BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 16 2005, pages 3435–3438
doi:10.1093/bioinformatics/bti537

Structural bioinformatics

FoldIndex©: a simple tool to predict whether a given protein sequence is intrinsically unfolded

Jaime Prilusky[1,†], Clifford E. Felder[2,†], Tzviya Zeev-Ben-Mordehai[2,3], Edwin H. Rydberg[2,3,5], Orna Man[2,4], Jacques S. Beckmann[4,6], Israel Silman[3] and Joel L. Sussman[2,*]

[1]Biological Services and [2]Department of Structural Biology, [3]Department of Neurobiology and [4]Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, Israel, [5]Department of Biochemistry, Instituto di Ricerche di Biologia Molecolare 'P. Angeletti' S.p.A. I-00040 Pomezia (Rome), Italy and [6]Service of Medical Genetics, Centre Hospitalier Universitaire Vaudois, CH-1011 Lausanne, Switzerland

Received on August 27, 2004; revised on May 10, 2005; accepted on June 8, 2005
Advance Access publication June 14, 2005

ABSTRACT

Summary: An easy-to-use, versatile and freely available graphic web server, FoldIndex© is described: it predicts if a given protein sequence is intrinsically unfolded implementing the algorithm of Uversky and co-workers, which is based on the average residue hydrophobicity and net charge of the sequence. FoldIndex© has an error rate comparable to that of more sophisticated fold prediction methods. Sliding windows permit identification of large regions within a protein that possess folding propensities different from those of the whole protein.

Availability: FoldIndex© can be accessed at http://bioportal.weizmann.ac.il/fldbin/findex

Contact: Joel.Sussman@weizmann.ac.il

Supplementary information: http://www.weizmann.ac.il/sb/faculty_pages/Sussman/papers/suppl/Prilusky_2005

A growing number of proteins have been found to be natively unfolded under physiological conditions (Dunker *et al*., 2000; Wright and Dyson, 1999; Zhang and Forman-Kay, 1997; Uversky *et al*., 2000; Bell *et al*., 2002; Schweers *et al*., 1994; Zeev-Ben-Mordehai *et al*., 2003; Vucetic *et al*., 2004). Uversky *et al*. (2000) described a simple method to predict whether a given protein assumes a defined fold or is intrinsically unfolded. It is based solely on the average hydrophobicity of its amino acids and on the absolute value of its net charge. Using as axes these two parameters, proteins determined experimentally to be folded and intrinsically unfolded ones are separated by a straight line described by a simple equation. This simple procedure allows rapid prediction of whether a given sequence is disordered or not.

Several other methods are available for predicting whether or not protein sequences are intrinsically unfolded. They include: PONDR (Romero *et al*., 1997; Dunker *et al*., 2002, http://www.disprot.org), NORSP (Liu and Rost, 2003, http://cubic.bioc.columbia.edu/services/NORSp), DisEMBL™(Linding *et al*., 2003a, http://dis.embl.de), DISOPRED (Ward *et al*., 2004, http://bioinf.cs.ucl.ac.uk/disopred/) and GlobPlot (Linding *et al*., 2003b, http://globplot.embl.de).

We have implemented the algorithm of Uversky *et al*. (2000) by transforming the equation of the boundary line separating 'folded' from 'disordered' proteins into a simple index, FoldIndex©, that discriminates between folded and intrinsically unfolded proteins. Uversky *et al*. (2000) defined the mean net charge, $|<R>|$, as the absolute value of the difference between the numbers of positively and negatively charged residues at pH 7.0, divided by the total residue number, and the mean hydrophobicity, $<H>$, as the sum of all residue hydrophobicities, divided by the total number of residues, using the Kyte/Doolittle scale (Kyte and Doolittle, 1982), rescaled to a range of 0–1. We rearranged their fold boundary equation $|<R>| = 2.785<H> - 1.151$ (Uversky *et al*., 2000) to yield the Fold Index© as:

$$I_{\mathrm{F}}^{\mathrm{KD}} = 2.785<H> - |<R>| - 1.151.$$

All positive values thus represent proteins (or domains) likely to be folded, and negative values represent those likely to be intrinsically unfolded.

To independently evaluate FoldIndex© and compare it to three other tools, we compiled our own sets of folded and unfolded proteins. The unfolded set consisted of 39 proteins (or domains) reported in the literature to be intrinsically unfolded: experimental data suggested that all are fully unfolded throughout their entire lengths. The 151 folded proteins were obtained from the OCA Protein Data Bank (PDB) browser at http://bioportal.weizmann.ac.il/oca, specifying only X-ray structures consisting of a single polypeptide chain, 50–200 residues long, with neither disulfide linkages nor non-protein elements, such as nucleic acids, heterogen groups or metal ions, to eliminate any possible external template that might help the polypeptide fold. Folded proteins with sequences missing >5%, or 10 residues, in the ATOM records relative to the SEQRES sequence, or those seen to be homologous duplicates at the 90% level by the nrdb90 tool (Holm and Sander, 1998), were excluded from the list of folded proteins.

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Table 1. Comparison of the efficacy of four algorithms in prediction of the folding status of sets of folded and intrinsically unfolded test proteins

	FoldIndex	DISOPRED	PONDR	GlobPlot
Intrinsically unfolded (39)				
Correct	30 (77%)	22 (56%)	28 (72%)	9 (23%)
Unassigned	0 (0%)	0 (0%)	0 (0%)	3 (8%)
Incorrect	9 (23%)	17 (44%)	11 (28%)	27 (69%)
Folded (151)				
Correct	133 (88%)	149 (99%)	140 (93%)	148 (98%)
Unassigned	3 (2%)	0 (0%)	2 (1%)	2 (1%)
Incorrect	15 (10%)	2 (1%)	9 (6%)	1 (1%)

The web tool FoldIndex©, implementing the original Uversky algorithm, provides a single score for the entire sequence, predicting whether it is folded or not. The other methods calculate a separate fold score for each individual residue. In order to compare the various methods we obtained a fold score for the entire sequence from the scores of the individual residues by calculating the arithmetic mean for PONDR and GlobPlot, and the geometric mean for DISOPRED. The use of the geometric mean for the DISOPRED score was deemed necessary because of the model on which this algorithm is based. It results in a highly skewed distribution with the range of scores for ordered residues being very narrow, relative to the range for disordered residues. We thus assumed that the DISOPRED scores for the residues of its training set, which contains mainly ordered residues, follow a log–normal distribution, making the geometric mean a more appropriate measure of whole sequence disorder rather than the arithmetic mean. Unless otherwise noted, the default settings were used with each method. For PONDR, the VL-XT method was used. It should be noted that for PONDR, DISOPRED and GlobPlot, scores above 0.5, 0.05 and 0, respectively, indicate a disordered protein. Scores very close to the boundary line, within ± 0.005 for FoldIndex©, DISOPRED and GlobPlot, and within ± 0.05 for PONDR VL-XT, were scored as unassigned, equivalent to no prediction. Table 1 summarizes the results obtained using the four procedures. Supplementary Material containing the lists of 39 intrinsically unfolded proteins and 151 folded proteins utilized in this comparison, as well as all individual scores, is available at http://www.weizmann.ac.il/sb/faculty_pages/Sussman/papers/suppl/Prilusky_2005

Inspection of these results shows that FoldIndex© performs slightly better than PONDR, and substantially better than DISOPRED and GlobPlot in correctly predicting the unfolded status of proteins shown experimentally to be intrinsically unfolded. In contrast, DISOPRED and GlobPlot perform best in predicting the folded status of proteins indeed known to be folded, with PONDR and FoldIndex© both being somewhat less successful. What might be the reason for these opposite trends? With respect to the intrinsically unfolded proteins, the FoldIndex© algorithm relies directly on data relating to charge and hydrophobicity, and PONDR makes substantial use of similar information relating to order-promoting and order-breaking amino acids. In contrast, DISOPRED and GlobPlot employ training sets that utilize disordered sequences in the PDB structures. It should be noted that the GlobPlot method is the most similar to that of FoldIndex© , inasmuch as it based on a simple 'disorder scale', namely the Russell/Linding Scale, which thus does not suffer from the possibility of 'over training' (Linding *et al*., 2003b).

The single global value for an entire sequence utilized by FoldIndex© and PONDR militates against successful prediction in a small number of cases for folded proteins, because fine structural matching may overcome the global physical chemical properties for borderline cases.

The *Drosophila* adhesion molecule, gliotactin, contains an extracellular domain with sequence homology to acetylcholinesterase, a transmembrane sequence and an intracellular sequence devoid of homology to any known protein sequence (Auld and Gilbert, 2005). FoldIndex© predicts that the first two segments, as expected, are folded, while the intracellular segment is predicted to be almost completely unfolded. This prediction for the latter segment was confirmed by physicochemical characterization (Zeev-Ben-Mordehai *et al*., 2003). Thus, FoldIndex© permits examination of the fold properties of overlapping segments, or sliding windows, within a sequence, to identify contiguously large regions with different fold properties than the protein as a whole. Figure 1 displays snapshots of the server outputs for three proteins shown experimentally to be folded, partially folded and intrinsically unfolded.

FoldIndex© can be used both as an interactive web tool and as an automated web service. The GUI (http://bioportal.weizmann.ac.il/fldbin/findex) generates a graph describing I_F^{KD} for a submitted sequence. It can superimpose a graph of the running averages of hydrophobic amino acids and charged amino acids, and generate a detailed report of values for each stage of the sliding window during the analysis. A file containing multiple sequences in 'fasta' format can also be uploaded, yielding a detailed report of the I_F^{KD} values for each sequence. The user can select the size of the sliding window or protein fragment size, step size to the next window, and specify the output format, as plain text, XML format (DTD, sample XML) or eFamily compatible format (http://www.sanger.ac.uk/xml/efamily/documentation/eFamily.html). These latter formats are particularly useful if a program or script is required to parse the results automatically. The output for this option is either displayed in the browser or e-mailed, depending on output size. The simplicity of the method permits rapid prediction of foldability for a large number of sequences. FoldIndex© is freely available for unlimited use by all classes of users, and returns the results right inside the web page within seconds. FoldIndex© should serve as a valuable tool for protein crystallographers, especially in the area of structural genomics/proteomics, by directing them to proteins or subsegments that are more likely to crystallize (Jaakola *et al*., 2005) and see, e.g. http://www.weizmann.ac.il/ISPC/biotools.html

FoldIndex© can be used as a web service for remote and automatic data processing by accessing http://bioportal.weizmann.ac.il/fldbin/findex?m=xml&sq=SEQUENCE where 'm' is either 'xml' (for XML output format) or 'efam' (for eFamily compatible format), and 'sq' is the one character code protein sequence, no spaces. A simple working perl script using this technique is available from http://bioportal.weizmann.ac.il/flddoc/example1.txt

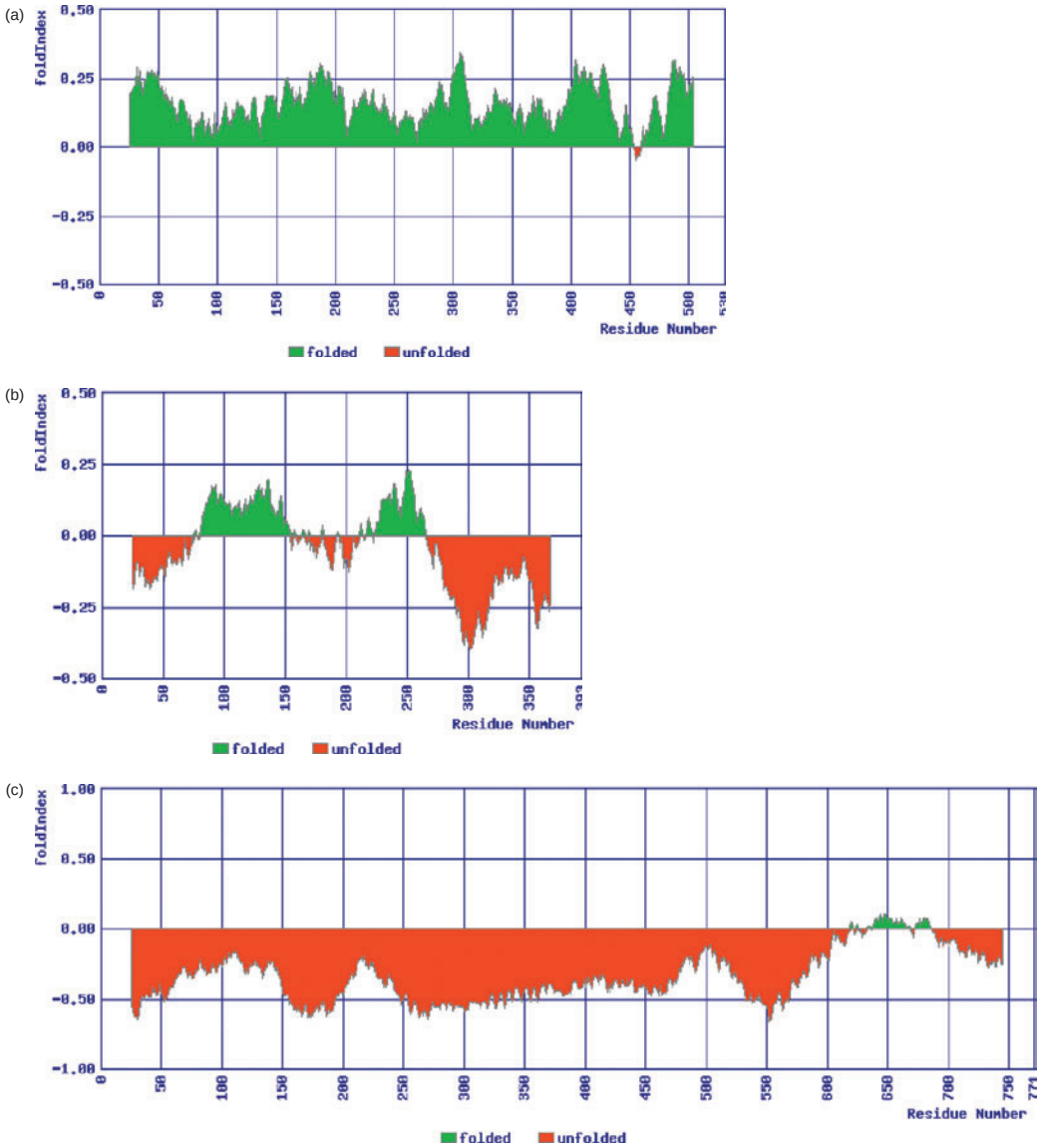

Fig. 1. FoldIndex plotted with window size 51, for three protein sequences. (**a**) Cat-Muscle (M1) Pyruvate Kinase is a well-folded three-domain structure (PDB-ID 1PKM; Swiss-Prot: P11979) (Allen and Muirhead, 1996). (**b**) The human p53 tumor suppressor protein contains large unstructured regions in its native state (PDB-ID 1TSR; Swiss-Prot: P04637) (Bell *et al*., 2002). (**c**) Chicken gizzard caldesmon is natively unfolded (Swiss-Prot: P12957) (Permyakov *et al*., 2003).

ACKNOWLEDGEMENTS

We thank both referees for their constructive suggestions. Support was by the EC Vth Framework 'SPINE' Project QLG2-CT-2002-00988, by an Israel Ministry of Science and Technology grant to the Israel Structural Proteomics Center, by the Divadol Foundation and the Minerva Foundation. J.L.S. is Morton and Gladys Pickman Professor of Structural Biology.

Conflict of Interest: none declared.

REFERENCES

Allen,S.C. and Muirhead,H. (1996) Refined three-dimensional structure of Cat-Muscle (M1) Pyruvate Kinase at a resolution of 2.6 Å. *Acta Crystallogr. D Biol. Crystallogr.*, **52**, 499–504.

Auld,V.J. and Gilbert,M.M. (2005) Evolution of CLAMs (Cholinesterase-Like Adhesion Molecules): structure and function during development. *Front. Biosci.*, **10**, 2177–2192.

Bell,S. *et al*. (2002) p53 contains large unstructured regions in its native state. *J. Mol. Biol.*, **322**, 917–927.

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

Dunker,A.K. *et al*. (2002) Intrinsic disorder and protein function. *Biochemistry*, **41**, 6573–6582.

Holm,L. and Sander,C. (1998) Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics*, **14**, 423–429.

Jaakola,V.-P. *et al*. (2005) G protein-coupled receptors show unusual patterns of intrinsic unfolding. *PEDS*, **18**, 103–110.

Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105–132.

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

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

Liu,J. and Rost,B. (2003) NORSp: predictions of long regions without regular secondary structure. *Nucleic Acids Res.*, **31**, 3833–3835.

Permyakov,S.E. *et al*. (2003) Natively unfolded C-terminal domain of caldesmon remains substantially unstructured after the effective binding to calmodulin. *Proteins*, **53**, 855–862.

Romero,P. *et al*. (1997) Sequence data analysis for long disordered regions prediction in the calcineurin family. *Genome Informatics*, **8**, 110–124.

Schweers,O. *et al*. (1994) Structural studies of tau protein and Alzheimer paired helical filaments show no evidence for beta-structure. *J. Biol. Chem.*, **269**, 24290–24297.

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

Vucetic,S. *et al*. (2004) DisProt: a database of protein disorder. *Bioinformatics*, **21**, 137–140.

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

Wright,P.E. and Dyson,H.J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure–function paradigm. *J. Mol. Biol.*, **293**, 321–331.

Zeev-Ben-Mordehai,T. *et al*. (2003) The intracellular domain of the *Drosophila* cholinesterase-like neural adhesion protein, gliotactin, is natively unfolded. *Proteins*, **53**, 758–767.

Zhang,O. and Forman-Kay,J.D. (1997) NMR studies of unfolded states of an SH3 domain in aqueous solution and denaturing conditions. *Biochemistry*, **36**, 3959–3970.