Analyzing Protein Disorder with IUPred2A

Gábor Erdős and Zsuzsanna Dosztányi

1Department of Biochemistry, MTA-ELTE Momentum Bioinformatics Research Group, ELTE Eötvös Loránd University, Budapest, Hungary
2Corresponding author: dosztanyi@caesar.elte.hu

IUPred2A is a combined prediction tool designed to discover intrinsically disordered or conditionally disordered proteins and protein regions. Intrinsically disordered regions exist without a well-defined three-dimensional structure in isolation but carry out important biological functions. Over the years, various prediction methods have been developed to characterize disordered regions. The existence of disordered segments can also be dependent on different factors such as binding partners or environmental traits like pH or redox potential, and recognizing such regions represents additional computational challenges.

In this article, we present detailed instructions on how to use IUPred2A, one of the most widely used tools for the prediction of disordered regions/proteins or conditionally disordered segments, and provide examples of how the predictions can be interpreted in different contexts. © 2020 The Authors.

Basic Protocol 1: Analyzing disorder propensity with IUPred2A online
Basic Protocol 2: Analyzing disordered binding regions using ANCHOR2
Support Protocol 1: Interpretation of the results
Basic Protocol 3: Analyzing redox-sensitive disordered regions
Support Protocol 2: Download options
Support Protocol 3: REST API for programmatic purposes
Basic Protocol 4: Using IUPred2A locally

Keywords: binding sites • bioinformatics • context dependent disorder • intrinsically disordered proteins • redox regulation

How to cite this article:
Erdős, G., & Dosztányi, Z. (2020). Analyzing protein disorder with IUPred2A. Current Protocols in Bioinformatics, 70, e99. doi: 10.1002/cpbi.99

INTRODUCTION

One of the groundbreaking discoveries of recent years in the field of molecular biology was the observation that a significant portion of various genomes encodes intrinsically disordered proteins and protein regions (IDPs/IDRs). These protein segments exist without well-defined three-dimensional structures, and are best characterized by an ensemble of highly flexible conformations in isolation, but still retain the ability to have important biological functions, defying the traditional structure-function paradigm (van der Lee et al., 2014; Wright & Dyson, 1999). Some of the functionalities of IDPs are a direct consequence of their disordered nature: they can serve as flexible linkers or rely on their entropic properties to act as springs, spacers, or bristles. IDPs can also mediate
interactions with other proteins, macromolecules such as DNA and RNA, or the membrane. A special class of functional sites located within disordered regions corresponds to short linear motifs that mediate interactions with specific globular domains through a few key residues. Many such cases are represented in the ELM database (see Current Protocols article: Gouw et al., 2017). From a structural point of view, disordered regions can undergo a disorder-to-order transition and adopt a more defined three-dimensional structure upon recognizing their specific partners. However, some IDRs retain a significant amount of flexibility even in the bound form, forming fuzzy complexes (Borgia et al., 2018; Miskei, Antal, & Fuxreiter, 2017). It is increasingly recognized that in addition to binding, disorder-to-order or order-to-disorder transitions can also be induced by post-translational modifications (PTMs) and various environmental conditions such as pH or change in redox potential (Jakob, Kriwacki, & Uversky, 2014). These disordered binding sites and other conditionally disordered regions are commonly found throughout signaling pathways or as a part of regulatory elements.

Due to the critical biological roles of IDPs, there is an ever-growing need for the development of prediction methods to recognize disordered regions in proteins and to identify their potential functions. While there are dedicated databases to collect experimentally verified disordered proteins, for example, the DisProt database (Hatos et al., 2020), in most cases computational tools are needed to start the characterization of disordered regions. IUPred is one of the most widely used prediction methods for protein disorder. The basic assumption behind this method is that disordered regions are composed of amino acids that cannot form favorable interactions with each other. This property of protein sequences is captured by an energy-estimation approach utilizing statistical potentials (Dosztányi, Csizmók, Tompa, & Simon, 2005). The ANCHOR method is based on a similar biophysical approach and recognizes specific regions that undergo disorder-to-order transition upon binding (Mészáros, Simon, & Dosztányi, 2009). ANCHOR estimates the amount of additional favorable interactions a residue might be able to form by interacting with a more structured environment. Both the original IUPred and ANCHOR packages are popular predictor methods in their categories due to their high speed and accuracy.

In a recently launched new installment, we combined IUPred with ANCHOR (hence the name IUPred2A; Mészáros, Erdős, & Dosztányi, 2018). In this new implementation, we reparametrized ANCHOR on newly available datasets, such as the comprehensive database of disordered proteins DisProt (Piovesan et al., 2017) and complexes of ordered and disordered proteins (Schad et al., 2018). Furthermore, we enhanced the capabilities of our web server by adding multiple new features. Importantly, we also expanded the conditional disorder prediction capabilities of IUPred2A by incorporating a new tool to predict regions that can undergo a disorder-to-order (or order-to-disorder) transition upon change in the redox potential of the environment. While this feature is in a highly experimental stage, our high-throughput analysis of whole proteomes indicates that redox potential-dependent switches in cells are more frequent and important than previously thought (Erdős, Mészáros, Reichmann, & Dosztányi, 2019).

In this article, we give an overview of how IUPred2A can be used in different contexts and how the results can be interpreted.
from PDB structures. These regions are typically short, only 3 to 10 residues long, but have traditionally represented important test cases for disorder-prediction methods (Jin & Dunbrack, 2005). The structured-domain option aims to help the identification of larger structural units even when the prediction profile is noisy. In addition to general disorder prediction, IUPred2A also offers predictions of certain types of conditionally disordered regions such as disordered binding regions or redox-regulated disordered regions. The main advantage of using the web server is that it provides a graphical representation of the prediction output, which presents additional features that can help in the interpretation of the results.

*Necessary Resources*

*Hardware*

While IUPred2A works best on desktop or laptop computers, it is also usable on tablets and smartphones. The device must have an active internet connection.

*Software*

Any modern browser with HTML5 and WebP support (for example Chrome 30+, Firefox 65+, Opera 19+).

*Input data*

The online application of IUPred2A accepts UniProt (UniProt Consortium, 2019) identifiers (both accessions and identifier codes are allowed) or a protein sequence in plain text or FASTA format.

*Submission page*

1. Visit the web page of IUPred2A (https://iupred2a.elte.hu, Fig. 1). To analyze a protein, either use a UniProt identifier (both accession and identifier codes are allowed) or a protein sequence in the respective entry box.
   a. If the UniProt identifier or accession code is given, the web server automatically collects the corresponding amino acid sequence and runs the prediction on it.
   b. The server accepts protein sequence in FASTA format or in plain text as well.
   c. The webserver of IUPred2A is able to analyze multiple proteins or even whole proteomes of smaller size. The user can upload proteins in a multi-FASTA-formatted file to the server, and provide an e-mail address. The results will be sent as an e-mail attachment to the given address in text format (for more information on text format, see Support Protocol 2). Please note that currently the maximum file size that can be uploaded is 1 Mb.
2. Select the type of analysis.

a. As default, “IUPred2 long disorder” and the “ANCHOR2” option from the “Context-dependent predictions” submenu are selected. To carry out a disorder prediction only, uncheck the “Context-dependent predictions” option (for context-dependent disorder methods, see Basic Protocols 2 and 3).

b. Option: Use “IUPred2 short disorder” for the prediction of short disordered regions

c. Option: Use “IUPred2 structured domains” for the prediction of structured domains in the query protein sequence. If this option is selected, a textbox will appear below the graphical interface. Inside the textbox, the number of predicted globular domains and their locations will be shown alongside a string representation of the protein sequence, where capital letters represent structured domains.

The short option is recommended when the main focus of the analysis is missing residues in a PDB structure, while the structured domain option can be used to identify larger structured segments that could be suitable for structure determination. In most scenarios, the long option is recommended.

3. Click on “Submit” to start the analysis.

Results page

4. After a couple of seconds, a new page loads showing the results. The main panel on the page shows a graphical representation of the analysis. Depending on the input, we can find information about the protein from UniProt above the graphical results. Clicking this header will lead us to the corresponding UniProt page. To the right of the title, a link to download the results can be found.

5. The main graphical panel displays the prediction profile corresponding to the submitted protein sequence. Predictions are provided for each residue in the sequence. Residues with scores above 0.5 are predicted as disordered, while scores below 0.5 indicate order.

6. There are four small icons on the left-hand side of the graph that enable the manipulation of the plot. By default, the graph can be panned along the x axis by dragging the plot, or zoomed in and out using the mouse wheel. These options can be turned on and off using their respective icons. Below these icons we also have an option to reset the plot to its original state. The figure can also be saved in standard “png” format.

The save figure option will only save the main disorder propensity figure. The complementary information plots will be omitted, but can be saved using screen capture.

7. In the case where the submitted input was a UniProt accession or identifier code, the graphical interface displays information collected from additional sources, including annotations about conserved sequence families, short linear motifs, experimental evidence for disordered status, and known PDB structures corresponding to a given entry. When a plain protein sequence is submitted, IUPred2A tries to map that sequence to a protein in UniProt. If the sequence only matches one single protein in the database, its identifier will be associated with the sequence, and it will be used for further analysis. If the UniProt entry corresponding to the submitted sequence cannot be unambiguously identified, only Pfam information is displayed.

8. Sequence family information is collected from the Pfam database (El-Gebali et al., 2019) and labeled as Pfam. Different types of sequence annotations such as Domains, Families, Motifs, and Repeats are displayed using associated colors.
It is worth noting that only Domain type Pfam family annotation gives a strong indication for order. Clicking on the box leads us directly to the site of the selected annotation in the Pfam database.

9. The line ELM collects experimentally verified short linear motifs from the Eukaryotic Linear Motif database (see Current Protocols article: Gouw et al., 2018).

10. The PTM lines contain information about post-translational modification sites (PTMs) collected from the PhosphoSitePlus database (Hornbeck et al., 2012). The upper line shows a display of different phosphorylation sites with Ser, Thr, and Tyr sites indicated by different colors, while the lower one shows acetylation and methylation sites colored by the type of the modification.

11. The EXP DIS line displays experimentally verified disordered regions from the DisProt (Hatos et al., 2020), DIBS (Schad et al., 2018), and MFIB (Fichó, Reményi, Simon, & Mészáros, 2017) databases. These annotations can overlap.

12. The subplot entitled PDB shows coverage of the residues for which a structure is available in the Protein Data Bank (Berman et al., 2000). When structure hits are found for the submitted protein, a “Show all structures” checkbox is present below the PDB subplot. Marking this checkbox will display the coverage of each individual PDB entry with its respective link to the database.

13. Hovering over the complementary plot boxes displays additional information about the selected annotation or structure, such as the exact positions, identifiers, and types.

An example of IUPred2 prediction is shown in Figure 2 for the human Lupus La protein (P05455) using the “IUPred2 long disorder” option. The C-terminal region of this protein has been shown experimentally to be disordered in isolation (Jacks et al., 2003). At the N-terminal part of the protein, the Pfam annotation indicates three domains. According to IUPred2A, the intervening regions are predicted as disordered. This scenario is supported by the fact that there is at least one structure for each domain solved independently of the rest of the protein.

ANALYZING DISORDERED BINDING REGIONS USING ANCHOR2

IUPred2A now incorporates ANCHOR2, which predicts regions that undergo a disorder-to-order transition upon binding to another protein. ANCHOR2 is one of the special cases of context-dependent disorder analysis.
**Necessary Resources**

See Basic Protocol 1

1. Fill in the query protein as described in Basic Protocol 1, and either leave the “Context-dependent predictions” selection box on default or select ANCHOR2.

2. Click on “Submit.” After a few seconds, the results will appear.

3. The plot shows the disordered sequence—prediction profile (long option on default) indicated in red, as well as the predicted disordered binding sites shown in blue. Predicted disordered binding sites are positions with score above 0.5.

4. The user can turn on/off both the IUPred or ANCHOR2 profile by clicking on the corresponding label at the top right corner in the plot.

An example output in Figure 3 shows a context-dependent prediction using ANCHOR2 for the protein with the UniProt accession P45985 corresponding to the human dual-specificity mitogen-activated protein kinase kinase 4. Since disordered binding regions are expected to occur within generally disordered regions, it is important to have an overview of which regions are predicted to be disordered. The C-terminal region of this protein is ordered, and contains a kinase domain indicated by a light blue box on the figure in the Pfam line spanning from residue 102 to 367. The corresponding region also has experimentally solved structures, as shown in the PDB line with a green box. The N-terminal region of the protein is predicted as disordered, indicated by the high IUPred values (red line), in agreement with experimental characterization (Delaforge et al., 2018). Disordered binding regions predicted by ANCHOR2 are shown by the blue line. We can see a peak in the ANCHOR2 prediction around residue 46. This region corresponds to experimentally verified MAPK docking motifs, as shown in the figure with purple boxes in the ELM line (Bardwell & Bardwell, 2015; Garai et al., 2012; Ho, Bardwell, Abdollahi, & Bardwell, 2003). Note that the IUPred prediction dips slightly below the 0.5 threshold in the region corresponding to the known binding motif. This is a very common feature when predicting short linear motifs.

**INTERPRETATION OF THE RESULTS**

The interpretation of order and disorder information can be quite complex. Many proteins have modular architecture, and in reality there is a continuum of structural state which cannot be fully represented in a simple prediction output. Here we present a use case to demonstrate how complementary information can be utilized to interpret the prediction profiles.
**Figure 4** IUpred2A analysis of the human catenin beta-1 protein.

**Materials**

See Basic Protocol 1

An example output is shown in Figure 4 using UniProt accession code P35222 corresponding to the human β-catenin protein

1. IUPred2A predicts most of the N-terminal region and the C-terminal region as disordered. There is also a small peak of predicted disorder for residues 547-558.

2. ANCHOR2 predicts disordered binding regions for the N-terminal and the C-terminal regions.

3. There are four Pfam annotations indicated, which correspond to armadillo repeat regions in the middle of the sequence. It is worth keeping in mind that repeat regions are often incompletely annotated in Pfam, so there could be additional repeat units.

4. Known motifs can be found at multiple locations, according to the ELM annotations. At the N-terminus there are multiple overlapping motif sites that correspond to phosphorylation sites, and a phosphodegron motif which regulates the degradation of the protein (Wu et al., 2003). There are three kinase docking sites annotated along the sequence in addition to the C-terminal PDZ binding motif. Motif sites are generally located within disordered regions, but can also occur within loop regions of ordered domains.

5. There are multiple PTM sites annotated for this protein. They are particularly enriched in the N-terminal phosphodegron site. Phosphorylation sites are also located within the docking sites, but at additional positions along the central region as well.

6. This protein contains two experimentally verified disordered regions based on annotation from the DIBS database.

7. There are multiple PDB structures corresponding to this protein, which cover different parts of the protein. We can see this after selecting “Show structures” at the bottom of the graphical interface.

8. While the presence of structure is often interpreted as an indication of order, this protein indicates that it can be more complicated. There are multiple structures corresponding to the N-terminal region. However, this region is predicted as disordered, which is corroborated by known linear motifs, multiple PTM sites, and, primarily, by the existence of experimental evidence for disorder. Taking a closer look at the PDB structures of this region, they all contain the region of β-catenin in complex.
Figure 5  Redox potential–sensitive region prediction of the Hsp33 protein from *E. coli* using the new feature of IUPred2A.

9. The middle section also has various solved structures (e.g., 1g3j, 3tx7). These cover all armadillo Pfam regions. Based on the well-formed structures, the region between 130 and 670 can be considered as mostly ordered.

10. However, the peak of predicted disorder is confirmed by the lack of electron density in the known structures. See https://www.rcsb.org/pdb/protein/P35222?addPDB=3SLA for more details.

11. The C-terminal region has PDB coverage from a single PDB structure (PDB code 2Z6H), and, on closer inspection, this region is actually missing from the electron density. Therefore, the structural annotation does not contradict the disorder status for the last 120 residues.

**ANALYZING REDOX-SENSITIVE DISORDERED REGIONS**

The new implementation of IUPred2A contains an additional feature to predict regions that undergo a disorder-to-order transition upon changes of the redox potential in the environment. The key residues of redox regulation are cysteines, which can provide large stabilizing effects through disulfide bond formation under oxidative conditions or through zinc ion coordination under reducing conditions. The lack of these stabilizing contributions can be modeled by mutating all cysteine residues to serines in the sequence. In most cases, the two scenarios are not significantly different. However, in the case of redox-sensitive disordered regions, redox_plus is expected to be mostly ordered, while redox_minus is expected to be mostly disordered. The actual region is calculated based on a heuristic segmentation approach (Mészáros et al., 2018).

**Necessary Resources**

See Basic Protocol 1

1. Fill in the query protein as described in Basic Protocol 1, and use the “Context-dependent predictions” selection box to select “Redox state.”

2. Click on “Submit.” After a few seconds, the results will appear. The plot shows two lines corresponding to redox_plus and redox_minus scenarios. An identified redox-sensitive disordered region will be indicated by shading the region between the two lines, and will also be highlighted by a box below the graph in a complementary plot entitled REDOX.
One of the best examples of a redox-sensitive disordered region is the Hsp33 chaperone protein from E. coli. Under normal circumstances, this Zn$^{2+}$ ion-binding protein is completely structured without activity. Under oxidative stress, the protein undergoes a conformational change due to the formation of intermolecular disulfide bonds. This leads to the release of the bound metal ion, and to the unfolding of the zinc-binding domain, exposing the substrate-binding surface that is required for the chaperone activity (Reichmann et al., 2012). This redox potential-sensitive switch can be found between residues 224 to 285, part of which is correctly predicted by IUPred2A, as shown in Figure 5.

DOWNLOAD OPTIONS

In addition to the graphical output, IUPred2A offers multiple formats in which to download the results. On each analysis, a button called “Download results” appears to the right of the header. Clicking on the button offers two options: Text or JSON. Choosing the “Text” button shows a plain text-formatted version of the analysis results. Each line in the header starts with a “#” symbol, to separate them from the relevant data. The results represent each residue on a separate line. Each line is split into columns. The first column contains the position of the given residue in the sequence; the second shows the one-letter IUPAC representation of the residue. The remaining columns are based on the type of analysis. They will show the disordered prediction score (IUPred2A), the prediction scores for IUPred and ANCHOR (ANCHOR2), or the Redox plus and Redox minus predictions (Redox). In the case of redox state calculations, an additional column is present with a binary classification. A 1 shows that the residue is within a redox potential-sensitive region, while a 0 represents the contrary.

Choosing the “JSON” option leads to a JSON representation of the results. This representation always contains a “sequence” key, with the protein sequence as value, as well as a “type” key with the corresponding analysis type as value. The rest of the keys are dependent on the type of the analysis. For IUPred2 calculations, the key “iupred2” will contain a list representation of IUPred2 scores; for context-dependent calculations, the “anchor2” key contains the ANCHOR2 scores; “iupred2_redox_plus” and “iupred2_redox_min” contain the redox state result scores; and the key “redox_sensitive_regions” shows the region boundaries in a list. The value of the “meta” key contains information about the publication of the method.

REST API FOR PROGRAMMATIC PURPOSES

IUPred2A is also accessible via REST API to enable automated/large-scale use. Requests should follow the syntax: https://iupred2a.elte.hu/iupred2a/::accession:: or https://iupred2a.elte.hu/iupred2a/::analysis_type::/::accession:: where ::accession:: must be a valid UniProt accession and ::analysis_type:: is either “long,” “short,” “glob,” “anchor,” or “redox,” depending on the type of analysis. In cases where “::analysis_type::” is not given, the default “long” will be used. If the requested URL ends with “.json” the output will be JSON type; in any other case it will be simple text.

For example:
https://iupred2a.elte.hu/iupred2a/q32p44
https://iupred2a.elte.hu/iupred2a/q32p44.json
https://iupred2a.elte.hu/iupred2a/short/q32p44

Here you can find a simple python script that shows the programmatic use of the IUPred2A REST API.
**Requirements**

Python3.6, python3-requests library version 2.22.0, python3-json library version 2.0.9

**Sample code**

```python
import requests
import json
URL = "https://iupred2a.elte.hu/iupred2a/anchor/q32p44.json"
response = json.loads(requests.get(URL).text)
print("\n".join(["{}	{}".format(response["sequence"]
[n], x) for n, x in enumerate(response["anchor2"])]))
```

**Results**

This example loads the results of an ANCHOR2 analysis for the human EML3 protein (Q32P44), and loads the JSON-formatted results as standard python dictionary into the variable “response.”

**USING IUPred2A LOCALLY**

The IUPred2A website offers a downloadable package (in .tar.gz format) for local use, which is free for academic users. Along with the IUPred2A software, the experimental “IUPred2A redox” can be downloaded as well. Both packages contain an executable file together with a python library for programmatic use.

**Necessary Resources**

**Hardware**

There are no definite hardware requirements for IUPred2A. It uses minimal amount of memory, and has no GPU requirements.

**Software**

- Python 3.x runtime environment
- Libraries: os, math, textwrap. These libraries are a part of the default Python 3 installation.

**Input data**

A protein sequence in either FASTA or plain text format

The downloaded package requires no installation process. To be able to use it, simply unpack the compressed directory. Depending on your choice, the executable file will either be called “iupred2a.py” for the standard version or “iupred2_redox.py” for the redox potential region prediction. To start a prediction, simply call the executable file of your choice in a terminal and pass the location of the protein sequence as an argument along with the type of prediction you want to carry out (either “long,” “short,” or “glob” for the prediction of long or short disorder or globular regions, respectively). The main package also contains the example sequence for the human p53 protein in FASTA format in a file called “P53_HUMAN.seq.”

**Example 1**

```
$ python3 iupred2a.py P53_HUMAN.seq long
```

Expected output:

```
# IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding
```
The output will be displayed on a table-like structure inside your chosen terminal in text format (for more information on display formats see Support Protocol 2).

To carry out an ANCHOR2 analysis, use the -a switch.

**Example 2**

```
$ python3 iupred2a.py -a P53_HUMAN.seq long
```

Expected output:

```
# IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding
# Balint Meszaros, Gabor Erdos, Zsuzsanna Dosztanyi
# Nucleic Acids Research 2018;46(W1):W329-W337.
#
# Prediction type: long
# Prediction output
# POS RES IUPRED2 ANCHOR2
1 M 0.9807 0.8488
2 E 0.9849 0.8331
3 E 0.9875 0.8187
4 P 0.9693 0.8124
5 Q 0.9657 0.8040
...
```

**Option:** IUPred2A requires multiple data files, which are a part of the downloadable package, and can be found inside the “data” folder. By default, this folder is at the same directory level as the executable; however, if you want to modify the structure, you can specify the location of the data files by using the -d switch.

Example:

```
$ python3 iupred2a.py -d /path/to/iupred2a/data/directory -a P53_HUMAN.seq long
```

For programmatic usage, the downloaded package also comes with a python library package called “iupred2a_lib.py.” There are multiple approaches to using this library. One is to include its absolute path in the PYTHONPATH environment variable on LINUX systems, or their equivalent in the python interpreter on Windows- or Mac-based machines. Once the path has been added to the environment variable, we can simply import the library and use the functions inside. The library contains four major functions and some minor helper functions. The first function is called “iupred.” It requires as an argument a protein sequence in a format of a standard python string. Note that this string
must only contain pure uppercase one-letter amino acid code, and not a FASTA-formatted sequence. The function also takes an optional argument called “mode,” where we can specify the type of analysis. The accepted options are “long,” “short,” or “glob” for the prediction of long or short disordered regions or structured domains, respectively. The default option is “long.” The function returns the predicted values in the form of a list of floating-point numbers. Note that if the “glob” optional argument is used, the function will return a tuple, where the first element will be the prediction scores while the second will be the string representation of the structured domains that we can see on the website, as described in Basic Protocol 1.

**Example 3**

```python
import iupred2a_lib
my_sequence = "MEEPQSDEPSVEPLSQTFSDDLWLLPPENNVLSPALPSQAMDDMLSPDDIEQ"  
WFTEDPGPDEAPRMEAPVAPAAPTAPAAPAPASWPLLSSVPSQKYTY"  
QGSYGFRGFLHSTAKTVSPTSPNLKMFCQLAKTCPVQLWVDSTPPFG"  
TRVYRAAIYQSQHMTEVRRCPHHERCSSDSGLAPQHLIREVGLRNVEY"  
LDDNTRPHSVEPEPVEGSCTTIHYNCNSCMSGGMNRRPILITIT"  
LEDSSGNLQGRSFVRCACPGDRRTEENRKKGEHPEHLPFGSTKA"  
LPNTTSSSPQPKKPRLGDGEYFTLQIRGRERVFMFRELNEALELKDAQAGKE"  
PFGSRAHSHKLKSGQSTSRHKKMLFKTEGPDS"
print(iupred2a_lib.iupred(my_sequence))
```

The function called “anchor2” works in the same fashion as “iupred.” It takes a sequence as input, and returns the predicted ANCHOR2 values in a list of floating-point numbers.

**Example 4**

```python
import iupred2a_lib
my_sequence = "MEEPQSDEPSVEPLSQTFSDDLWLLPPENNVLSPALPSQAMDDMLSPDDI"  
EQWFTEDPGPDEAPRMEAPVAPAAPTAPAAPAPASWPLLSSVPSQKTV"  
QGSYGFRGFLHSTAKTVSPTSPNLKMFCQLAKTCPVQLWVDSTPPFG"  
TRVYRAAIYQSQHMTEVRRCPHHERCSSDSGLAPQHLIREVGLRNVEY"  
LDDNTRPHSVEPEPVEGSCTTIHYNCNSCMSGGMNRRPILITIT"  
LEDSSGNLQGRSFVRCACPGDRRTEENRKKGEHPEHLPFGSTKA"  
LPNTTSSSPQPKKPRLGDGEYFTLQIRGRERVFMFRELNEALELKDAQAGKE"  
PFGSRAHSHKLKSGQSTSRHKKMLFKTEGPDS"
print(iupred2a_lib.anchor2(my_sequence))
```

This library also contains a function to predict the experimental redox-sensitive disordered regions. The main function is called “get_redox_regions,” which requires two arguments. The first must be a list of floating-point numbers from an iupred prediction where the cysteine residues have been switched to serines (the result of the helper function “iupred2_redox”), and the second must be the result of a standard iupred prediction (the function “iupred2”). The results will be represented in a standard python dictionary, where each key is the starting position of the region and each value is the end of the corresponding starting point.

**Example 5**

```python
import iupred2a_lib
my_sequence = "Erdos and Dosztanyi"
```
"MEEPQSDPSVEPPLSQETFSDKLPLLNPENVSLPSQAMDDLMLSPDD
IEQWFTEDPGDEAPRMPEAAPPVAPAAPAPTAPAPAPAPSPWSPLSSV
PSQKTYQSQ4GRLHSLGTASKVCTYSPALNMFCQLAKTCPVQL
WVDSTPPGTVRAMAIYKQSMTECVRRCPHHERCSDSDGLAPPQH
LIRVEGNLRVYLDNRNTFHRHSVYPYEPPEGSDCTTITYNMCNSS
CMGMNRRPILTIITLEDSSGNLLGRNSFEVRCACPGDRRTEEEENL
RRKGEFHEHLPFSSTKRALPPNNTSSSPQFKKKPLDGEYFTLQIRGQER
FEMFRELNEALELKDQAQAGKEPGGSRAHSKLKSKKGQLSRHKKLMF
KTEGFPSD"

iupred_result = iupred2a_lib.iupred(my_sequence)
iupred_redox = iupred2a_lib.iupred_redox(my_sequence)
print(iupred2a_lib.get_redox_regions(iupred_redox,
iupred_result))

**Example 6: Genome-wide analysis**

Using the downloaded package, we can carry out a genome-wide analysis using IUPred2A. The following example will download the entire human proteome from the UniProt database and calculate the frequency of disordered residues and frequency of proteins that have at least one disordered segment longer than 30 residues in the proteome. Please note that this example may take more than 5 min, depending on the computer.

**Hardware**

Internet connection

**Software**

Python 3.x

Libraries: iupred2a_lib, gzip, ftplib, itertools

```python
import ftplib
import gzip
import iupred2a_lib
from itertools import groupby
ftp = ftplib.FTP("ftp.uniprot.org")
ftp.login()
ftp_file = "pub/databases/uniprot/current_release/\nknowledgebase/reference_proteomes/Eukaryota/\nUP000005640_9606.fasta.gz"
ftp.retrbinary("RETR" + ftp_file,
open("human_proteome.fasta", "wb").write)
proteome = {}
header ="
with gzip.open("human_proteome.fasta", mode="rt") as file_handler:
    for line in file_handler:
        if line.startswith(""):
            header = line
            proteome[header] = ""
        elif header:
            proteome[header] += line.strip()
        disordered_residues = 0
        disordered_proteins = 0
        all_residues = 0
```
for header, seq in proteome.items():
    iupred_scores = iupred2a_lib.iupred(seq, "long")
    if max([sum([1 for _ in y]) if x == 1 else 0 for x, y in groupby([1 if x>=0.5 else 0 for x in iupred_scores])]) > 30:
        disordered_proteins += 1
    disordered_residues += sum([1 for x in iupred_scores if x>=0.5])
    all_residues += len(seq)
print("Fraction of disordered residues:
{:.2f}".format(disordered_residues/all_residues))
print("Fraction of disordered proteins:
{:.2f}".format(disordered_proteins/len(proteome)))

Acknowledgments
This work was completed in the ELTE Thematic Excellence Programme (ED-18-1-2019-0030) supported by the Hungarian Ministry for Innovation and Technology. This project has received funding from the European Union’s Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 778247. E. G. also acknowledges funding from the Joseph Cours Scholarship.

Literature Cited
Bardwell, A. J., & Bardwell, L. (2015). Two hydrophobic residues can determine the specificity of mitogen-activated protein kinase docking interactions. The Journal of Biological Chemistry, 290(44), 26661–26674. doi: 10.1074/jbc.M115.691436.
Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., … Bourne, P. E. (2000). The protein data bank. Nucleic Acids Research, 28(1), 235–242. doi: 10.1093/nar/28.1.235.
Borgia, A., Borgia, M. B., Bugge, K., Kissling, V. M., Heidarsson, P. O., Fernandes, C. B., … Schuler, B. (2018). Extreme disorder in an ultrahigh-affinity protein complex. Nature, 555(7694), 61–66. doi: 10.1038/nature25762.
Delaforge, E., Kragelj, J., Tengo, L., Palencia, A., Milles, S., Bouvignies, G., … Jensen, M. R. (2018). Deciphering the dynamic interaction profile of an intrinsically disordered protein by NMR exchange spectroscopy. Journal of the American Chemical Society, 140(3), 1148–1158. doi: 10.1021/jacs.7b12407.
Dosztányi, Z., Csizmok, V., Tompa, P., & Simon, I. (2005). The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. Journal of Molecular Biology, 347(4), 827–839. doi: 10.1016/j.jmb.2005.01.071.
El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., … Finn, R. D. (2019). The Pfam protein families database in 2019. Nucleic Acids Research, 47(D1), D427–D432. doi: 10.1093/nar/gky995.

Erdős, G., Mészáros, B., Reichmann, D., & Dosztányi, Z. (2019). Large-scale analysis of redox-sensitive conditionally disordered protein regions reveals their widespread nature and key roles in high-level eukaryotic processes. Proteomics, 19(6), e1800070. doi: 10.1002/pmic.201800070.
Fichó, E., Reményi, I., Simon, I., & Mészáros, B. (2017). MFIB: A repository of protein complexes with mutual folding induced by binding. Bioinformatics, 33(22), 3682–3684. doi: 10.1093/bioinformatics/btx486.
Garai, Á., Zeke, A., Gógl, G., Türö, I., Fördős, F., Blankenburg, H., … Reményi, A. (2012). Specificity of linear motifs that bind to a common mitogen-activated protein kinase docking groove. Science Signaling, 5(245), ra74. doi: 10.1126/scisignal.2003004.
Gouw, M., Michael, S., Sámano-Sánchez, H., Kumar, M., Zeke, A., Lang, B., … Gibson, T. J. (2018). The eukaryotic linear motif resource—2018 update. Nucleic Acids Research, 46(D1), D248–D344. doi: 10.1093/nar/gkx1077.
Gouw, M., Sámano-Sánchez, H., Van Roey, K., Diella, F., Gibson, T. J., & Dinkel, H. (2017). Exploring short linear motifs using the ELM database and tools. Current Protocols in Bioinformatics, 58, 8.22.1–8.22.35.
Hatos, A., Hajdu-Soltész, B., Monzon, A. M., Palopoli, N., Álvarez, L., Aykac-Fas, B., … Piovesan, D. (2020). DisProt: Intrinsic protein disorder annotation in 2020. Nucleic Acids Research, 48(D1), D269–D276.
Ho, D. T., Bardwell, A. J., Abdollahi, M., & Bardwell, L. (2003). A docking site in MKK4 mediates high affinity binding to JNK MAPKs and competes with similar docking sites in JNK substrates. The Journal of Biological Chemistry, 278(35), 32662–32672. doi: 10.1074/jbc.M304229200.
Hornbeck, P. V., Kornhauser, J. M., Tkachev, S., Zhang, B., Skrzypek, E., Murray, B., … Sullivan, M. (2012). PhosphoSitePlus: A comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. Nucleic Acids Research, 40(Database issue), D261–D270. doi: 10.1093/nar/gkr1122.
Jacks, A., Babon, J., Kelly, G., Manolaridis, I., Cary, P. D., Curry, S., & Conte, M. R. (2003). Structure of the C-terminal domain of human Laprotein reveals a novel RNA recognition motif coupled to a helical nuclear retention element. *Structure, 11*(7), 833–843. doi: 10.1016/S0969-2126(03)00121-7.

Jakob, U., Kriwacki, R., & Uversky, V. N. (2014). Conditionally and transiently disordered proteins: Awakening cryptic disorder to regulate protein function. *Chemical Reviews, 114*(13), 6779–6805. doi: 10.1021/cr400459c.

Jin, Y., & Dunbrack, R. L., Jr. (2005). Assessment of disorder predictions in CASP6. *Proteins, 61*(Suppl 7), 167–175. doi: 10.1002/prot.20734.

Mészáros, B., Erdos, G., & Dosztányi, Z. (2018). IUPred2A: Context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Research, 46*(W1), W329–W337. doi: 10.1093/nar/gky384.

Mészáros, B., Simon, I., & Dosztányi, Z. (2009). Prediction of protein binding regions in disordered proteins. *PLoS Computational Biology, 5*(5), e1000376. doi: 10.1371/journal.pcbi.1000376.

Miskei, M., Antal, C., & Fuxreiter, M. (2017). FuzDB: Database of fuzzy complexes, a tool to develop stochastic structure-function relationships for protein complexes and higher-order assemblies. *Nucleic Acids Research, 45*(D1), D228–D235. doi: 10.1093/nar/gkw1019.

Piovesan, D., Tabaro, F., Mičetić, I., Necci, M., Quaglia, F., Oldfield, C. J., … Tosatto, S. C. E. (2017). DisProt 7.0: A major update of the database of disordered proteins. *Nucleic Acids Research, 45*(D1), D219–D227. doi: 10.1093/nar/gkw1056.

Reichmann, D., Xu, Y., Cremers, C. M., Ilbert, M., Mittelman, R., Fitzgerald, M. C., & Jakob, U. (2012). Order out of disorder: Working cycle of an intrinsically unfolded chaperone. *Cell, 148*(5), 947–957. doi: 10.1016/j.cell.2012.01.045.

Schad, E., Fichő, E., Pancsa, R., Simon, I., Dosztányi, Z., & Mészáros, B. (2018). DIBS: A repository of disordered binding sites mediating interactions with ordered proteins. *Bioinformatics, 34*(3), 535–537. doi: 10.1093/bioinformatics/btx640.

UniProt Consortium. (2019). UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Research, 47*(D1), D506–D515. doi: 10.1093/nar/gkx1049.

van der Lee, R., Buljan, M., Lang, B., Weatheritt, R. J., Daughdrill, G. W., Dunker, A. K., … Babu, M. M. (2014). Classification of intrinsically disordered regions and proteins. *Chemical Reviews, 114*(13), 6589–6631. doi: 10.1021/cr400525m.

Wright, P. E., & Dyson, H. J. (1999). Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm. *Journal of Molecular Biology, 293*(2), 321–331. doi: 10.1006/jmbi.1999.3110.

Wu, G., Xu, G., Schulman, B. A., Jeffrey, P. D., Harper, J. W., & Pavletich, N. P. (2003). Structure of a beta-TrCP1-Skp1-beta-catenin complex: Destruction motif binding and lysine specificity of the SCF(beta-TrCP1) ubiquitin ligase. *Molecular Cell, 11*(6), 1445–1456. doi: 10.1016/S1097-2765(03)00234-X.