FoldUnfold: web server for the prediction of disordered regions in protein chain

Oxana V. Galzitskaya*, Sergiy O. Garbuzynskiy, Michail Yu. Lobanov

Institute of Protein Research, Russian Academy of Sciences, 142290, Pushchino, Moscow Region, Russia

Associate Editor: Dmitrij Frishman

ABSTRACT

Summary: Identification of disordered regions in polypeptide chains is very important because such regions are essential for protein function. A new parameter, namely mean packing density of residues has been introduced to detect disordered regions in a protein sequence. We have demonstrated that regions with weak expected packing density would be responsible for the appearance of disordered regions. Our method (FoldUnfold) has been tested on datasets of globular proteins (559 proteins) and long disordered protein segments (129 proteins) and showed improved performance over some other widely used methods, such as DISOPRED, PONDR VL3H, IUPred, GlobPlot.

Availability: The FoldUnfold server is available for users at http://skuld.protres.ru/~mlobanov/ogu/ogu.cgi. There is a link to our server through the web-site of DisProt (http://www.disprot.org/predictors.php).

Contact: ogalzit@vega.protres.ru

BACKGROUND

The formation of a sufficient number of interactions is necessary to compensate the loss of conformational entropy during the protein folding process. Therefore, structural uniqueness of native protein is a result of the balance between the conformational entropy and the energy of residue interactions. It seems that disordered regions in a protein chain do not have a sufficient amount of interactions to compensate the loss of conformational entropy resulting from the formation of a globular state (Galzitskaya et al., 2000). Therefore, their enhanced stabilization can be achieved by additional interactions with other agents or by oligomerization.

It was shown that disordered regions are involved in DNA-binding and other types of molecular recognition and a large portion of the sequences of natively unfolded proteins contain segments of low complexity and high predicted flexibility (Wootton, 1994; Romero et al., 1998; Wright and Dyson, 1999; Galzitskaya et al., 2000; Obradovic et al., 2003; Radivojac et al., 2004). Also it was indicated that a combination of low overall hydrophobicity and a large net charge represents a structural feature of natively unfolded proteins in comparison with small globular proteins (Uversky et al., 2000). Now there are several widely used methods to predict disordered regions in proteins: GlobPlot (Linding et al., 2003) is a simple propensity-based approach evaluating the tendency of residues to be in a regular secondary structure; PONDR VL3H (Obradovic et al., 2003) was trained to distinguish experimentally verified disordered proteins from globular proteins by various machine learning approaches; DISOPRED (Ward et al., 2004) was trained to specifically recognize regions missing in X-ray structures; IUPred (Dosztanyi et al., 2005) assigns the order/disorder status to residues on the basis of their ability to form favorable pairwise contacts. We were the first who used such parameter as the number of contacts per residue to distinguish folded and natively unfolded proteins (Galzitskaya et al., 2004). We have extended our method to predict disordered regions and made comparison with the above mentioned methods (Galzitskaya et al., 2006). It has been demonstrated that our method is the best among widely used methods.

INTRODUCTION

Mean packing density was calculated for each amino acid residue from the database of 5829 three-dimensional structures as an average number of close residues (within the given distance). In our case a residue will be considered close to the given residue if any pair of their heavy atoms is at a distance less than 8 Å excluding the neighboring residues. The mean packing density in a globular state for each of 20 types of amino acid residues is presented in our work (Galzitskaya et al., 2006).

To detect disordered regions, we construct a packing density profile of the expected packing density for the protein sequence. The calculations are based on a sliding window averaging technique. First, the expected packing density is determined for each residue (it equals to the average packing density observed for this type of residue in a globular state); mean, these numbers are averaged inside the window and assigned to the central residue of the window. The value of the averaged expected packing density for every position of the polypeptide chain provides the packing density profile.

Our method has been tested on datasets of globular proteins (559 proteins) and long disordered protein segments (129 proteins) (Dosztanyi et al., 2005). A receiver operator characteristic (ROC) curve for our method has been obtained (Galzitskaya et al., 2006) to determine a threshold for our method. The true positive rate was calculated as the percentage of residues predicted as disordered on the set of the disordered proteins and segments; the false positive rate is the percentage of predicted disordered residues on the set of globular proteins. Our method showed improved performance over some other widely used methods, such as DISOPRED (Ward et al., 2004) PONDR VL3H (Obradovic et al., 2003), IUPred (Dosztanyi et al., 2005), GlobPlot (Linding et al., 2003) (see Table 1).

* To whom correspondence should be addressed.
The FoldUnfold server takes amino acid sequence in FASTA format as an input and calculates the expected packing density profiles along the sequence. We used this property, that is the mean packing density, to predict the state of protein with an unknown three-dimensional structure: either folded or unfolded (in other words, disordered). If the expected mean packing density in protein is less than 20.4 and the size of this segment is equal or larger than the size of the window used, the program can predict unfolded regions of size equal or greater than the window-size used.

We have also made predictions of disordered regions in 129 proteins (Dosztanyi et al., 2005) using recently published method RONN (Yang et al., 2005). True positive rate for this method (0.765 if averaging is done over residues and 0.694 if averaging is done over proteins) does not exceed that of our method (0.851 and 0.716, respectively, see Table 1). Comparison of our method with other new published methods (PONDR VSL2 (Obradovic et al., 2005), PreLink (Coeytaux and Poupon, 2005), SPRITZ (Vullo et al., 2006)) will be done in next publications.

**ACKNOWLEDGEMENTS**

This work was supported by the program MCB RAS, by the Russian Foundation for Basic Research (grant 05-04-48750), by the Howard Hughes Medical Institute (55005607), and by INTAS grant (05-1000004-7747).

**Conflict of Interest:** none declared

**REFERENCES**


