Research Square Preprint: Please note that this article has not completed peer review.

AFPredictor: a computational screening protocol for antifreeze/ice-structuring proteins

CURRENT STATUS: POSTED

Andrew C. Doxey
University of Waterloo

Brendan J. McConkey
University of Waterloo

DOI:
10.1038/nprot.2006.213

SUBJECT AREAS
Biochemistry *Computational biology and bioinformatics*

KEYWORDS
antifreeze protein, ice-structuring protein, ice-binding surface, algorithm, surface feature recognition, structural bioinformatics

Introduction

Antifreeze/ice-structuring proteins (AFPs, ISPs) have evolved independently in a wide range of organisms including bacteria, plants and fish. Despite their independent origins and diverse folds, AFPs bind to a common substrate (ice) albeit on different surfaces and in different orientations [1]. In theory, AFPs should possess common structural characteristics responsible for ice-binding activity, independent of protein fold. This motivated the development of a surface-based pattern detection algorithm, AFPredictor, to analyze and rank structural characteristics of AFPs [2].

Using AFPredictor, it was demonstrated that ‘ordered surface carbons’ (OSCs) are a distinguishing feature of AFPs and, more specifically, their ice-binding surfaces [2]. AFPredictor identified AFPs from within a large set of structures with greater than 99% specificity. Furthermore, it was used to identify a novel ice-binding protein by screening a library of homology modeled structures based on cDNA sequences obtained from cold-acclimated winter rye (Secale cereale).

AFPredictor is freely available on request from the authors. The parameters (residue types, cutoff values, etc.) are fully modifiable. The following protocol describes how to use AFPredictor to detect AFPs and their ice-binding residues, and score results against a background non-redundant dataset.

Equipment

1. A personal computer (Macintosh, Unix/Linux or Windows operating system)
2. Input structures in PDB file format, from experimental data or structural models
3. Local installation of the Perl programming language (available from http://www.perl.org)
4. AFPredictor program and Vsurface program module [3].

Procedure

Applying the algorithm to a set of input structures

1. If not already installed, download and install the Perl programming language (http://www.perl.org).
2. Prepare a set of input structures in Brookhaven PDB file format. Structures may be downloaded from the Protein Data Bank (http://www.rcsb.org). X-ray structures are preferred over NMR

structures due to increased accuracy of sidechain positions. If starting with a library of sequences, generate a set of 3D homology-models using comparative-modeling software, such as SWISS-MODEL (http://swissmodel.expasy.org/SWISS-MODEL.html, Modeller (http://www.salilab.org/modeller/), etc. NOTE: The performance of the algorithm is partly dependent on the resolution of the input structures, and better results will be obtained using high quality structural data or homology models. Homology models should be constructed with high quality templates (e.g. X-ray structures) and checked for errors using software such as WHATCHECK [4].

3. Run AFPredictor on the input structures. Unless specified otherwise, the program will use default values for the parameters. The usage for AFPredictor is shown in Fig. 1. If the option –d is specified, the # OSCs, FASA and TASA scores (see [2] for a description) will be printed as the program runs and output in the file (output.txt). If the option –f is specified, additional information will be output including the residues and atoms corresponding to predicted OSCs as well as their solvent-accessibility as calculated by Vsurface [3]. Option –o will highlight predicted OSCs in an output PDB file. Atoms are highlighted in the B-factor column of the PDB file and can be viewed using molecular visualization software. For example, in Deep View (Swiss-PDB viewer) (http://www.expasy.org/spdbv/), the AFP output PDB file can be loaded, a molecular surface calculated, and the molecular surface layer colored by B-factor. Using the default Swiss-PDB color scheme, OSCs will appear red and all other atoms will appear blue (Fig. 2). Command-line use of AFPredictor is shown below:

perl AFPredictor.pl [-Options] {-f <PDB file> or –d <directory>}

Applying the algorithm to a background dataset (optional)

At this point, the user might wish to rank the scored test structures from step 3 against a background dataset such as a non-redundant set of PDB files. If default parameters have been used, the user may compare with the previous results compiled in the file 2006results.txt. In this case, steps 4-7 may be omitted. However, if any parameters used in step 3 have been changed from the default values or the user wishes to use a different data set for comparison, proceed to step 4.

4. Create a list of non-redundant protein sequences for comparison. This may be done using the PISCES server [5] accessible at http://dunbrack.fccc.edu/PISCES.php. It is important that redundant sequences/structures be removed from the background dataset because multiple instances of the same structure may statistically skew the final results. Ensure that incomplete (e.g. alpha-carbon only) PDB files are removed in the culling procedure.
5. Retrieve the corresponding structures from the PDB and place all files in the same directory (e.g. /nonredundantPDB/). All files should be in PDB format and have the extension .pdb. A PDB web download utility is available at http://www.rcsb.org/pdb/static.do?p=download/web_download/download.jsp.

NOTE: If the PDB was culled 'by chain' instead of 'by entry' in step 4, retrieve only the specified chain for each PDB file. If this approach is used, the algorithm may analyze surfaces that are internal in the native structure. If this is unwanted, cull the PDB 'by entry' (in step 4), obtain and use a monomeric structure database, or process PDB files through the PQS server (http://pqs.ebi.ac.uk/).

6. (Optional) Remove hydrogen atoms from all PDB files. Doing so will improve the overall runtime of the algorithm, especially if the structural database is large (>500 structures).
7. Run the algorithm on the non-redundant PDB directory (specify the parameter –d). An example command (using default parameters) is shown below.

```
perl AFPredictor.pl -d /nonredundantPDB/
```

8. Sort the results from step 3 and step 7 by FASA or TASA. This results in a final FASA or TASA ranking. A percentile score for each structure can be calculated as: 1-ranking / # structures. This should not be interpreted as the strength of ice-binding activity or likelihood that a structure is an ice-binding protein, but as a relative measure of the surface area occupied by predicted OSCs.

Critical Steps

Step 3: In choosing parameter values, it is important to consider the resolution of input structures. The value for VTHRESH (option -v) controls the stringency of vector-matching and should be relaxed

(increased) for lower quality input structures. Note that this will increase the rate of false positive identifications as well. The choices for parameter values might also depend on the particular scenario in which AFPredictor is being used. For example, a simple use of AFPredictor is to predict the putative ice-binding surface of a protein already known to have ice-binding activity. In this case, the user might run the program on the protein structure with relaxed parameters to increase the probability of detecting all contributing OSCs, at the expense of an increased false positive rate. Subsequent iterations could increase the stringency for more precise predictions. For screening a large structural database for potentially highly active ice-binding proteins, the parameters can be set to a high stringency to reduce the number of false positives (though some true positives may also be filtered out in the process).

References

1. Jia, Z. & Davies, P.L. Antifreeze proteins: an unusual receptor-ligand interaction. *Trends Biochem. Sci.* **27**, 101-106 (2002).
2. Doxey, A.C., Yaish, M.W., Griffith, M. & McConkey, B.J. Ordered surface carbons distinguish antifreeze proteins and their ice-binding regions. *Nat. Biotechnol.* **24**, 852-855 (2006).
3. McConkey, B.J., Sobolev, V. & Edelman, M. Quantification of protein surfaces, volumes and atom-atom contacts using a constrained Voronoi procedure. *Bioinformatics* **18**, 1365-1373 (2002).
4. Hooft, R.W., Vriend, G., Sander, C. & Abola, E.E. Errors in protein structures. *Nature* **381**, 272 (1996).
5. Wang, G. & Dunbrack, R.L., Jr. PISCES: a protein sequence culling server. *Bioinformatics* **19**, 1589-1591 (2003).

Acknowledgements

This work was supported by grants from the National Science and Engineering Research Council (NSERC) to B.J.M., and an NSERC Canada Graduate Scholarship held by A.C.D.

Figures

```
Usage: perl AFPredictor.pl [-Options] {-f <PDB file> or -d
<directory>}

e.g. AFPredictor -t 15 -r 7.66 -q 4.5 –v 1.13 -x KRHFW -f
lwfa.pdb -o results.pdb
```

-t <value>	Solvent-accessibility (SAS) threshold to define surface carbons. Default = 15.1 square angstroms.
-r <value>	Maximum length of allowed vectors (Vmax). Default = 7.66 angstroms.
-q <value>	Minimum length of allowed vectors (Vmin). Default = 4.5 angstroms.
-v <value>	Vector-matching cutoff (V_{THRESH}). ‖V1-V2‖ <= Vthresh. Default = 1.09 angstroms.
-c <value>	Distance threshold for clustering of OSCs. Default = 7.66 angstroms.
-p <value>	Polar atom distance cutoff. OSCs must be within this distance of a polar atom. Default = 4 angstroms.
-h <value>	SAS threshold for polar atoms. Default = 0 square angstroms.
-x <list>	List of omitted residue types (e.g. KHD) in single-letter amino acid format. Default = RDEHKFWY
-o <file>	This option will output a PDB file <file> in which OSCs have been marked. The OSCs are marked by setting the beta-factor score to 100. All other beta-factors are set to 0.
-d <directory>	The algorithm will be run on all (*.pdb) files in the specified directory.
-f <file>	The algorithm will be applied only to the specified PDB file.

Figure 1

Usage of AFPredictor with description of options and default parameters.

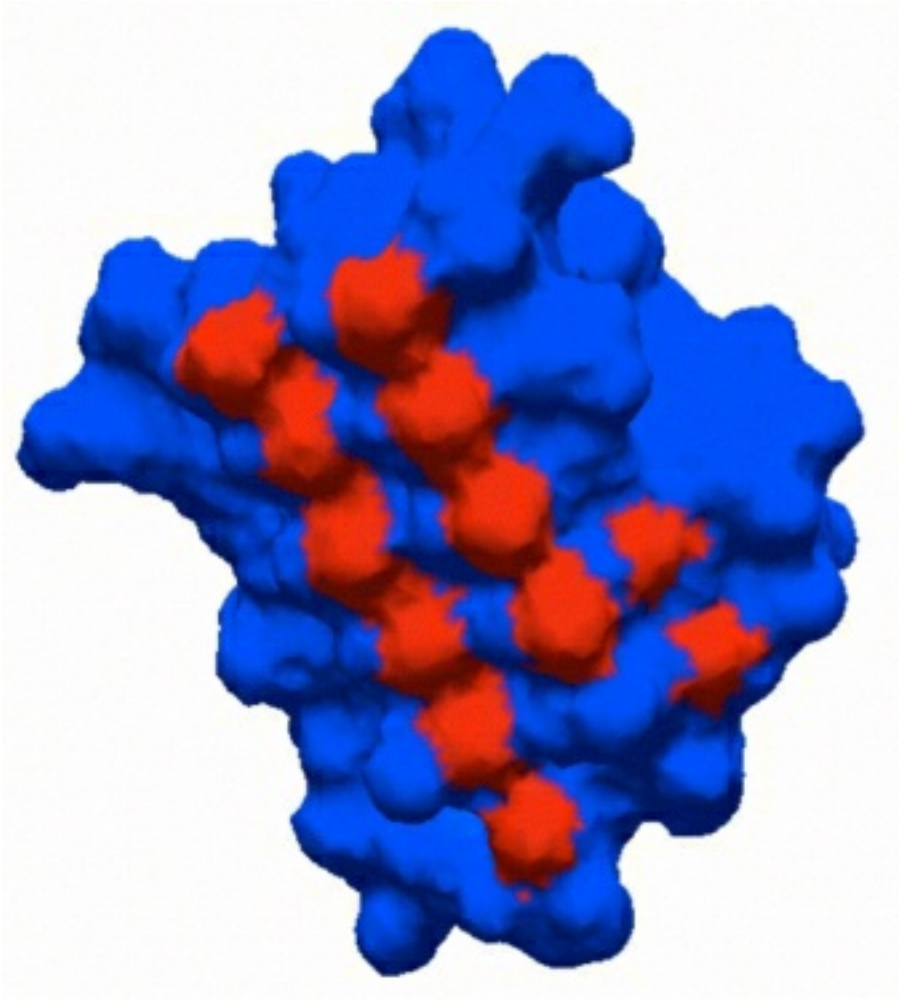

Figure 2

Sample AFPredictor PDB output file showing ordered surface carbons, visualized using DeepView spdbv 3.7. The structure shown is a beta-helical antifreeze protein from spruce budworm, PDB ID 1LOS.

Ordered surface carbons distinguish antifreeze proteins and their ice-binding regions

by Doxey, A.C. et al.

Nature Biotechnology (14 August, 2006)