Nucleic Acids Research, 2003, Vol. 31, No. 13 ***3833–3835***
DOI: 10.1093/nar/gkg515

NORSp: predictions of long regions without regular secondary structure

Jinfeng Liu[1,3,4] and Burkhard Rost[1,2,3,*]

[1]CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168th Street BB217, [2]Columbia University Center for Computational Biology and Bioinformatics (C2B2), Russ Berrie Pavilion, 1150 St Nicholas Avenue, [3]North East Structural Genomics Consortium (NESG), Department of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168th Street BB217 and [4]Department of Pharmacology, Columbia University, 630 West 168th Street, New York, NY 10032, USA

Received February 14, 2003; Revised and Accepted March 17, 2003

ABSTRACT

Many structurally flexible regions play important roles in biological processes. It has been shown that extended loopy regions are very abundant in the protein universe and that they have been conserved through evolution. Here, we present NORSp, a publicly available predictor for disordered regions in protein. Specifically, NORSp predicts long regions with NO Regular Secondary structure. Upon user submission of a protein sequence, NORSp will analyse the protein for its secondary structure, presence of transmembrane helices and coiled-coil. It will then return email to the user about the presence and position of disordered regions. NORSp can be accessed from http://cubic.bioc.columbia.edu/services/NORSp/.

INTRODUCTION

Irregular structures mediate function

The three-dimensional (3D) structure of a protein is assumed to largely determine its biological function. The first decades of rapid progress in the experimental determination of 3D structures by X-ray crystallography (1) focused on determining 'rigid' structures at high resolution. Recently, a new type of structure has emerged with very long regions that appear to adopt regular structure only upon binding to substrates or other proteins (2); they are referred to as floppy, natively disordered, natively unfolded or loopy (3,4–7). It seems that these irregular regions are important for function.

Predicting irregular structures

Structural irregularity can be studied from several aspects: one class of 'natively disordered' regions was defined as the regions invisible in electron density maps of X-ray diffraction, presumably since the flexibility keeps them from crystallising into well-ordered structures. These regions sometimes are associated with regions with 'compositional bias' or 'low sequence complexity' (8–10). Another class is characterised by proteins that appear unfolded by CD measurements (5). Previously, we investigated the problem of disordered proteins from a structure-oriented perspective and studied extended regions of very low regular secondary structure (helix or strand) content (NORS) (3). We showed that NORS regions are particularly abundant in eukaryotic proteomes, conserved during evolution, over-represented in regulatory function category and important in protein–protein interactions. These results were in agreement with studies that predicted 'natively disordered regions' through neural networks (11).

Here, we introduced a web-based interface to make our method of predicting NORS regions publicly accessible. The method can be useful for biologists in several ways. For example, crystallographers can check whether their proteins contain NORS regions and make the decision about whether to proceed with the experiments since NORS proteins may be difficult to crystallise, as demonstrated by their low occurrence in PDB (3). Biologists interested in protein structure–function relationship may also find it interesting to verify whether the protein–protein interaction sites coincide with NORS regions.

DESIGN AND IMPLEMENTATION

Definition of NORS

We defined NORS regions as segments of >70 consecutive residues with $<12\%$ of the residues in helix, strand or coiled-coil regions and with at least one segment of 10 adjacent residues exposed to solvent. We identify such NORS regions by merging predictions of secondary structure, transmembrane helices and coiled-coil regions. We pre-calculate this information as well as NORS regions for each protein in >60 completely sequenced genomes (Fig. 1), and have included them in our PEP database (12) through a searchable SRS (13)

*To whom correspondence should be addressed. Tel: +1 2123054018; Fax: +1 2123057932; Email: rost@columbia.edu

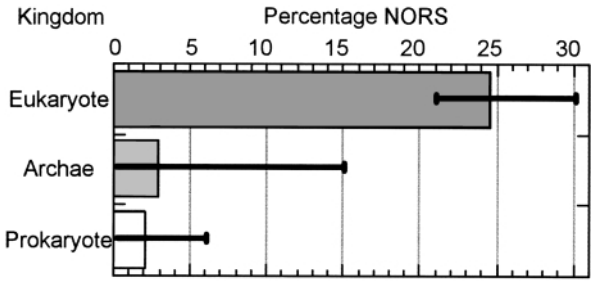

Figure 1. NORS proteins are much more abundant in eukaryotes than in prokaryotes and archae-bacteria. Shown in the graph are the average percentages of NORS proteins in the three kingdoms. Error bars indicate the maximum and minimum values.

interface (http://cubic.bioc.columbia.edu/db/PEP/). NORS information has also been used in our target selection process for North East Structural Genomics Consortium (14) to exclude proteins likely to pose problems to crystallisation.

Prediction by NORSp

Protein sequences submitted to our web site are subjected to the following steps. (a) Build sequence profile through a database search with an automated, iterated PSI-BLAST (15). (b) Secondary structure and solvent accessibility are predicted by PROFphd (16), membrane helices are predicted by the PHDhtm (17) using the PSI-BLAST profiles. (c) Coiled-coil regions are predicted by COILS (18). (d) The secondary structure, membrane helices and coiled-coil information are then combined to calculate the structural content for each sequence window of a certain length, and NORS regions are identified when the structural content is below the given threshold; overlapping NORS regions are joined. Technically, to obtain most of these intermediate results, NORSp utilises the same engine which is behind the PredictProtein server (19) (http://cubic.bioc.columbia.edu/predictprotein/).

INPUT, OUTPUT AND ADVANCED OPTIONS

Input

The input to NORSp is protein sequence; proteins shorter than 70 residues are returned unprocessed. Currently, the valid input format is a sequence in one-letter residue code or a FASTA-format. The sequence can be entered into the sequence text box or uploaded from users' local disk.

Output

Users have the option of receiving 'succinct' output, which only shows the position of the NORS region in the context of the submitted sequence, or 'verbose' output, which includes the intermediate data used by NORSp: secondary structure, solvent accessibility, transmembrane helices and coiled-coil prediction. By default, the results will be in plain text (ASCII) format. However, HTML formatted results can also be requested that can be displayed in any web browser. Due to concerns about file size and user mailbox overflow, the results will normally be available to download from our website and only URLs are sent to the users by email unless users request the full results being sent directly.

Recommendation and advanced options

We determined the particular threshold used to define NORS regions in order to minimise the false positive rate as determined by manually inspecting PDB proteins (3). This conservative solution implies that the vast majority of NORS regions that we detect are likely to constitute structurally irregular, floppy, loopy or natively disordered regions. However, we supposedly miss many such regions in our predictions. Users who are aware of this may be interested in changing the threshold to see which regions may be good candidates for irregular regions although not detected by our default. We provide three options for advanced users: the size of sequence window for calculating secondary structure content (default = 70), maximum of secondary structure content (default = 12%) and the minimum length of consecutive exposed residues (default = 10).

ACKNOWLEDGEMENTS

We are grateful to Hepan Tan (Columbia) for his help in developing the tool. This work was supported by grants 1-P50-GM62413-01 and RO1-GM63029-01 from the National Institute of Health (NIH). Last, but not least, thanks go to all those who deposit their experimental data in public databases and to those who maintain these databases.

REFERENCES

1. Hendrickson,W.A. (1991) Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. *Science*, **254**, 51–58.
2. Wright,P.E. and Dyson,H.J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.*, **293**, 321–331.
3. Liu,J., Tan,H. and Rost,B. (2002) Loopy proteins appear conserved in evolution. *J. Mol. Biol.*, **322**, 53–64.
4. Dunker,A.K. and Obradovic,Z. (2001) The protein trinity-linking function and disorder. *Nat. Biotechnol.*, **19**, 805–806.
5. Uversky,V.N., Gillespie,J.R. and Fink,A.L. (2000) Why are 'natively unfolded' proteins unstructured under physiologic conditions? *Proteins*, **41**, 415–427.
6. Zetina,C.R. (2001) A conserved helix-unfolding motif in the naturally unfolded proteins. *Proteins*, **44**, 479–483.
7. Namba,K. (2001) Roles of partly unfolded conformations in macromolecular self-assembly. *Genes Cells*, **6**, 1–12.
8. Dunker,A.K., Garner,E., Guilliot,S., Romero,P., Albrecht,K., Hart,J., Obradovic,Z., Kissinger,C. and Villafranca,J.E. (1998) Protein disorder and the evolution of molecular recognition: theory, predictions and observations. *Pac. Symp. Biocomput.*, 473–484.
9. Wootton,J.C. and Federhen,S. (1996) Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.*, **266**, 554–571.
10. Dunker,A.K., Lawson,J.D., Brown,C.J., Williams,R.M., Romero,P., Oh,J.S., Oldfield,C.J., Campen,A.M., Ratliff,C.M., Hipps,K.W. *et al.* (2001) Intrinsically disordered protein. *J. Mol. Graph. Model.*, **19**, 26–59.
11. Romero,P., Obradovic,Z., Li,X., Garner,E.C., Brown,C.J. and Dunker,A.K. (2001) Sequence complexity of disordered protein. *Proteins*, **42**, 38–48.
12. Carter,P., Liu,J. and Rost,B. (2003) PEP: predictions for entire proteomes. *Nucleic Acids Res.*, **31**, 410–413.
13. Etzold,T. and Argos,P. (1993) SRS–an indexing and retrieval tool for flat file data libraries. *Comput. Appl. Biosci.*, **9**, 49–57.

14. Liu,J. and Rost,B. (2002) Target space for structural genomics revisited. *Bioinformatics*, **18**, 922–933.
15. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
16. Rost,B. (2001) Review: protein secondary structure prediction continues to rise. *J. Struct. Biol.*, **134**, 204–218.
17. Rost,B., Casadio,R. and Fariselli,P. (1996) Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci.*, **5**, 1704–1718.
18. Lupas,A. (1996) Prediction and analyis of coiled-coil structures. *Methods Enzymol.*, **266**, 513–525.
19. Rost,B. and Liu,J. (2003) The PredictProtein server. *Nucleic Acids Res.*, **31**, 3300–3304.