Sciences, John Maynard Smith Building, University of Sussex, Falmer, Brighton, BN1 9QG, UK and ⁴EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

Received on May 23, 2005; revised on June 16, 2005; accepted on July 6, 2005
Advance Access publication July 19, 2005

ABSTRACT
Summary: HTHquery is a web-based service to determine if a protein structure has a helix-turn-helix structural motif which could bind to DNA. It is based on a similarity with a set of structural templates, the accessibility of a putative structural motif and a positive electrostatic potential in the neighbourhood of a putative binding region. A set of scores are computed, based on each template, using a linear predictor. From the training set used, the predictor has a true positive rate of 83.5% and a false positive rate of 0.8%. The emphasis for the website is on providing a straightforward interface which can be easily used by a bench-based scientist.
Availability: HTHquery is implemented using a set of Perl scripts and C program and can be accessed freely on the website http://www.ebi.ac.uk/thornton-srv/databases/HTHquery
Contact: Hugh.Shanahan@physics.org

1 INTRODUCTION
Robust methods to detect DNA-binding proteins from structures of unknown function are important for structural biology. This is not a trivial task given the submission of over 300 non-redundant entries in the Protein Data Bank (PDB) by Structural Genomics Initiatives (Todd et al., 2005) and that it has been estimated that 6–7% of the genes in a eukaryotic genome have a DNA-binding function (Luscombe et al., 2001). In the recent past, we have focussed on identifying a particular class of DNA-binding proteins which have a helix-turn-helix (HTH) structural motif. This is a relatively abundant class of proteins; approximately one third of known DNA-binding protein structural families have this motif. Initially the search was based on locating contiguous fragments of Cα positions within a protein that are similar to a set of templates constructed from a set of known DNA-binding proteins with an HTH motif (Jones et al., 2003). This was demonstrated to be a quite powerful technique, as the HTH motif is an example of convergent evolution, so a template can detect DNA-binding proteins with HTH that do not exhibit any sequential similarity. The addition of electrostatics improved the accuracy of this approach substantially (Shanahan et al., 2004).

The relevant observables in this calculation (similarity to a template, accessibility of a putative binding region, electrostatic potential in the neighbourhood of a putative binding region) were computed for a set of 79 known DNA-binding protein chains with an HTH motif and 490 protein chains from the PDB that do not bind to DNA and have an RMSD with a structural template that is <2.5 Å. The protein chains were chosen to be non-redundant by using the CATH structural database (Pearl et al., 2005). In particular, no protein chain shares the same six digit CATH code (i.e. the sequence similarity is <95%). A linear predictor was constructed which had a true positive rate of 83.5% and a false positive rate of 0.8%. A projection of the data in two of the most discriminating observables can be found in Figure 1.

These methods have been implemented in the website HTHquery. The goal of HTHquery is to test if a particular structures of a protein has an HTH motif which could bind to DNA. The web interface can be run easily and simply with default parameter settings and can be used by anyone with a valid protein structure in a PDB file format. No specialist bioinformatics knowledge is required.

2 DESCRIPTION
The website is designed for structural biologists to submit protein structures and test if the structure has an HTH motif which can bind to DNA. The interface is extremely simple, in that the user submits a single protein structure in a PDB format, or can simply enter a PDB code. The user can then select the set of protein chains that he wishes to examine. By default all are examined. If the structure already contains nucleic acid chains, then the user is notified of this and they are not examined.

A putative HTH motif is determined for each protein chain by scanning the protein chain with a set of seven structural templates and determining the region of the chain with the smallest RMSD. This is computed using the Kabsch algorithm (Kabsch, 1976). For each template, the region with the minimum RMSD is stored. The accessible surface area for each putative motif is then computed using NACCESS (Hubbard and Thornton, 1993) and the electrostatic motif score (EMS) is computed using the methods described by (Shanahan et al., 2004). These values are then passed through the linear predictor described above.

The linear predictor effectively divides the 3D parameter space into two regions, which we refer to as ‘true’ and ‘false’ separated...
by a 2D plane which maximizes the number of proteins which do or
do not have a DNA-binding HTH motif respectively. The predictor
returns a score which is proportional to $1/[1 + \exp(-d)]$, where
$d$ is the distance from this plane ($d$ is positive if it lies in the true
region). A score ranges from 9 to $-9$, where 9 indicates it is deep in
the true region and $-9$ if it is deep in the false region. All template
searches which return a score that is $>3$ is flagged as being a definite
hit and listed on the main results page. The results for the above
observables are listed from each template search and it is possible
to launch a Rasmd window (Sayle and Milner-White, 1995) which
will either display the backbone of the region of the protein which
is taken to be as a DNA-binding protein with an HTH motif or a
space-filled representation showing the electrostatic score for each
solvent accessible atom. Each of the terms in the results page are
copiously annotated with help pages. There are links to separate
web pages which list all template searches where the linear predictor
score lies between $-3$ and $3$ (a marginal hit) and $<-3$ (where we
are confident there is no DNA-binding HTH motif). We stress that
the linear predictor score simply reflects the distance from the plane
that best divides the observed data and does not reflect a statistical
confidence.

The web pages are written with HTML and Javascript. The code for
computing the observables, computing the linear predictor score and
outputting the results uses a combination of Perl and C subroutines
(used for the computationally expensive calculations). A direct inter-
face between Perl and C is provided using the Swig library (Beazley
electrostatic potential is computed using Delphi (Rocchia et al.,
2001). The electrostatic calculation is computed using a simplified
charge set, and as a result it is not necessary to protonate the structure
beforehand. Details of the electrostatic calculation can be found in
Shanahan et al. (2004). Although the calculation of the electrostatic
potential has been simplified such that it is not necessary to protonate
the structure, it is not recommended to submit protein structures with
main or side chain heavy atoms missing.

Calculations for a typical protein chain usually takes the order of
minutes. There are help pages to explain HTHquery in detail, includ-
ing a specific example to work through. Structures with multiple sets
of coordinates, i.e. those generated from NMR studies, cannot be
dealt with and it is recommended that an averaged set of positions
be submitted instead. However, modelled structures with all heavy
protein atoms included may be submitted.

The interface is designed for submission of one structure at a time,
and batch submissions cannot be made. Requests for batch submit-
sions should be directed to the authors. HTHquery is not a part
of a specific functional genomics pipeline; however requests for a
SOAP (Gudgin et al., 2003, http://www.w3.org/TR/soap12-part1/), or
similar interface will be considered if the authors are contacted.

ACKNOWLEDGEMENTS

H.P.S. was supported by an MRC/PPARC training fellowship and
S.J. was supported by a US Department of Energy grant (DE-FG02-
96ER62166).

Conflict of Interest: none declared.

REFERENCES

O’Reilly Perl Conference 2.0, Aug. 17-20, San Jose, CA.
Jones,S. et al. (2003) Using structural motif templates to identify proteins with dna
Pearl,F. et al. (2005) The CATH Domain Structure Database and related resources
Gene3D and DHS provide comprehensive domain family information for genome
Rocchia,W. et al. (2001) Extending the applicability of the nonlinear poisson-boltzmann
equation: Multiple dielectric constants and multivalent ions. J. Phys. Chem. B, 105,
6507–6514.
Sayle,R.A. and Milner-White,E.J. (1995) RasMol: biomolecular graphics for all. TIBS,
20, 374–376.
Todd,A.E. et al. (2005) Progress of structural genomics initiatives: an analysis of solved

Fig. 1. Electrostatic motif score against RMSD for the true (i.e. those
proteins with a DNA-binding HTH motif) and false sets used to train the
linear predictor. A number of true proteins have an RMSD that is $>2.5\AA$.