W160–W163 *Nucleic Acids Research, 2005, Vol. 33, Web Server issue*
doi:10.1093/nar/gki381

Scooby-domain: prediction of globular domains in protein sequence

Richard A. George[1,2], Kuang Lin[3] and Jaap Heringa[4,*]

[1]Inpharmatica Ltd, 60 Charlotte Street, London W1T 2NU UK, [2]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK, [3]Division of Mathematical Biology, National Institute for Medical Research, The Ridgeway, Mill Hill NW7 1AA, UK and [4]Centre for Integrative Bioinformatics (IBIVU), Faculty of Sciences and Faculty of Earth and Life Sciences, Vrije Universiteit, De Boelelaan 1081a, 1081HV Amsterdam, The Netherlands

Received February 14, 2005; Revised and Accepted March 8, 2005

ABSTRACT

Scooby-domain (sequence hydrophobicity predicts domains) is a fast and simple method to identify globular domains in protein sequence, based on the observed lengths and hydrophobicities of domains from proteins with known tertiary structure. The prediction method successfully identifies sequence regions that will form a globular structure and those that are likely to be unstructured. The method does not rely on homology searches and, therefore, can identify previously unknown domains for structural elucidation. Scooby-domain is available as a Java applet at http://ibivu.cs.vu.nl/programs/scoobywww. It may be used to visualize local properties within a protein sequence, such as average hydrophobicity, secondary structure propensity and domain boundaries, as well as being a method for fast domain assignment of large sequence sets.

INTRODUCTION

The suggestion that there might be a relationship between the ratio of hydrophobic and hydrophilic residues and molecular structure was first noted by Waugh (1) and later observed by Fisher (2). The globular structure of a protein cannot be achieved by any combination of amino acids, as certain principles of structure must be obeyed. Proteins with too many hydrophobic residues will aggregate in solution: a small polypeptide cannot tolerate >30% of hydrophobic residues (3). Furthermore, a largely hydrophilic protein will fail to form a stable hydrophobic core (4).

Long polypeptides will fold into compact, semi-independent, structural units called domains (5). Given the observed random distribution of hydrophobic residues in proteins (6), domain formation appears to be the optimal solution for a large protein to bury its hydrophobic residues while keeping hydrophilic residues at the surface (7). Consequently, there are no observed protein structures of >250 residues that contain a single hydrophobic core (8).

Methods to correctly define domains in protein sequence are extremely important in many areas of biology. Successful domain delineation would enable: the correct design of soluble constructs for high throughput structural genomics, the design of site directed mutagenesis experiments, the optimization of secondary structure prediction and threading methods; and comparative sequence analysis (9).

The Scooby-domain (sequence hydrophobicity predicts domains) algorithm identifies the location of domains in a protein query sequence based on the distribution of observed lengths and hydrophobicities in domains with known 3D structure. Scooby-domain uses a multilevel smoothing window to average the hydrophobic content of domain-sized regions in a sequence of unknown structure. Using the window length and average hydrophobicity, the probability that the region can fold into a domain is then calculated and regions that are likely to be unstructured are also identified. Scooby-domain is available as a Java applet that can be used to visualize local properties of a protein sequence, such as average hydrophobicity and secondary structure propensity, as well as being a tool to manually assign domain boundaries. A web server is also available to automatically assign domain boundaries to a query sequence.

METHODS

Multilevel smoothing window

Hydrophobicity plots were designed to display the distribution of hydrophobic and hydrophilic residues along a protein sequence and are useful to identify transmembrane regions or antigenic sites (10,11). To generate a hydrophobicity plot for a sequence each residue is first assigned a value of

*To whom correspondence should be addressed. Tel: +31 2059 87649; Fax: +31 2059 87653; Email: heringa@cs.vu.nl

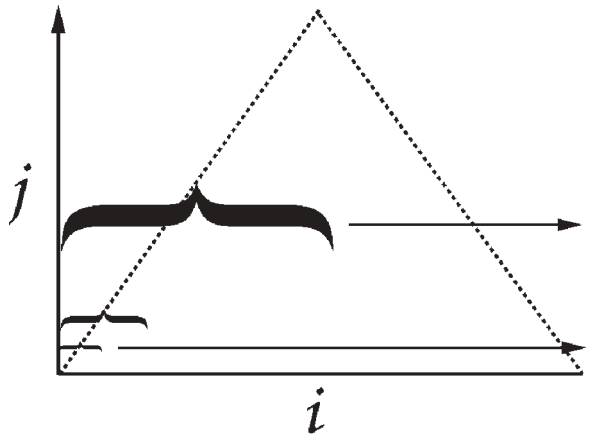

Figure 1. Multilevel smoothing window. Each smoothing window sums the properties of the residues it encapsulates along a sequence, and places the value at its central position. This leads to a 2D matrix, where the value at cell (i, j) is the average property encapsulated by a window of size j that is centred at residue position i. The matrix has a triangular shape, the apex of which will correspond to a window size equal to the length of the sequence or the maximum window size.

hydrophobicity, and then a smoothing window, of a given size, is scanned along the sequence. Starting at the N-terminus of the sequence, the average hydrophobicity of the amino acids encapsulated by the window is calculated and the value is plotted at the centre of the window. The window then moves along the sequence, one residue at a time, calculating the average hydrophobicity of the residues it encapsulates until the window reaches the end of the sequence. A window size of 19 residues is useful for identifying transmembrane regions and a window size of 7 is useful to identify surface regions.

Scooby-domain applies a multilevel smoothing window to visualize properties of amino acids in a sequence, which means that it uses windows of increasing size, starting from a length of three residues. Each smoothing window calculates the average property of the residues it encapsulates, placing the value at its central position. This leads to a 2D matrix, where the value at cell (i, j) is the average hydrophobicity encapsulated by a window of size j that is centred at residue position i. The matrix has a triangular shape, the apex of which will correspond to a window size equal to the length of the sequence or the maximum window size (Figure 1). This adds a new dimension to the traditional hydrophobicity plot and is useful for identifying both local and global properties for a protein sequence.

Generating a domain probability matrix for a query sequence

Scooby-domain uses the multilevel smoothing window to predict the location of domains in a query sequence. A window size, representing the length of a putative domain, is incremented starting from the smallest domain size observed in the database to the largest domain size. Based on the window length and its average hydrophobicity, the probability that it can fold into a domain is found directly from the distribution of domain size and hydrophobicity, calculated using the S-level domain representatives from the CATH domain database (12). For each domain, percentage hydrophobicity is calculated using a binary hydrophobicity scale, where 11 amino acid types are considered as hydrophobic: Ala, Cys, Phe, Gly, Ile, Leu, Met, Pro, Val, Trp and Tyr (6). Visualization of the Scooby-domain probability matrix for a sequence can be used to effectively identify regions that are likely to fold into domains or are likely to be unstructured (Figure 2).

Automatic domain boundary assignment

The Scooby-domain web server performs fast, automatic, domain annotation by identifying the most domain-like regions in the query sequence. The highest probability in the domain probability matrix represents the first predicted domain. The corresponding stretch of sequence for this domain is removed from the sequence. Therefore, the first predicted domain will always have a continuous sequence and further domain predictions can encompass discontinuous domains. If the excised domain is at a central position in the sequence, the resulting N- and C-termini fragments are rejoined and the probability matrix recalculated as before. The second highest probability is then found and the corresponding sub-sequence removed.

AVAILABILITY

Scooby-domain Java applet

The Scooby-domain algorithm is available as a Java applet (http://ibivu.cs.vu.nl/programs/scoobywww) to both visualize a domain probability matrix and to analyse local sequence features for a query sequence (Figure 2). Once a query sequence has been entered, several options are available from the drop-down menu.

- 'Domain': creates a domain probability matrix that can be used to visually assign domains.
- 'HP Runs': first identifies non-random regions in a sequence, e.g. long stretches of hydrophilic residues, that are unlikely to form a globular structure and then creates a domain probability matrix.
- 'Binary HP': plots the average hydrophobicity for each smoothing window using the binary assignment described above.
- 'Eisenberg': plots the average hydrophobicity for each smoothing window using the residue hydrophobicity scale described by Eisenberg (13).
- 'Alpha': plots the propensity for a sequence region to be α-helical based on the residue propensities described by Chou and Fasman (14).
- 'Beta': plots the propensity for a sequence region to be a β-strand based on the residue propensities described by Chou and Fasman (14).
- 'Coil': plots the propensity for a sequence region to be in a random coil conformation based on the residue propensities described by Chou and Fasman (14).
- 'Linker': plots the propensity of a sequence region to be an inter-domain linker based on the linker propensities described in the linker database (15).
- 'Walkies': plots a random walk representation of the query sequence. A random walk moves up for every hydrophobic residue and down for every hydrophilic residue encountered as it moves along the sequence.

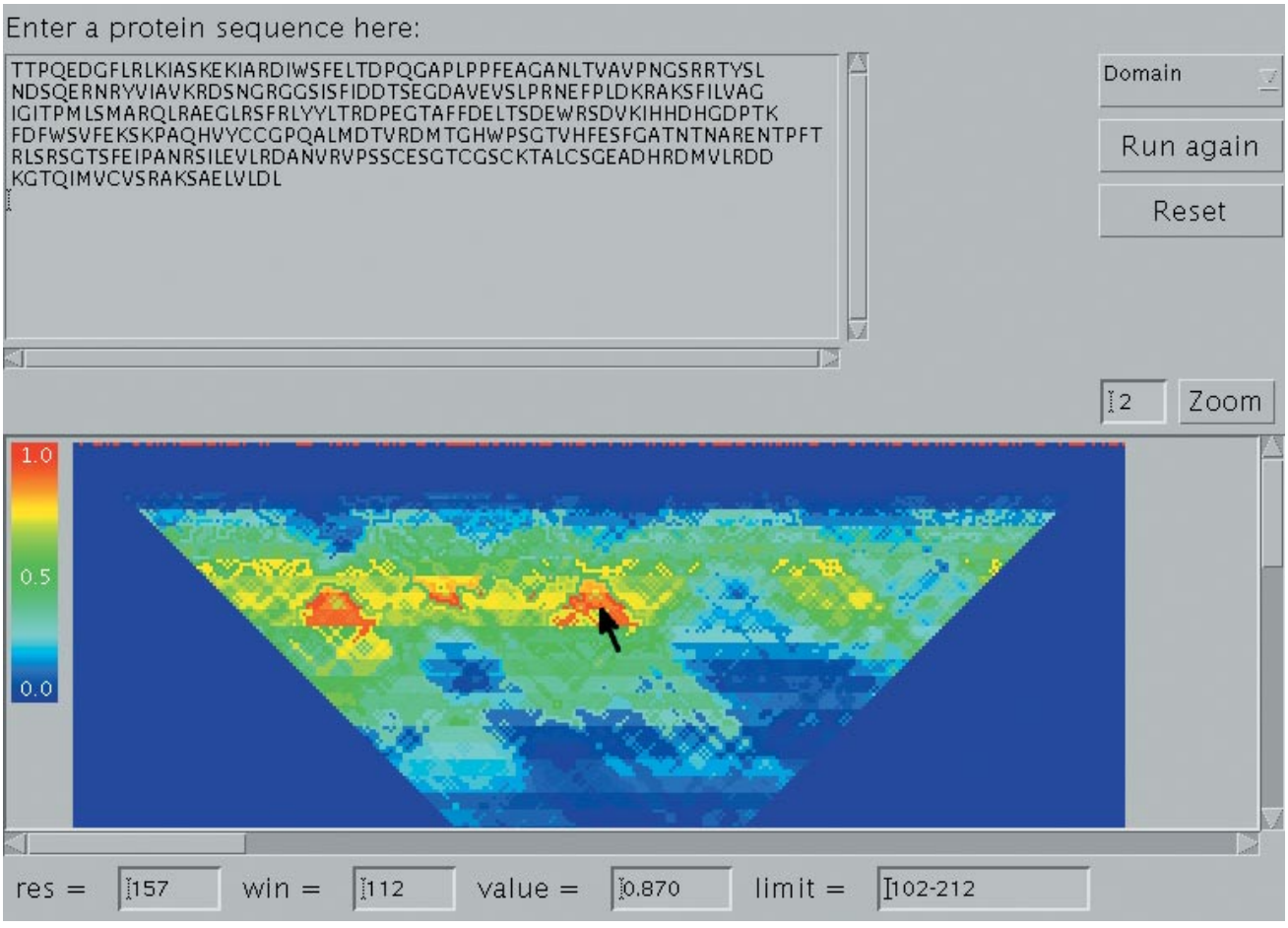

Figure 2. Scooby-domain Java applet. The Scooby-domain plot shows the probability for regions in a sequence to fold into a domain. Average hydrophobicities are converted into domain probability scores by referring to the observed distribution of domain sizes and hydrophobicities, i.e. given an average hydrophobicity and window length the probability that it can fold into a domain is found directly from the observed data. Regions coloured red represent the centre of a putative domain. The mouse-pointer highlights a likely domain between residues 102 and 212. Please note that the multilevel smoothing window is plotted upside-down when compared to Figure 1.

All calculated probabilities and propensities are normalized to values between zero and one, and are identified using a colour gradient of blue to red. Therefore, a region with a high probability of domain occurrence will appear red in the 'Domain' probability plot, while a region of low probability will appear blue. Using the mouse pointer, it is possible to identify the residue position, the window length being used, the probability value at that position, and the start and end points (limits) of the window. The limits will represent the domain boundaries when analysing a domain probability plot.

Scooby-domain prediction server

The option to automatically delineate domains for a query sequence is also available. The prediction server is simple to use, requiring a single query sequence and a few parameters. The 'N- and C-termini weighting' option can be used to encourage the domain-cutting algorithm to begin domain assignments at the start or end of the query sequence. The option to include external domain data, such as the boundary predictions from Domination (9), is also available. The additional boundary information can be fed directly into the domain probability matrix to help improve predictions. Domain predictions for a query sequence are presented with the highest scoring prediction first, out of 10 possible results. A GIF image of the domain probability matrix is displayed below the predictions and is available for download as a postscript image file.

The Scooby-domain prediction algorithm performed relatively well on a test set of 193 multidomain proteins, correctly predicting the location of over half (113/224) of the domain boundaries within an error of ±20 residues. The predictions are, for simple cases, accurate. But errors in domain boundary prediction are expected due to the intrinsic simplicity of our method. Further improvements can be achieved by searching the domain databases (16) and adding the results from other domain prediction methods. Using the option to include predictions made by Domination (9) improved the Scooby-domain performance to 60%.

CONCLUSIONS

The multilevel smoothing window applied by Scooby-domain is a useful tool to visualize local and global properties of amino acids in a protein sequence. The Java applet can also be used to

identify if a region in a protein can form a globular structure or is likely to be unstructured.

Methods to predict the location of domains are extremely important. Percentage hydrophobicity and domain size is a good predictor of domain location and has been applied to predict domain boundaries by the Scooby-domain algorithm. Domain prediction algorithms that have utilized the hydrophobicity of proteins, such as SnapDragon (7), and the constraints of domain size, such as Domain Guess by Size (17), have shown some success but are limited to small proteins, often with only two or three domains. Scooby-domain can quickly locate domains in a protein sequence, regardless of its length.

Precise identification of domain boundaries is a very difficult problem. Here we have presented a simple method which shows promising results. However, predictions are not accurate enough to be exclusively reliable. It is advised that other methods are used in combination with the Scooby-domain algorithm such as those methods that rely on homology searches of the domain databases. The Scooby-domain algorithm is flexible, in that it can accept boundary predictions from other sources to improve its prediction success.

ACKNOWLEDGEMENTS

The authors thank Victor Simossis and Jens Kleinjung for testing the web server. The Open Access publication charges for this article were waived by Oxford University Press.

Conflict of interest statement. None declared.

REFERENCES

1. Waugh,D.F. (1954) Protein–protein interactions. *Adv. Protein Chem.*, **9**, 325–437.
2. Fisher,H.F. (1964) A limiting law relating the size and shape of protein molecules to their composition. *Biochemistry*, **51**, 1285–1291.
3. Van Holde,K.E. (1966) The molecular architecture of multichain proteins. *Molecular Architecture in Cell Physiology*. Society of General Physiologists, pp. 81–96.
4. Dill,K.A. (1985) Theory for the folding and stability of globular proteins. *Biochemistry*, **24**, 1501–1509.
5. Richardson,J.S. (1981) The anatomy and taxonomy of protein structure. *Adv. Protein Chem.*, **34**, 167–339.
6. White,S.H. and Jacobs,R.E. (1990) Statistical distribution of hydrophobic residues along the length of protein chains. Implications for protein folding and evolution. *Biophys. J.*, **57**, 911–921.
7. George,R.A. and Heringa,J. (2002) SnapDRAGON: a method to delineate protein structural domains from sequence data. *J. Mol. Biol.*, **316**, 839–851.
8. Garel,J. (1992) Folding of large proteins: multidomain and multisubunit proteins. In Creighton,T. (ed.), *Protein Folding*. 1st edn. W.H. Freeman and Company, New York, pp. 405–454.
9. George,R.A. and Heringa,J. (2002) Protein domain identification and improved sequence similarity searching using PSI-BLAST. *Proteins*, **48**, 672–681.
10. Hopp,T.P. and Woods,K.R. (1981) Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl Acad. Sci. USA*, **78**, 3824–3828.
11. Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105–132.
12. Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) CATH—a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.
13. Eisenberg,D. (1984) Three-dimensional structure of membrane and surface proteins. *Annu. Rev. Biochem.*, **53**, 595–623.
14. Chou,P.Y. and Fasman,G.D. (1974) Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. *Biochemistry*, **13**, 211–222.
15. George,R.A. and Heringa,J. (2002) An analysis of protein domain linkers: their classification and role in protein folding. *Protein Eng.*, **15**, 871–879.
16. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bradley,P., Bork,P., Bucher,P., Cerutti,L. *et al.* (2005) InterPro, progress and status in 2005. *Nucleic Acids Res.*, **33**, 201–205.
17. Wheelan,S.J., Marchler-Bauer,A. and Bryant,S.H. (2000) Domain size distributions can predict domain boundaries. *Bioinformatics*, **16**, 613–618.