# Intrinsically disordered proteins

## Prediction and annotation

# How common is protein disorder?

- Around 50% of human proteins have long disordered regions
- Around 30% of residues in the human proteome are predicted as disordered
- Disorder content increases with evolutionary complexity

Eukaryotes > Prokaryotes

# Where can we find disordered proteins?

In the literature

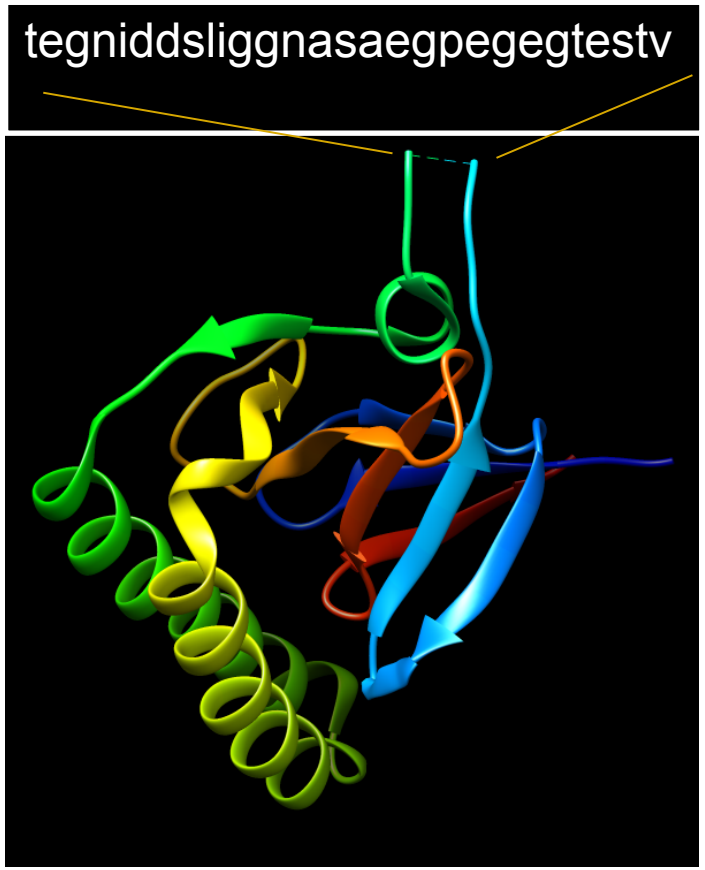
Failed attempts to crystallize
Lack of NMR signals
Heat stability
Protease sensitivity
Increased molecular volume
“Freaky” sequences …

***Disprot database:***

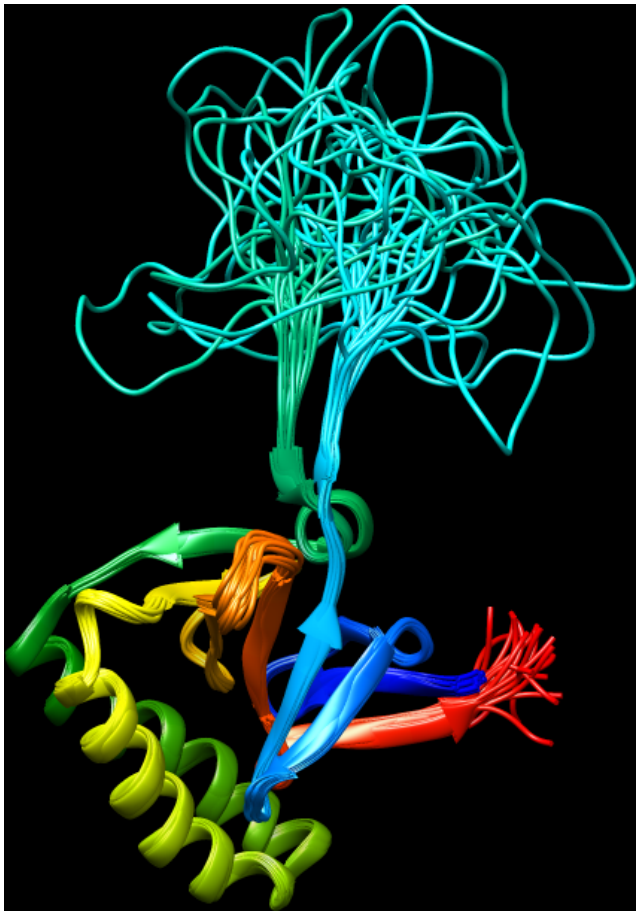
www.disprot.org

# Where can we find disordered proteins?

In the PDB



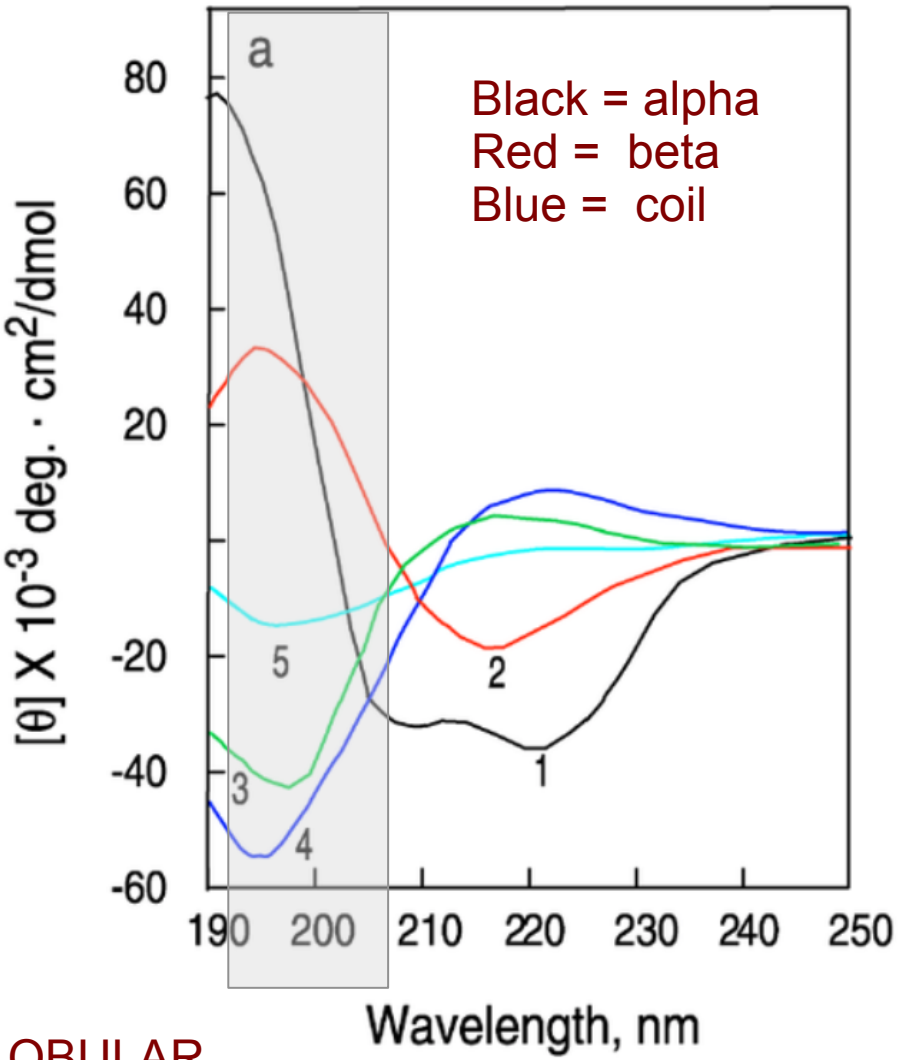

Missing electron density regions from the PDB

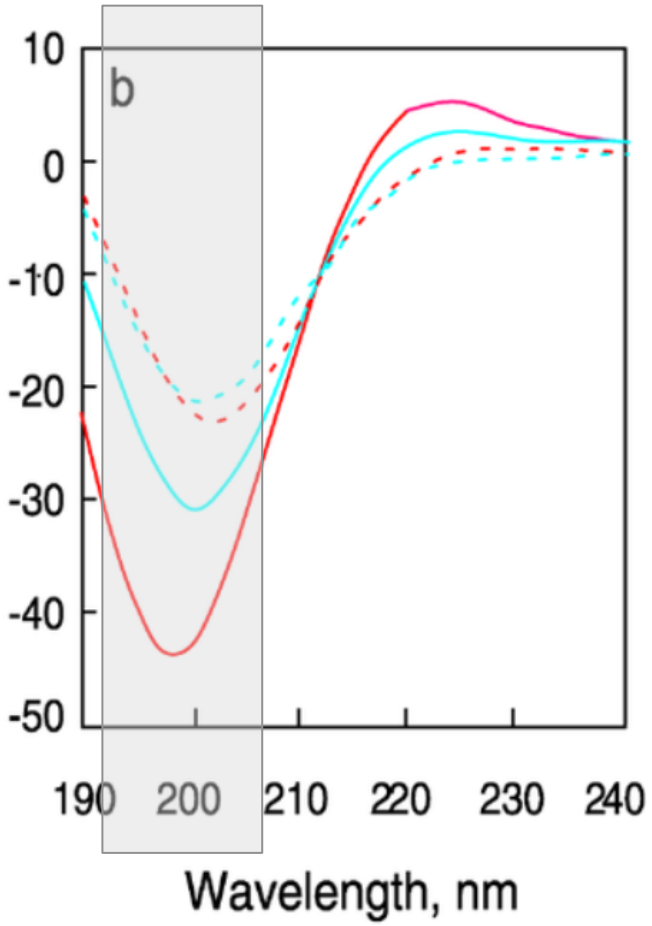


NMR structures with large structural variations

# Lack of regular secondary structure

## Far-UV Circular Dichroism (CD)




GLOBULAR

Positive signal at ~190nm

Minimum at ~210-220nm

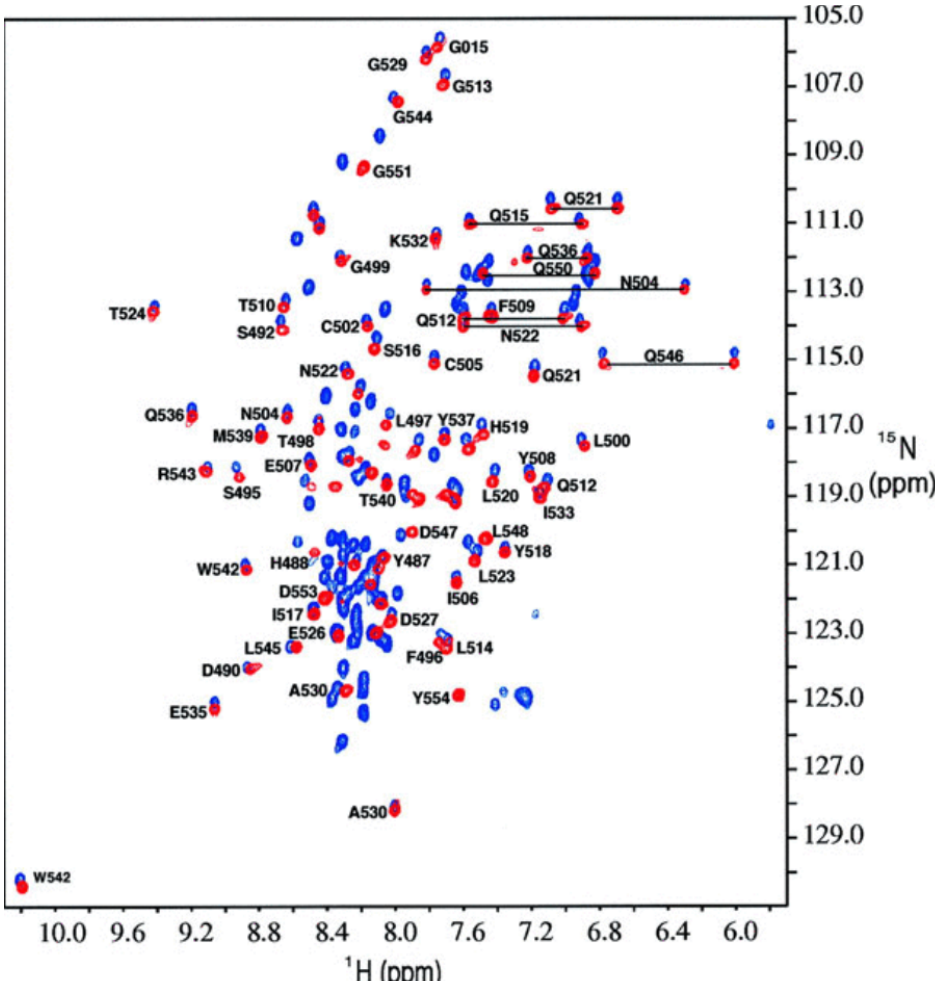
DISORDERED

Minimum at ~200nm

Low signal at 220nm

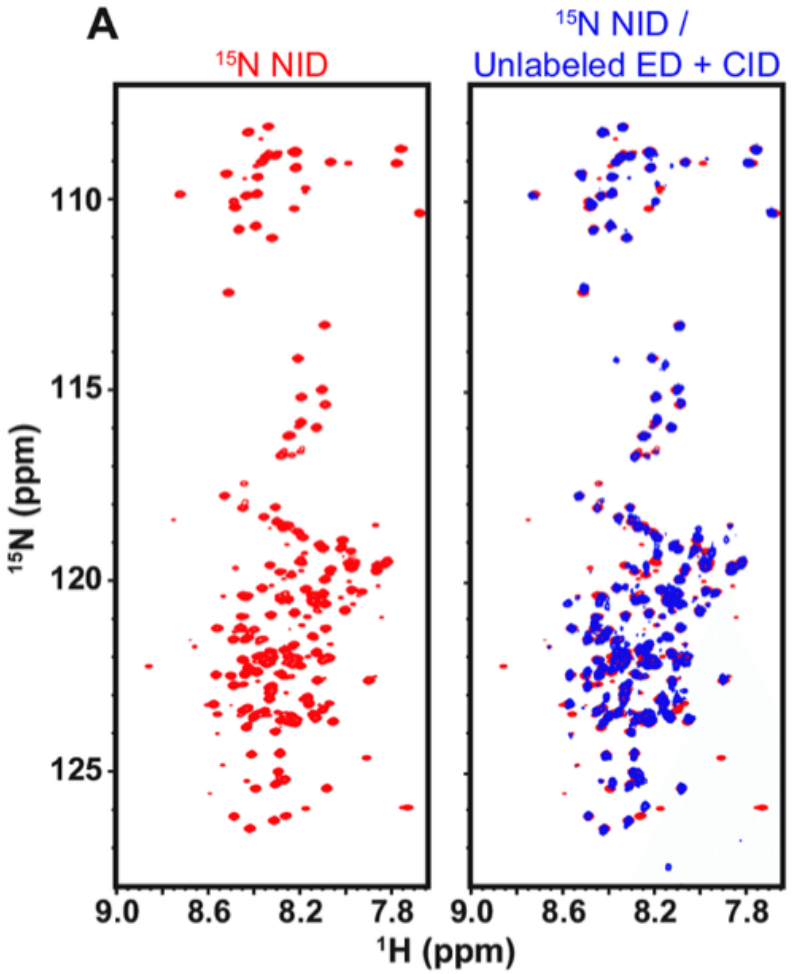
# Poor Dispersion in NMR experiments

## Globular




~ 4 ppm

## Disordered



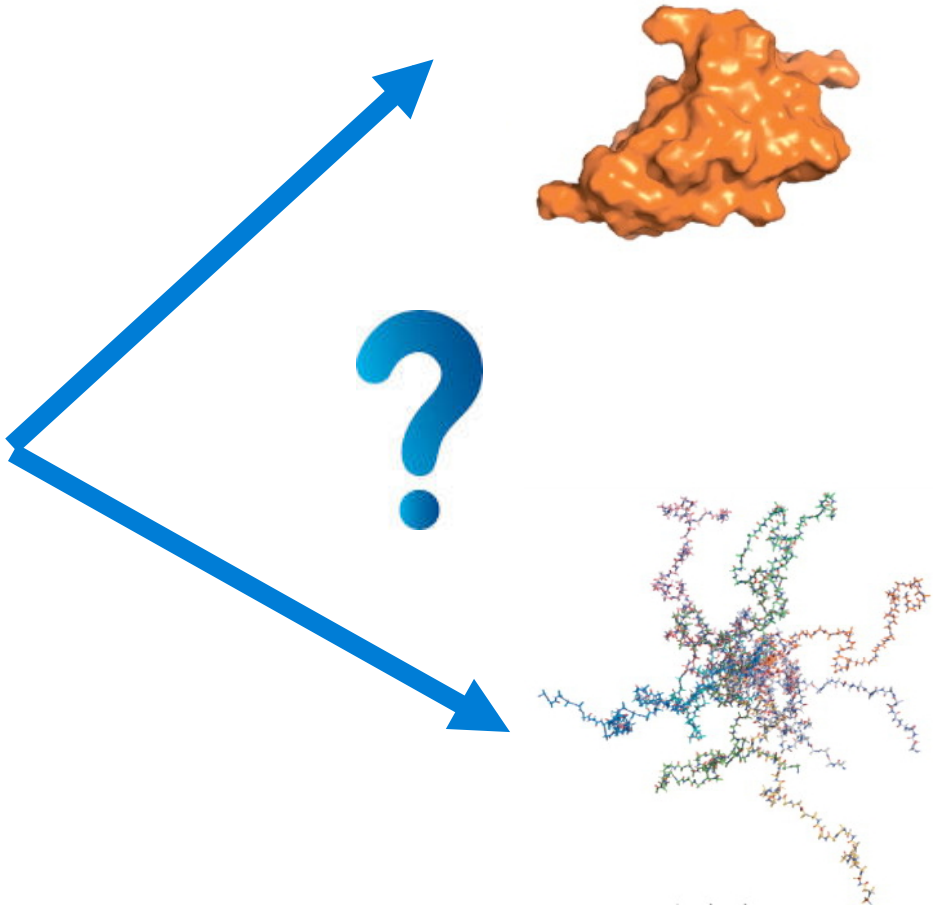

~ 1 ppm

# Sequence properties of IDPs

- Amino acid compositional bias
- High proportion of polar and charged amino acids (Gln, Ser, Pro, Glu, Lys)
- Low proportion of bulky, hydrophobhic amino acids (Val, Leu, Ile, Met, Phe, Trp, Tyr)
- Low sequence complexity
- Signature sequences identifying disordered proteins

# Protein disorder is encoded in the amino acid sequence

**TDVEAAVNSLVNLYLQAS**
**YLS**



***How can we discriminate ordered and disordered regions ?***

# Prediction of protein disorder

***Can we discriminate ordered and disordered regions ?***

- Training sets:
  - Ordered structures come from the PDB
  - Short and Long disorder
    - PDB (L<30)
    - DisProt (L>=30)

    *The two types of datasets differ not just in their lengths*
  - Training sets are small
  - Unbalanced datasets

# Prediction of protein disorder

- Disordered residues can be predicted from the amino acid sequence
  - ~ 80% at the residue level
- Methods can be specific to certain type of disorder
  - accordingly, accuracies vary depending on datasets

# Prediction methods for protein disorder

Over 50 methods …

- Based on amino acid propensity scales or on simplified biophysical models
  - **GlobPlot**, FoldIndex, FoldUnfold, **IUPred**, UCON, **TOP-IDP**

- Machine learning approaches
  - PONDR VL-XT, VL3, **VSL2, FIT**; Disopred; POODLE S and L ; DisEMBL; DisPSSMP; PrDOS, DisPro, OnD-CRF, POODLE-W, RONN, …

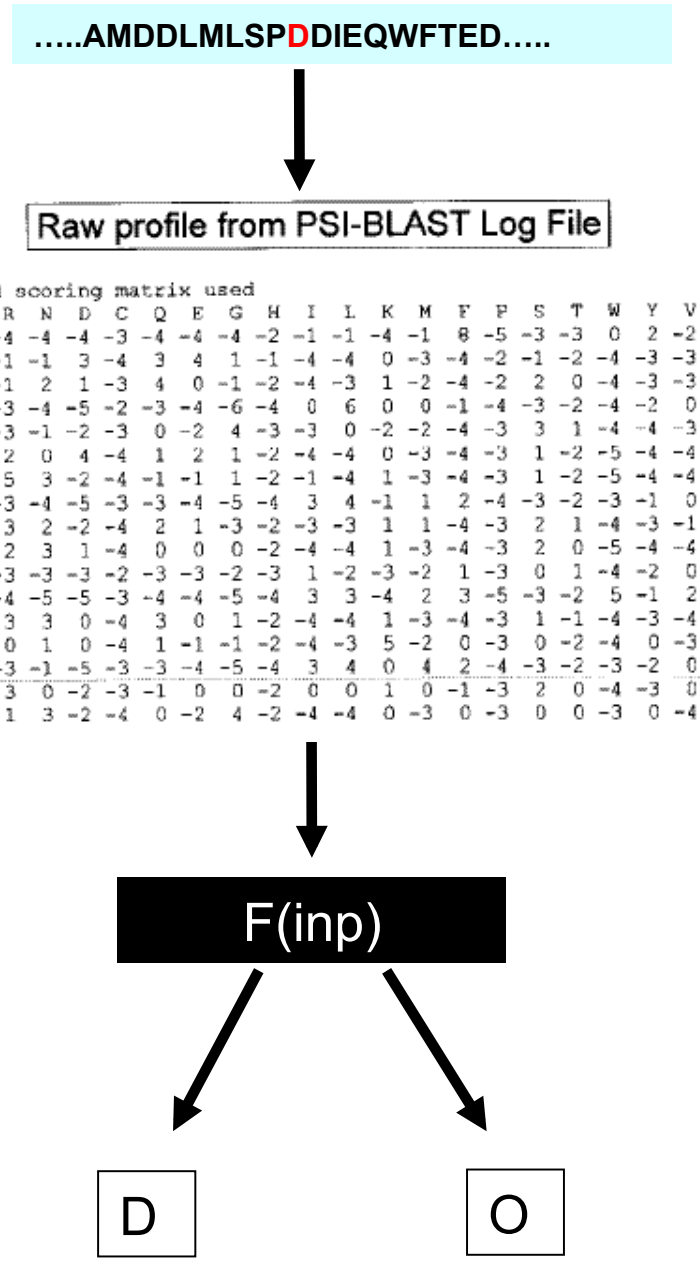
# DISOPRED
# PONDR

Trained in missing residues from X-ray structures

Trained with NMR data

SVM with linear kernel

Assign label: D or O




Ward et al. J.Mol. Biol. 2004; 337:653

# IUPred

- Globular proteins form many favorable interactions to ensure the stability of the structure
- Disordered protein cannot form enough favourable interactions

Energy estimation method

Based on globular proteins

No training on disordered proteins

Dosztanyi (2005) JMB 347, 827

PROTEIN

## Structure

```
MODEL        1
ATOM      1  N   MET A  23       2.191  28.312  -4.381
ATOM      2  CA  MET A  23       2.394  27.327  -3.305
ATOM      3  C   MET A  23       3.514  26.377  -3.706
ATOM      4  O   MET A  23       3.589  25.977  -4.867
ATOM      5  CB  MET A  23       1.128  26.503  -3.033
ATOM      6  CG  MET A  23       0.025  27.305  -2.344
ATOM      7  SD  MET A  23      -1.456  26.318  -2.038
ATOM      8  CE  MET A  23      -2.566  27.602  -1.402
ATOM      9 1H   MET A  23       2.034  27.828  -5.254
ATOM     10 2H   MET A  23       1.397  28.910  -4.199
ATOM     11 3H   MET A  23       3.017  28.882  -4.497
```

Calculated energy per residue

## Sequence

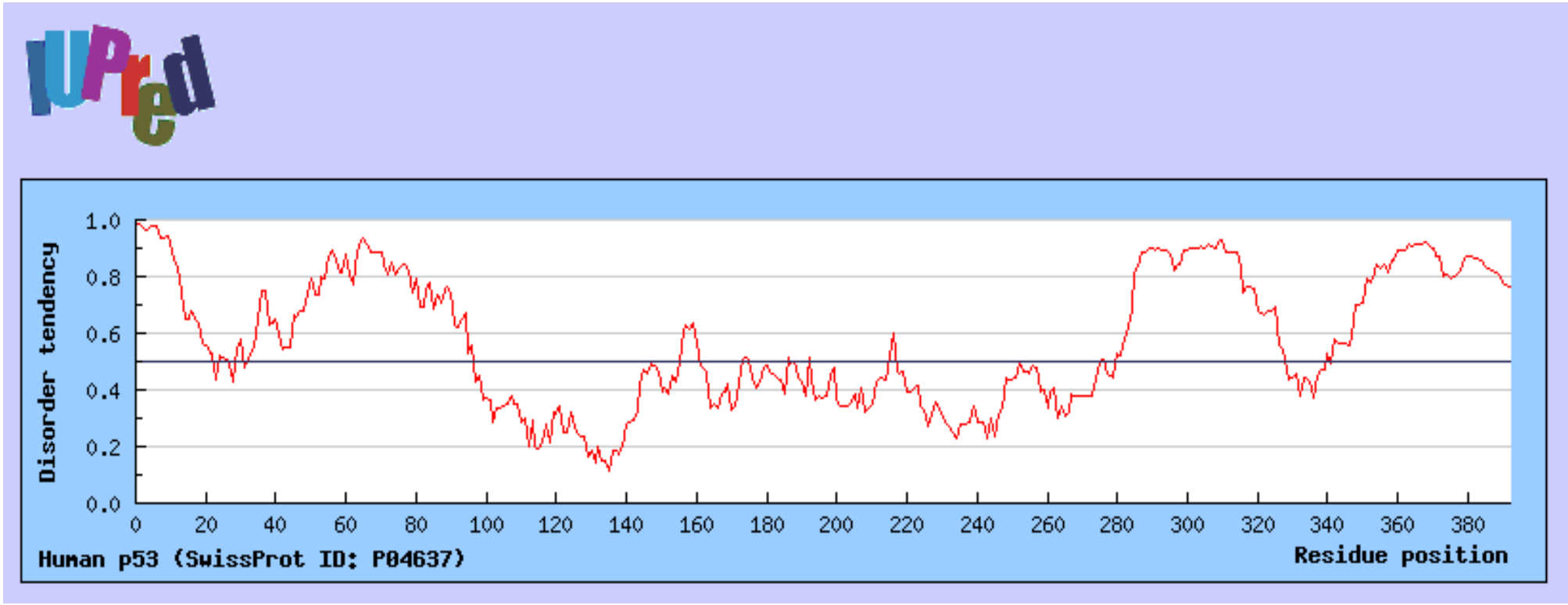
```
MKVPPHSIEA EQSVLGGLML
DNERWDDVAE RVVADDFYTR
PHRHIFTEMA RLQESGSPID
LITLAESLER QGQLDSVGGF
AYLAELSKNT PSAANISAYA
DIVRERAVVR EMIS
```

Amino acid composition (***n***)

| | |
|---|---|
| A | 10.5 |
| C | 0.0 |
| D | 7.0 |
| E | 9.6 |
| F | 2.6 |
| G | 5.3 |
| H | 2.6 |
| I | 6.1 |
| K | 1.8 |
| L | 8.8 |
| M | 3.5 |
| N | 2.6 |
| P | 4.4 |
| Q | 3.5 |
| R | 7.9 |
| S | 8.8 |
| T | 3.5 |
| V | 7.9 |
| W | 0.9 |
| Y | 2.6 |

Estimated energy per residue

# Typical output



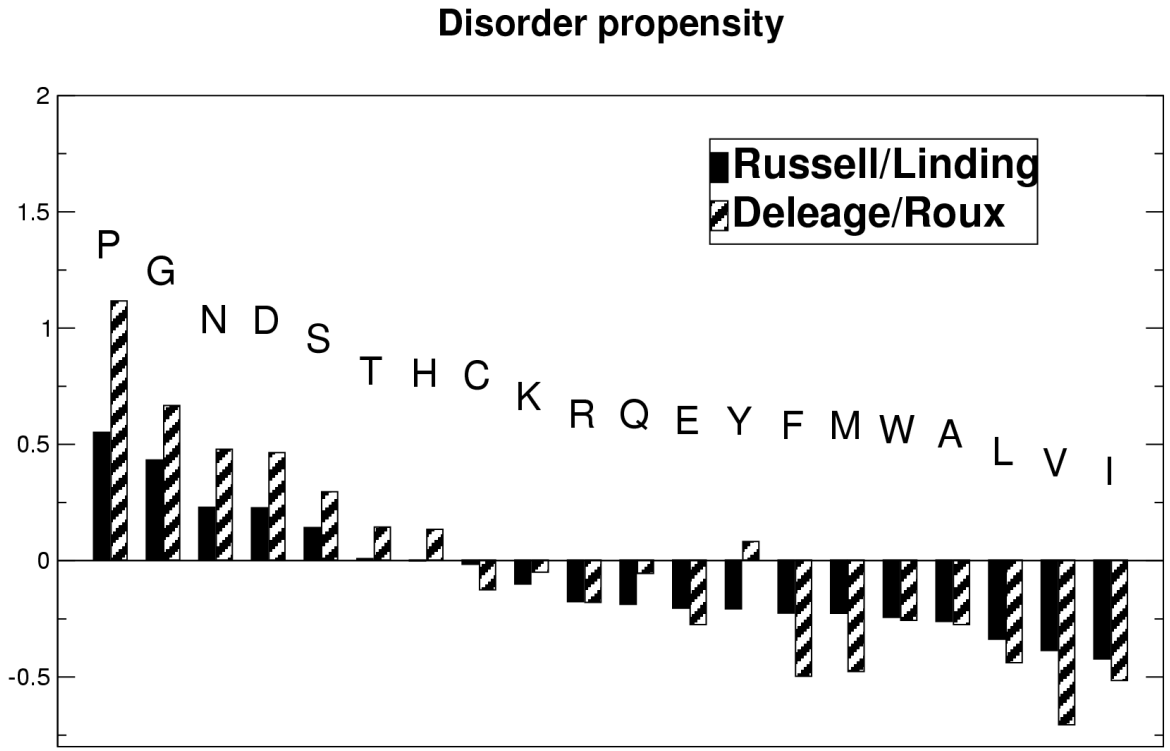

IUPRED > 0.4-0.5 = DISORDERED

# GlobPlot

Globular proteins form regular secondary structures, and different amino acids have different propensity to be in them

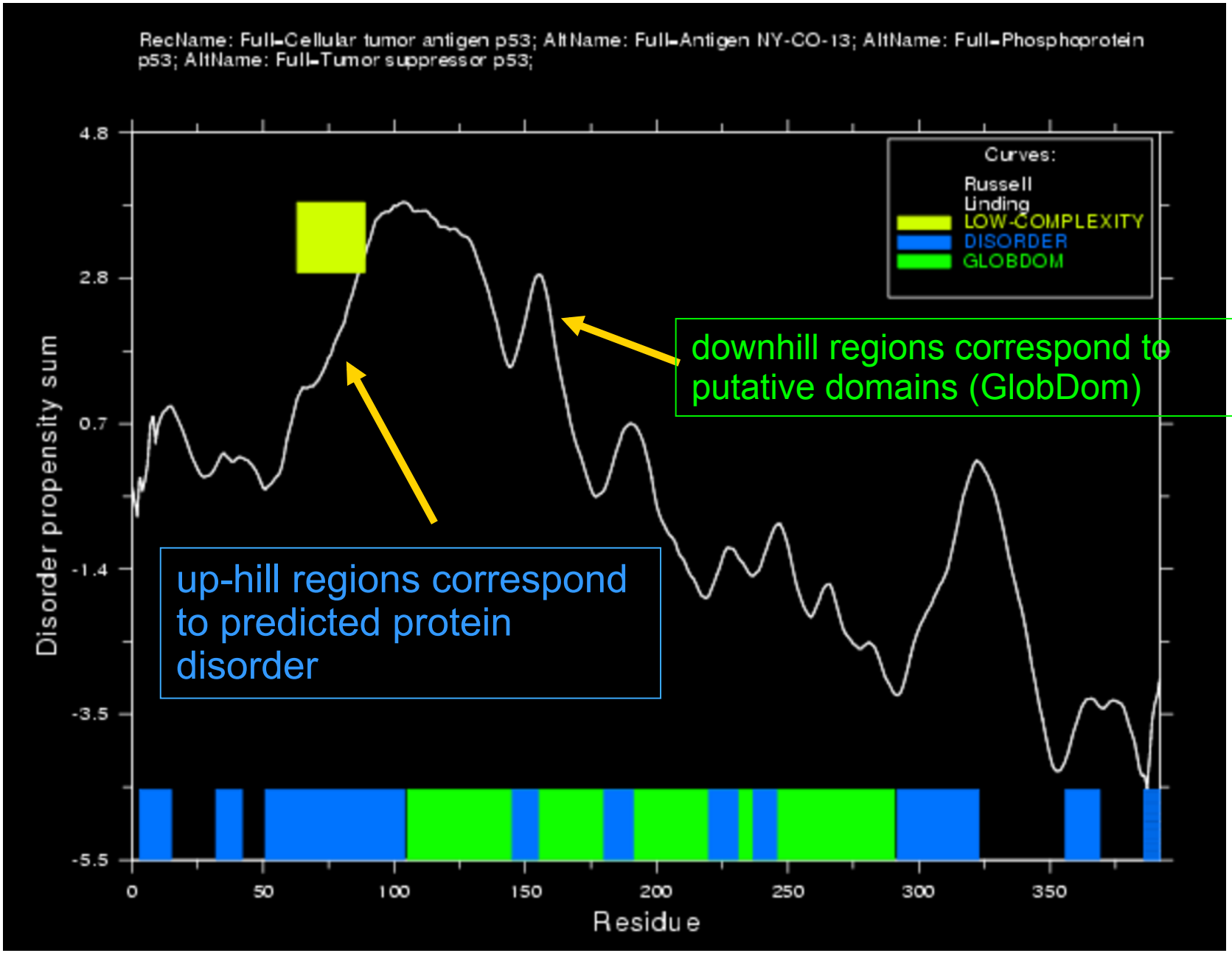
Compare the propensity of amino acids:

- to be in coil (irregular) structure.
- to be in regular secondary structure elements

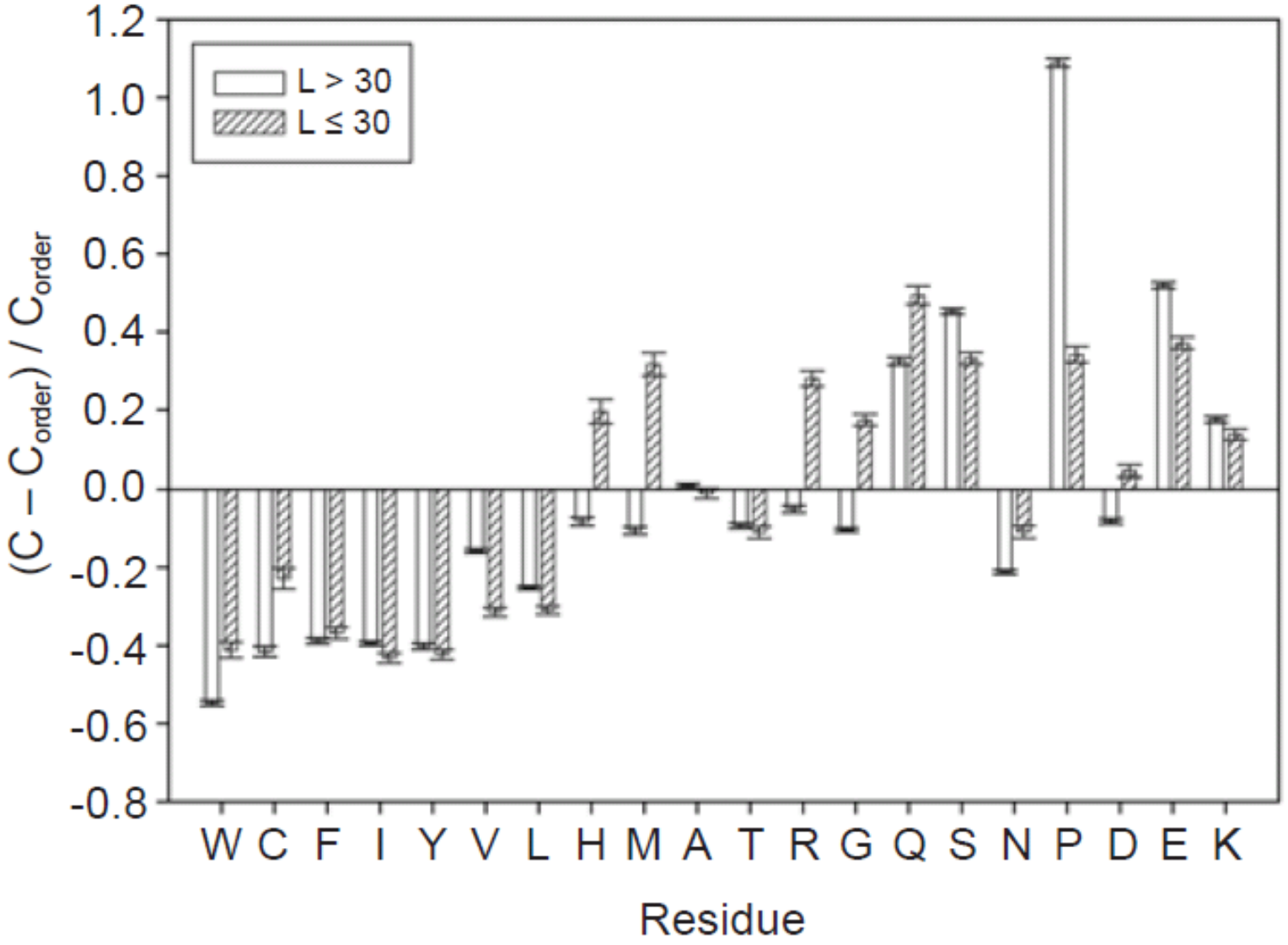



**Linding (2003) *NAR* 31, 3701**

# GlobPlot


RecName: Full=Cellular tumor antigen p53; AltName: Full=Antigen NY-CO-13; AltName: Full=Phosphoprotein p53; AltName: Full=Tumor suppressor p53;
Curves:
Russell
Linding
LOW-COMPLEXITY
DISORDER
GLOBDOM
downhill regions correspond to putative domains (GlobDom)
up-hill regions correspond to predicted protein disorder
Disorder propensity sum
4.8
2.8
0.7
-1.4
-3.5
-5.5
0
50
100
150
200
250
300
350
Residue

# Different flavors of disorder




Short and long disordered regions have different compositional biases

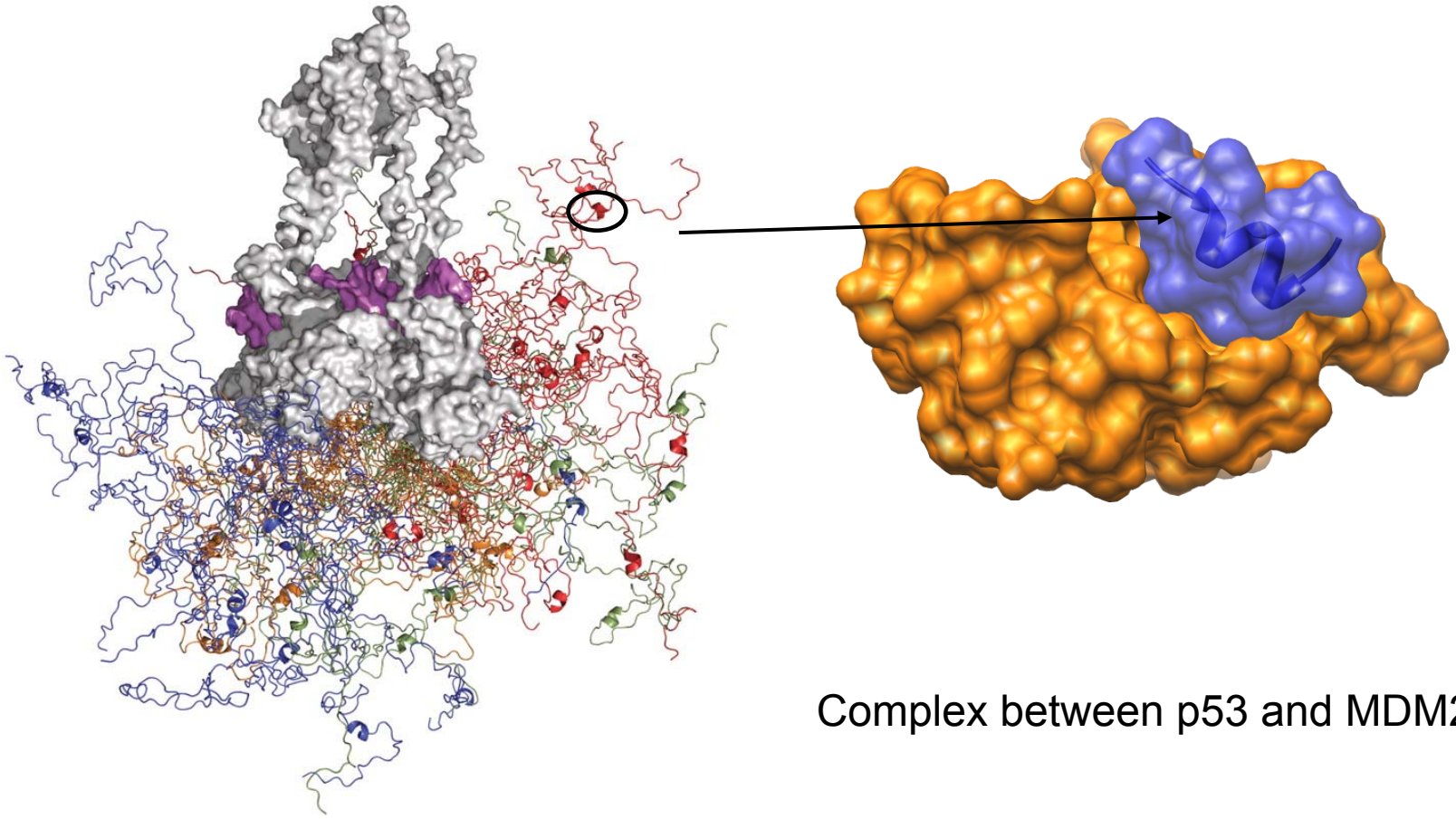
# Genome level annotations

- Combining experiments and predictions: **METASERVERS**
  - MobiDB: http://mobidb.bio.unipd.it
  - D2P2: http://d2p2.pro
  - IDEAL: http://www.ideal.force.cs.is.nagoya-u.ac.jp/IDEAL/
- Multiple predictors
- How to resolve contradicting experiments/ predictions?
  - Majority rules

# Functions of IDPs

I Entropic chains
II Linkers
III Molecular recognition
IV Protein modifications (e.g. phosphorylation)
V Assembly of large multiprotein complexes

# Protein interactions of IDPs



Complex between p53 and MDM2

# Binding regions within IDPs

- Complexes of IDPs in the PDB: ~ 200
- Known instances: ~ 2 000
- Estimated number of such interactions in the human proteome: ~ 100 000

- Experimental characterization is very difficult
- Computational methods

# Binding regions within IDPs

- **SLIMs: Short linear motifs**

  3-11 residues long, average size 6-7 residues

- **Disordered binding regions, Morfs**

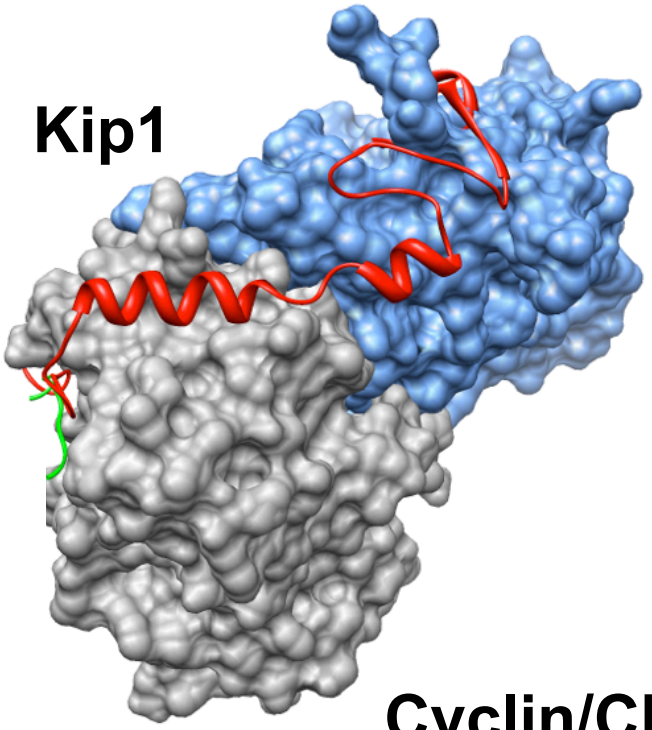
  undergo disorder to order transition upon binding

  usually less then 30 residues, can be up to 70

- **Intrinsically disordered domains**

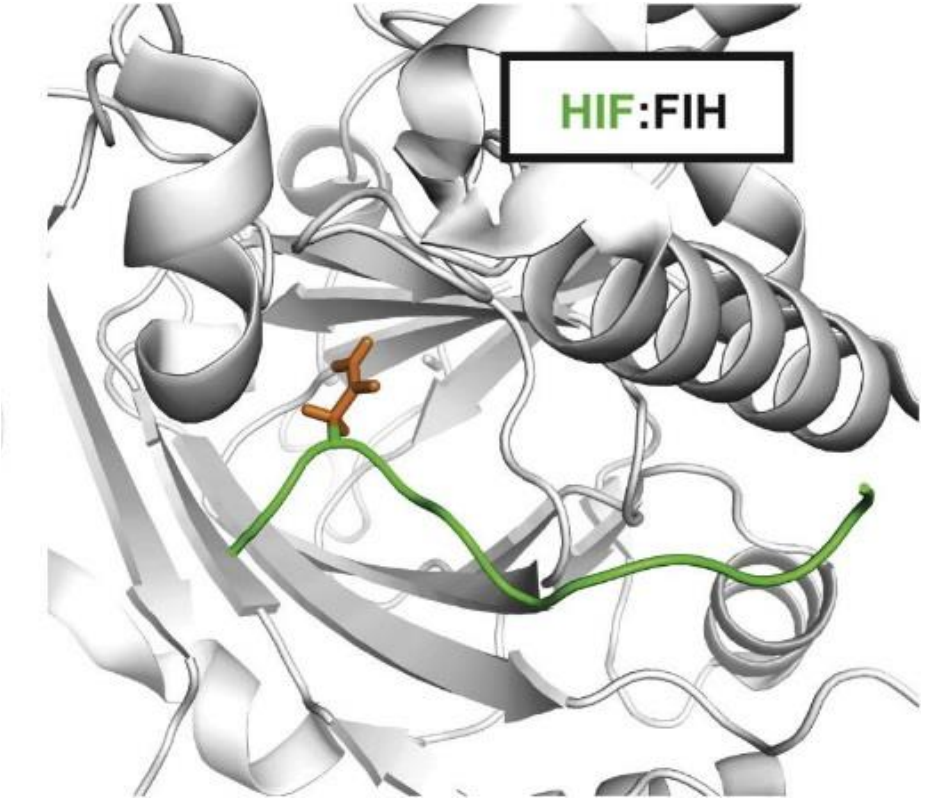
  evolutionary conserved disordered segments

## Intrinsically disordered regions (IDRs)




~ 20-50 residues

## Short Linear Motifs SLIMs




~ 3-10 residues

# Bioinformatical approaches

(~10, as opposed to the more than 50 disorder prediction methods)

- Biophysical properties (***ANCHOR***)
- Machine Learning methods

  ***(MorfPred, Morf$_{chibi}$, DISOPRED3)***
- Linear motifs

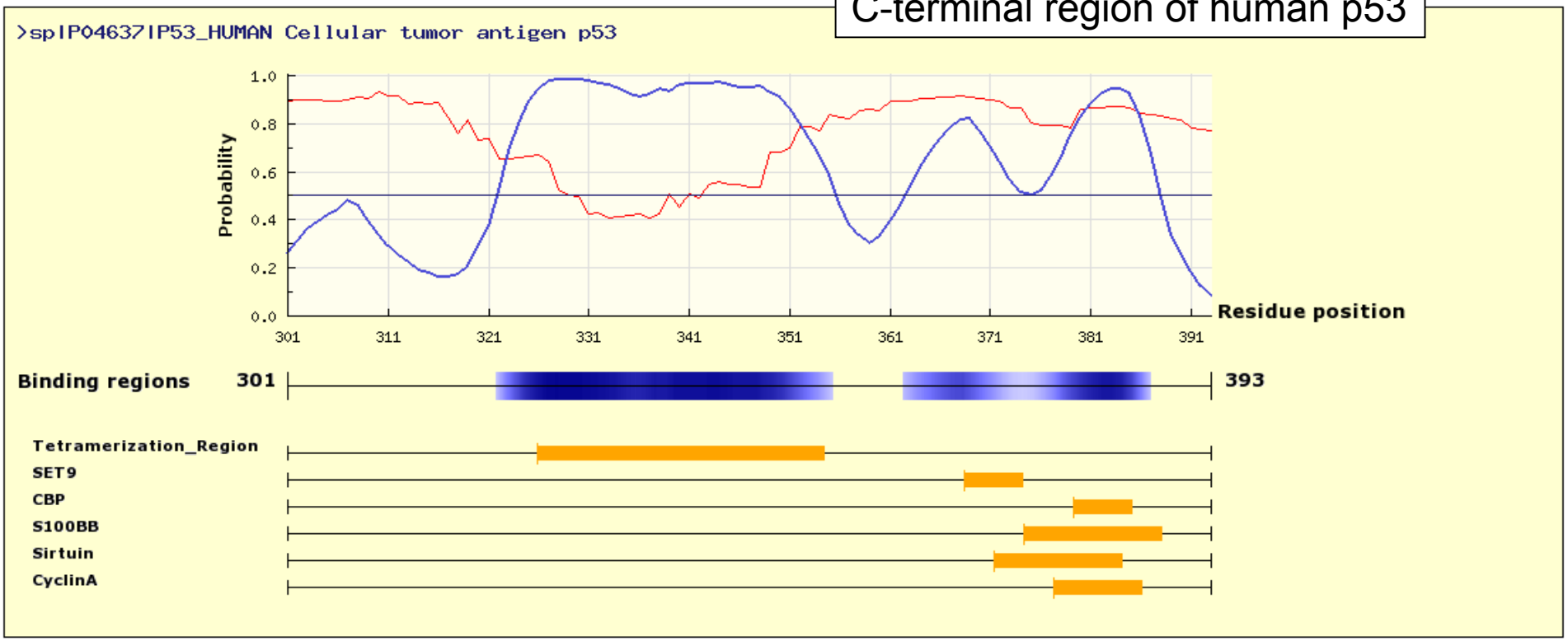
  ***(Regular Expression, PSSMs)***
- Conservations patterns (**SlimPrints, PhyloHMM** )

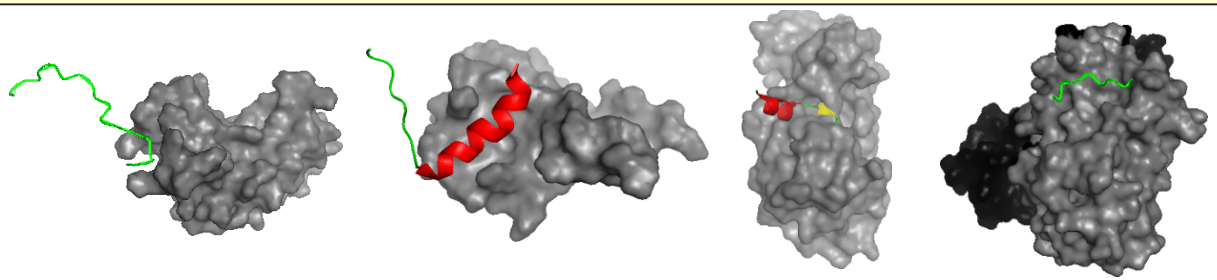
# Prediction of disordered binding regions – ANCHOR

- What discriminates disordered binding regions?
  - Cannot form enough favorable interactions with their sequential environment
  - It is favorable for them to interact with a globular protein
- Based on simplified physical model
  - Based on an energy estimation method using statistical potentials
  - Captures sequential context

# ANCHOR

C-terminal region of human p53



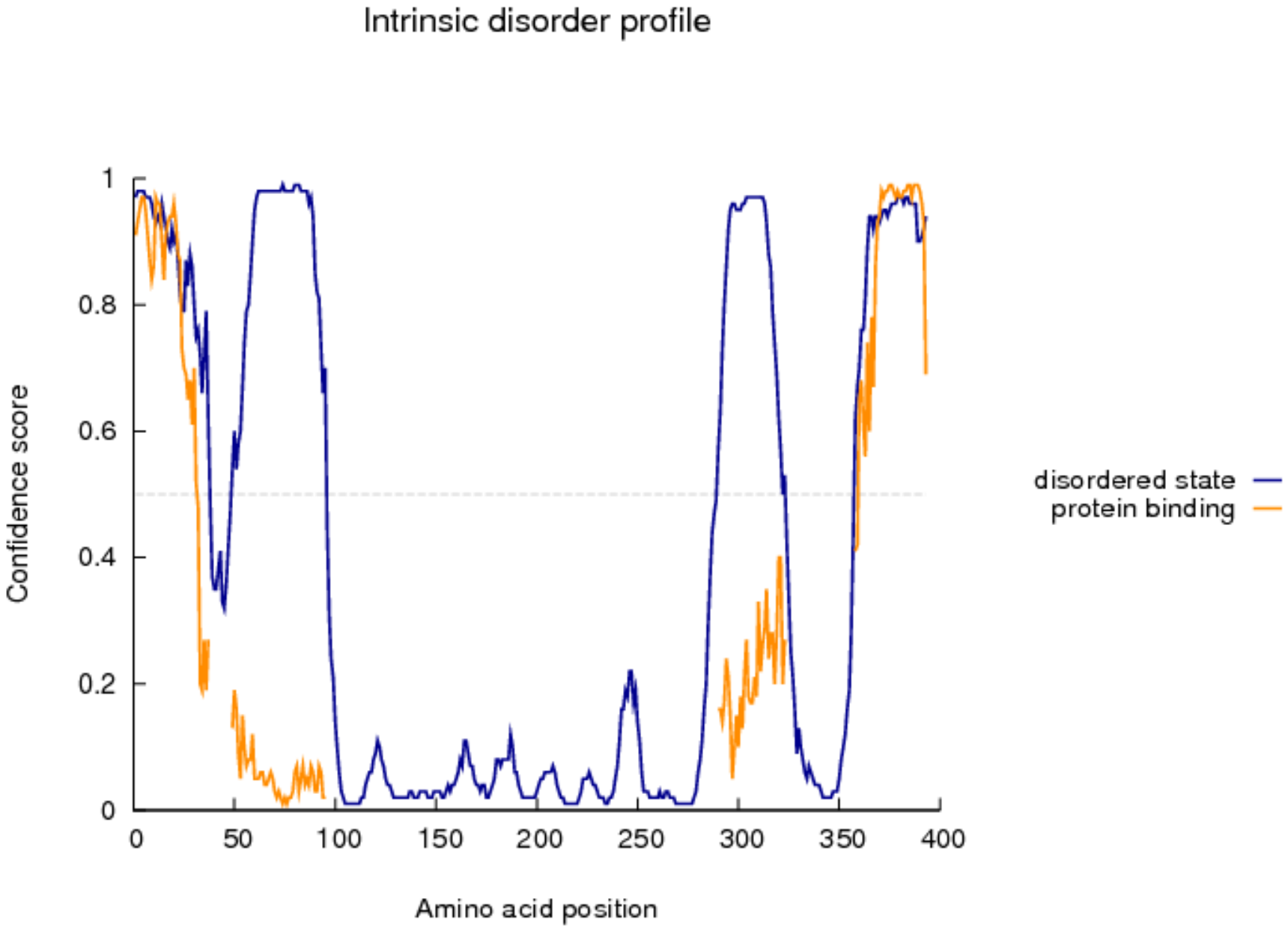
Dosztanyi et al. Bioinformatics. 2009;25:2745

# DISOPRED3

- Uses three SVMs
  - Simple sequence profile
  - PSI-Blast profiles (very slow)
  - PSI-Blast profiles with global features
- trained on short chains in complex

Jones DT, Cozzetto D. Bioinformatics. 2015; 31:857

# DISOPRED3


Intrinsic disorder profile
Confidence score
1
0.8
0.6
0.4
0.2
0
0
50
100
150
200
250
300
350
400
Amino acid position
disordered state
protein binding