BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 16 2005, pages 3433–3434
doi:10.1093/bioinformatics/bti541

Structural bioinformatics

IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content

Zsuzsanna Dosztányi*, Veronika Csizmok, Peter Tompa and István Simon

Institute of Enzymology, BRC, Hungarian Academy of Sciences, PO Box 7, H-1518 Budapest, Hungary

Received on March 24, 2005; revised on May 27, 2005; accepted on June 13, 2005
Advance Access publication June 14, 2005

ABSTRACT

Summary: Intrinsically unstructured/disordered proteins and domains (IUPs) lack a well-defined three-dimensional structure under native conditions. The IUPred server presents a novel algorithm for predicting such regions from amino acid sequences by estimating their total pairwise interresidue interaction energy, based on the assumption that IUP sequences do not fold due to their inability to form sufficient stabilizing interresidue interactions. Optional to the prediction are built-in parameter sets optimized for predicting short or long disordered regions and structured domains.

Availability: The IUPred server is available for academic users at http://iupred.enzim.hu

Contact: zsuzsa@enzim.hu

INTRODUCTION

Instrinsically unstructured proteins exist as an ensemble of alternative conformations, in contrast to folded, globular proteins that have unique native structure. Significant fraction of known genomes encode for proteins with regions of disordered structure. In some eukaryotic genomes >20% of the coded residues are predicted as disordered (Dunker *et al*., 2000; Ward *et al*., 2004a). In many cases a protein is fully disordered, while in many other cases there are long disordered segments in otherwise ordered, folded proteins (Tompa, 2002; Dyson and Wright, 2005). Despite their lack of a well-defined globular structure, these proteins carry out basic functions (Iakoucheva *et al*., 2002; Ward *et al*., 2004a), mostly associated with signal transduction, cell-cycle regulation and transcription. Several methods have been developed to predict the disordered character from amino acid sequences. Some are based on the special amino acid composition of fully disordered proteins, i.e. the abundance of hydrophilic residues and a high net charge (Uversky *et al*., 2000; Vucetic *et al*., 2003), whereas others use various machine learning approaches trained on specific datasets (Obradovic *et al*., 2003; Ward *et al*., 2004a; Linding *et al*., 2003b). Recently, it was suggested that these sequences do not have the capacity to properly wrap backbone hydrogen bonds (Fernandez and Berry, 2004), which has also been shown to be important for protein stability.

BACKGROUND

Our method is footed on the physical explanation of the ordered/disordered nature of proteins. Globular proteins make a large number of interresidue interactions, providing the stabilizing energy to overcome the entropy loss during folding (Garbuzynskiy *et al*., 2004). In contrast, intrinsically unstructured/disordered proteins and domains (IUPs) have special sequences that do not have the capacity to form sufficient interresidue interactions. To discriminate between ordered and disordered regions in proteins, we have developed a new approach that estimates the potential of polypeptides to form such stabilizing contacts by using a statistical interaction potential (Thomas and Dill, 1996; Dosztányi *et al*., 2005). It was shown that the sum of interaction energies can be estimated by a quadratic expression in the amino acid composition, which takes into account that the contribution of an amino acid to order/disorder depends not only on its own chemical type, but also on its potential interaction partners (Dosztányi *et al*., 2005).

The calculation involves a 20 × 20 energy predictor matrix, parameterized by a statistical method to approach the expected pairwise energy of globular proteins of known structure. Comparing globular proteins and disordered ones, a clear separation of their energy content is found (Dosztányi *et al*., 2005). As no training on disordered proteins is involved, this distinction underlines that the lack of a well-defined three-dimensional structure is an intrinsic property of certain evolved proteins. This approach was turned into a position-specific method to predict protein disorder by considering only the local sequential environment of residues within 2–100 residues in either direction. The score is then smoothed over a window-size of 21. This prediction method (IUPred), when tested on datasets of globular proteins and long disordered protein segments, showed improved performance over some other widely used methods, such as DISOPRED2 (Ward *et al*., 2004a,b) and PONDR VL3H (Obradovic *et al*., 2003).

THE IUPred SERVER

The web server takes a single amino acid sequence as an input and calculates the pairwise energy profile along the sequence. The energy values are then transformed into a probabilistic score ranging from 0 (complete order) to 1 (complete disorder). Residues with a score above 0.5 can be regarded as disordered. Optional is the prediction of long disorder, short disorder, and structured domains, each using slightly different parameters. The main profile of our server is to predict context-independent global disorder that encompasses at least 30 consecutive residues of predicted disorder. A different set of parameters is suited for predicting short, probably context-dependent, disordered regions such as missing residues in the X-ray

*To whom correspondence should be addressed.

© The Author 2005. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oupjournals.org

Downloaded from http://bioinformatics.oxfordjournals.org/ by guest on October 24, 2014

structure of an otherwise globular protein. For this application the sequential neighborhood of only 25 residues is considered. As chain termini of globular proteins are often disordered in X-ray structures, this is taken into account by an end-adjustment parameter that favors disorder prediction at the ends.

The dependable identification of ordered regions is a crucial step in target selection for structural studies and structural genomics projects (Linding *et al*., 2003a). Finding putative structured domains suitable for stucture determination is another potential application of this server. In this case the algorithm takes the energy profile and finds continuous regions confidently predicted ordered. Neighboring regions close to each other are merged, while regions shorter than the minimal domain size of at least 30 residues are ignored. When this prediction type is selected, the region(s) predicted to correspond to structured/globular domains are returned.

The core program to calculate the pairwise energy profile and disorder probability is written in C, the web server is written in PHP. The calculation of the energy profile is based on single sequence, without time-consuming alignment calculations. To further facilitate the easy accessibility for scripting, a simple text output is generated on default. However, the user can also request a graphical output. The plot shows the disorder tendency of each residue along the sequence. The plot is generated by the JpGraph software (JpGraph, 2005, http://www.aditus.nu/jpgraph/) on the fly, without storing the graphical images on the local machine. When the prediction type of structured domains is selected, these are highlighted on the plot by thick lines. For long sequences, the graph is shown for fragments of user-defined fixed length, 500 on default.

ACKNOWLEDGEMENTS

This work has been sponsored by grants GVOP-3.1.1.-2004-05-0143/3.0, OTKA F043609, T049073, and NKFP MediChem2 1/A/005/2004. Z.D. and P.T. were supported by the Bolyai János Scholarship. P.T. would like to acknowledge the support of the International Senior Research Fellowship GR067595 from the Wellcome Trust.

Conflict of Interest: none declared.

REFERENCES

Dosztányi,Z. *et al*. (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.*, **347**, 827–839.

Dunker,A.K. *et al*. (2000) Intrinsic protein disorder in complete genomes. *Genome Inform. Ser. Workshop Genome Inform.*, **11**, 161–171.

Dyson,H.J. and Wright,P.E. (2005) Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.*, **6**, 197–208.

Fernandez,A. and Berry,R.S. (2004) Molecular dimension explored in evolution to promote proteomic complexity. *Proc. Natl Acad. Sci. USA*, **101**, 13460–13465.

Garbuzynskiy,S.O. *et al*. (2004) To be folded or to be unfolded? *Protein Sci.*, **13**, 2871–2877.

Iakoucheva,L.M. *et al*. (2002) Intrinsic disorder in cell-signaling and cancer-associated proteins. *J. Mol. Biol.*, **323**, 573–584.

JpGraph (2005) *JpGraph*. Aditus Consulting.

Linding,R. *et al*. (2003a) GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Res.*, **31**, 3701–3708.

Linding,R. *et al*. (2003b) Protein disorder prediction: implications for structural proteomics. *Structure (Camb)*, **11**, 1453–1459.

Obradovic,Z. *et al*. (2003) Predicting intrinsic disorder from amino acid sequence. *Proteins*, **53** (Suppl. 6), 566–572.

Thomas,P.D. and Dill,K.A. (1996) An iterative method for extracting energy-like quantities from protein structures. *Proc. Natl Acad. Sci. USA*, **93**, 11628–11633.

Tompa,P. (2002) Intrinsically unstructured proteins. *Trends Biochem. Sci.*, **27**, 527–533.

Uversky,V.N. *et al*. (2000) Why are 'natively unfolded' proteins unstructured under physiologic conditions? *Proteins*, **41**, 415–427.

Vucetic,S. *et al*. (2003) Flavors of protein disorder. *Proteins*, **52**, 573–584.

Ward,J.J. *et al*. (2004a) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.*, **337**, 635–645.

Ward,J.J. *et al*. (2004b) The DISOPRED server for the prediction of protein disorder. *Bioinformatics*, **20**, 2138–2139.

Downloaded from http://bioinformatics.oxfordjournals.org/ by guest on October 24, 2014