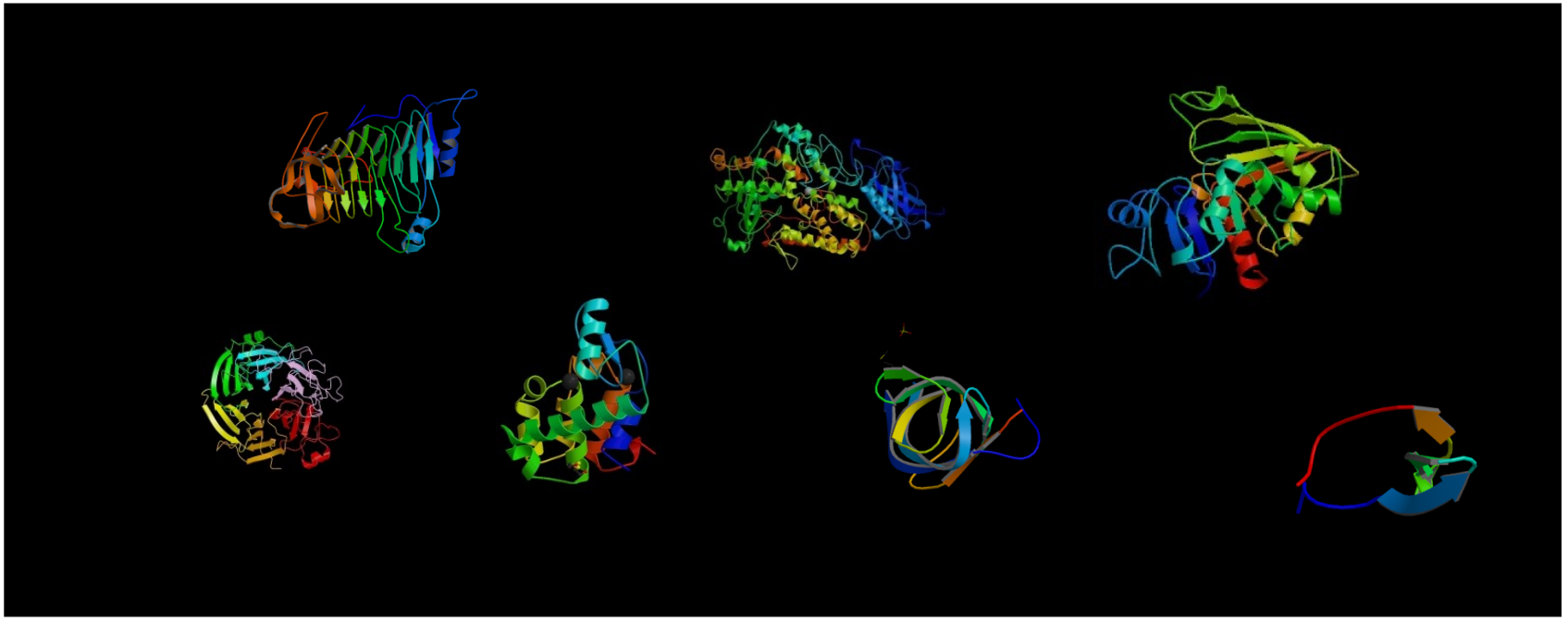
# 9.
# Intrinsically disordered proteins and protein binding regions

# IDPs

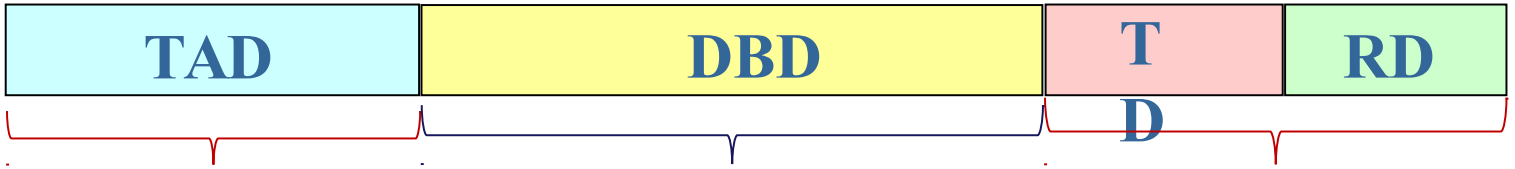
- Intrinsically disordered proteins/regions (IDPs/IDRs)
- Do not adopt a well-defined structure in isolation under native-like conditions
- Highly flexible ensembles
- Functional proteins
- Involved in various diseases

# Protein Structure/Function Paradigm

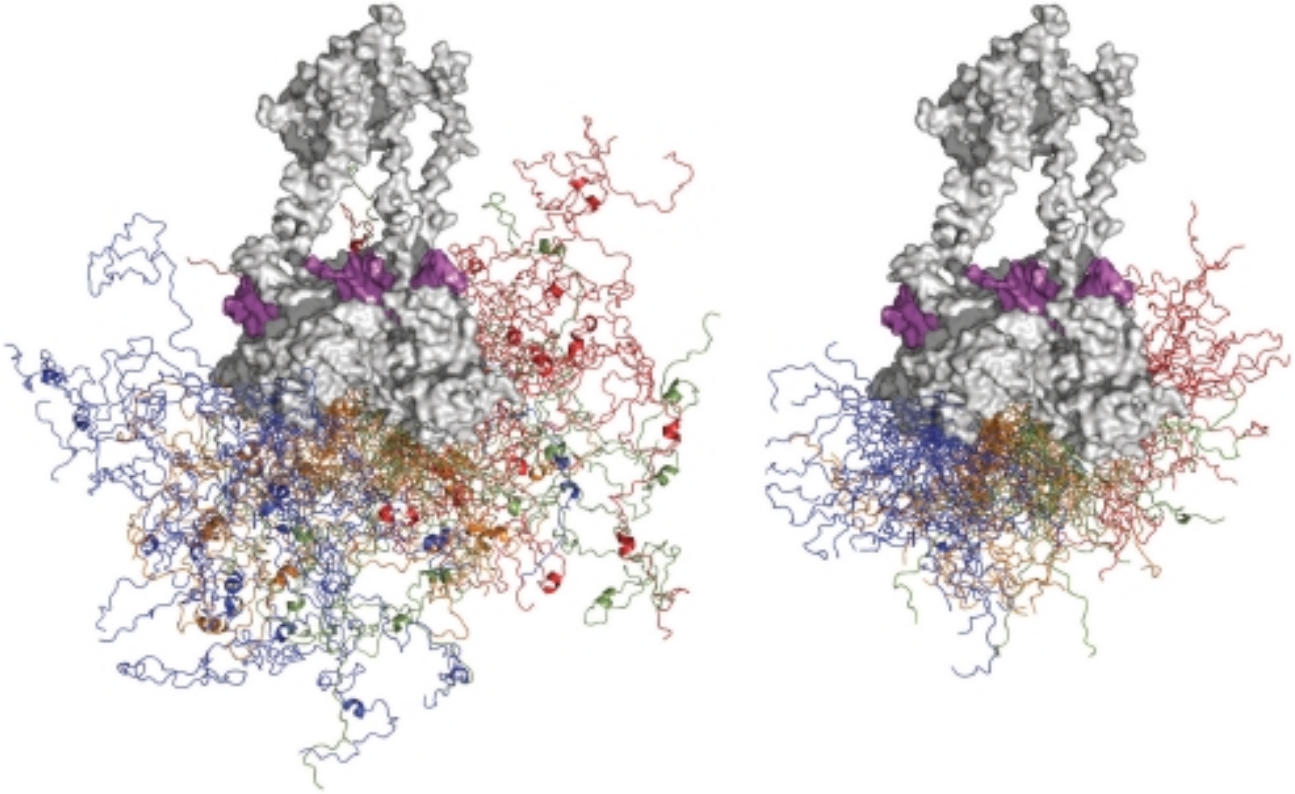


Dominant view: 3D structure is a prerequisite for protein function

# p53 tumor suppressor

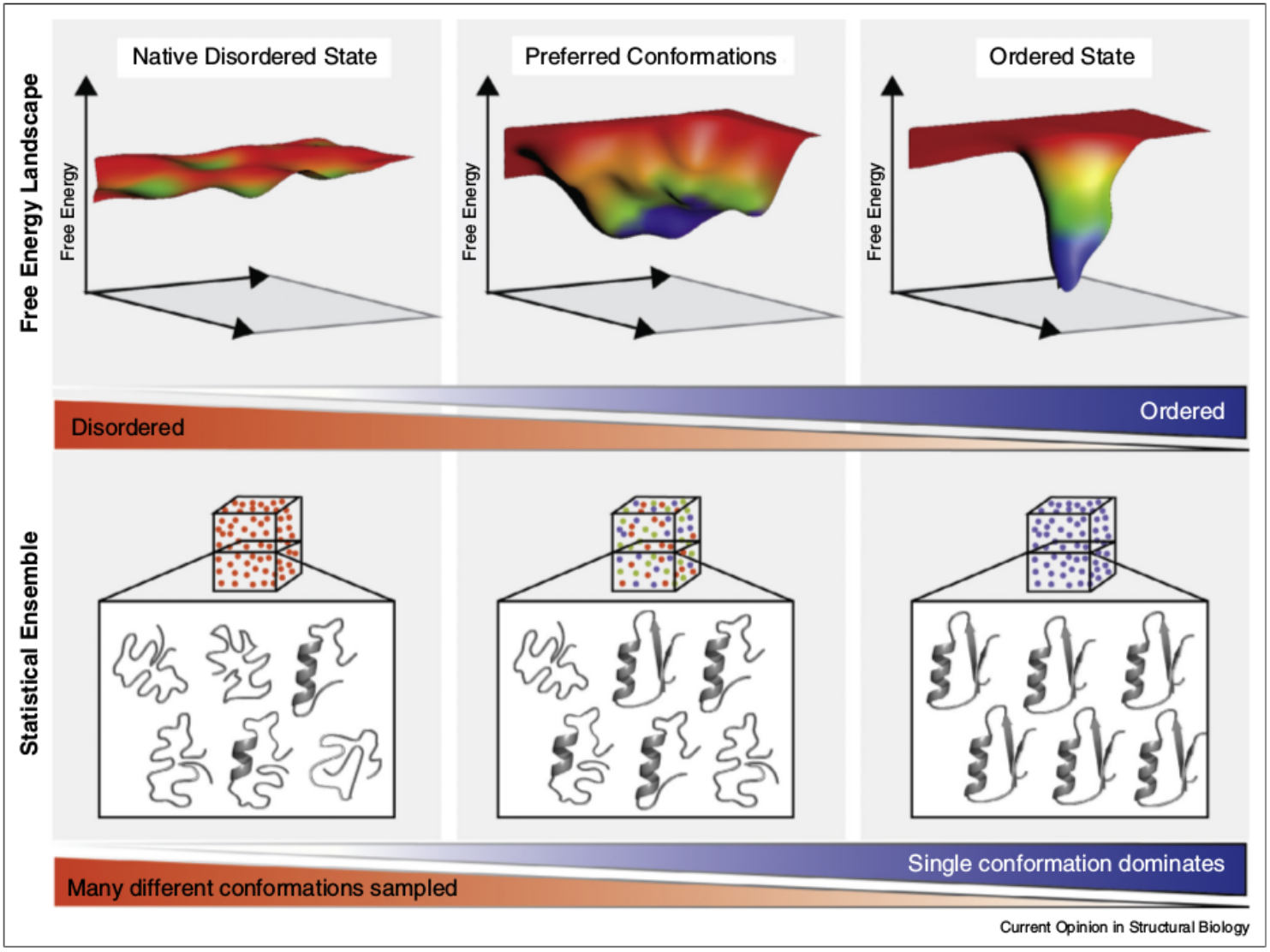



disordered ordered disordered



Wells et al. PNAS 2008; 105: 5762

# Funnels




Flock et al Curr Opin Struct Biol. 2014; 26:62

# Experimental detection of disorder

In the literature

Failed attempts to crystallize

Lack of NMR signals

Heat stability

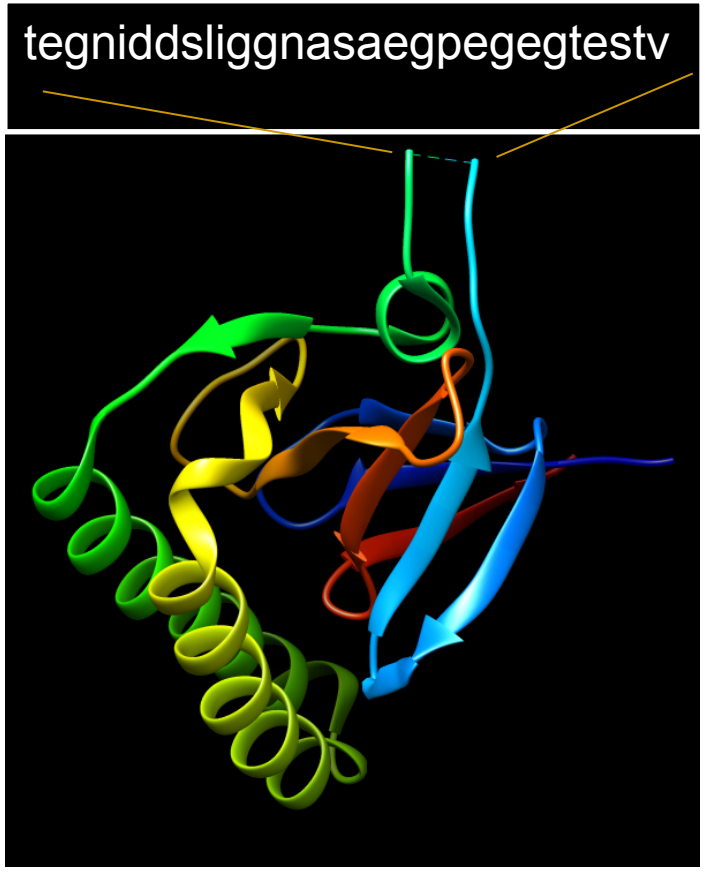
Protease sensitivity

Increased molecular volume

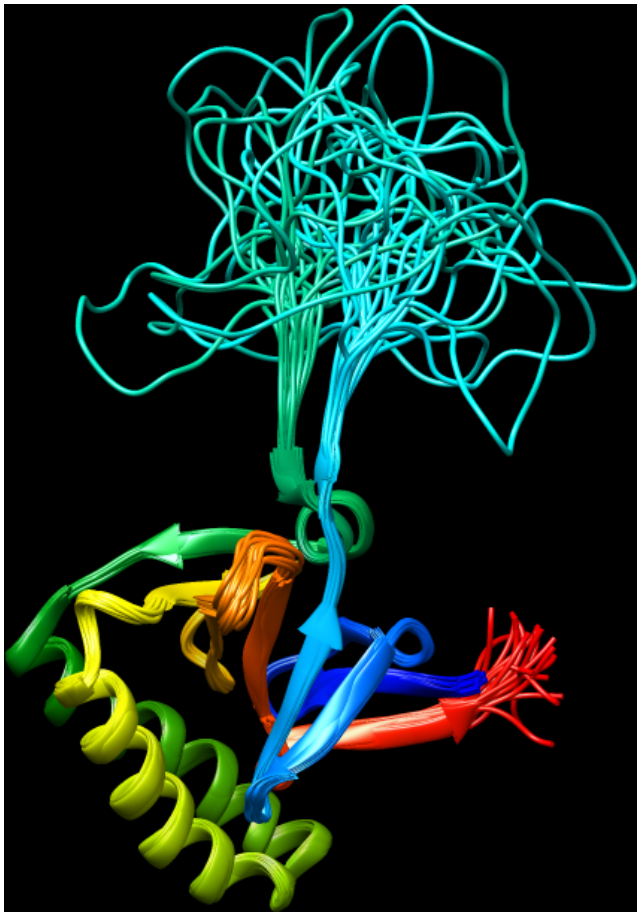
“Freaky” sequences …

# Where can we find disordered proteins?

In the PDB




Missing electron density regions from the PDB



NMR structures with large structural variations

DisProt – Database of protein disorder – Mozilla Firefox

DisProt - Database of...

disprot.bio.unipd.it

http://www.disprot.org/

PubMed | USDHUF=X Basic Char... | Budapest, Hungary For... | DISORDER | Tools | oz-masterclass - home | Computational analysis...

Browse | Search | About | Help | Statistics | Feedback

## New Version

DisProt 7 v0.3 26-09-2016

The **old DisProt** is still available!

## Statistics

**Proteins** 803

**Regions** 2167

## Start

You can do a **complex search** from the Browse page or use **Blast** from the Search page.

## Citing DisProt

Piovesan D et al. **DisProt 7.0: a major update of the database of disordered proteins**
Nucleic Acids Res., 2016.

Go to PubMed | Go to NAR

# Welcome to DisProt

DisPot is a **community resource** annotating protein sequences for intrinsically **disorder regions** from the literature.

It classifies intrinsic disorder based on **experimental methods** and three ontologies for **molecular function, transition and binding partner**.

# Sequence properties of disordered proteins

- 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***

# 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

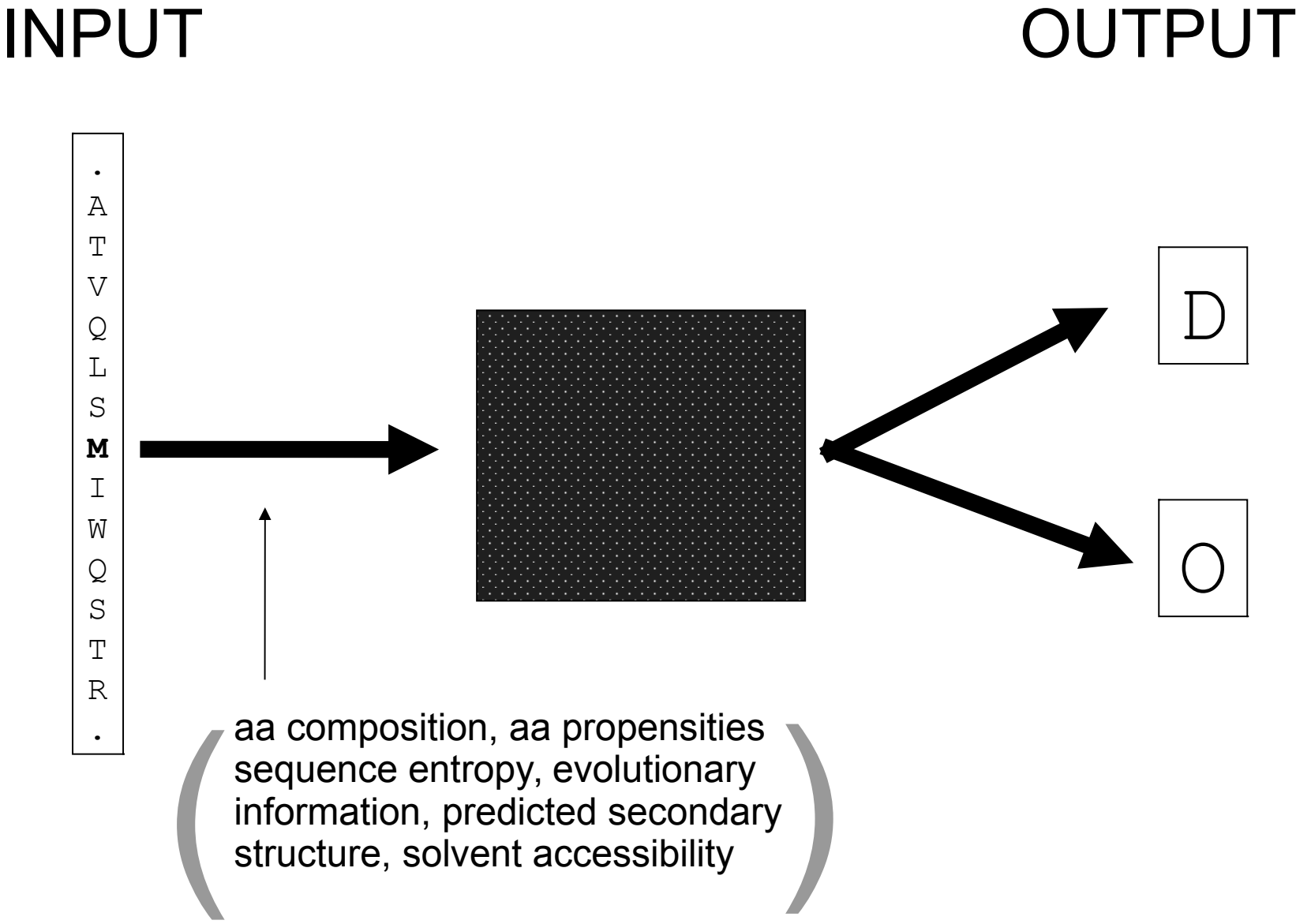
# Prediction methods for protein disorder

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, …

# Machine learning approaches



# IUPred

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

Energy estimation method

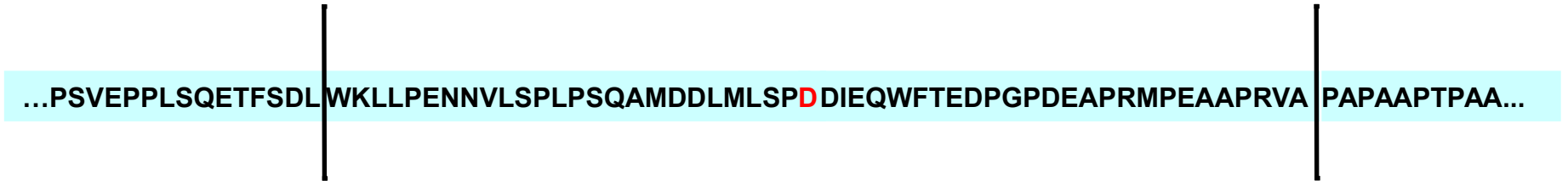
Based on globular proteins

No training on disordered proteins

Dosztanyi (2005) JMB 347, 827

# Predicting protein disorder - IUPred

- The algorithm:




*Based only on the composition of environment of D's we try to predict if it is in a disordered region or not:*

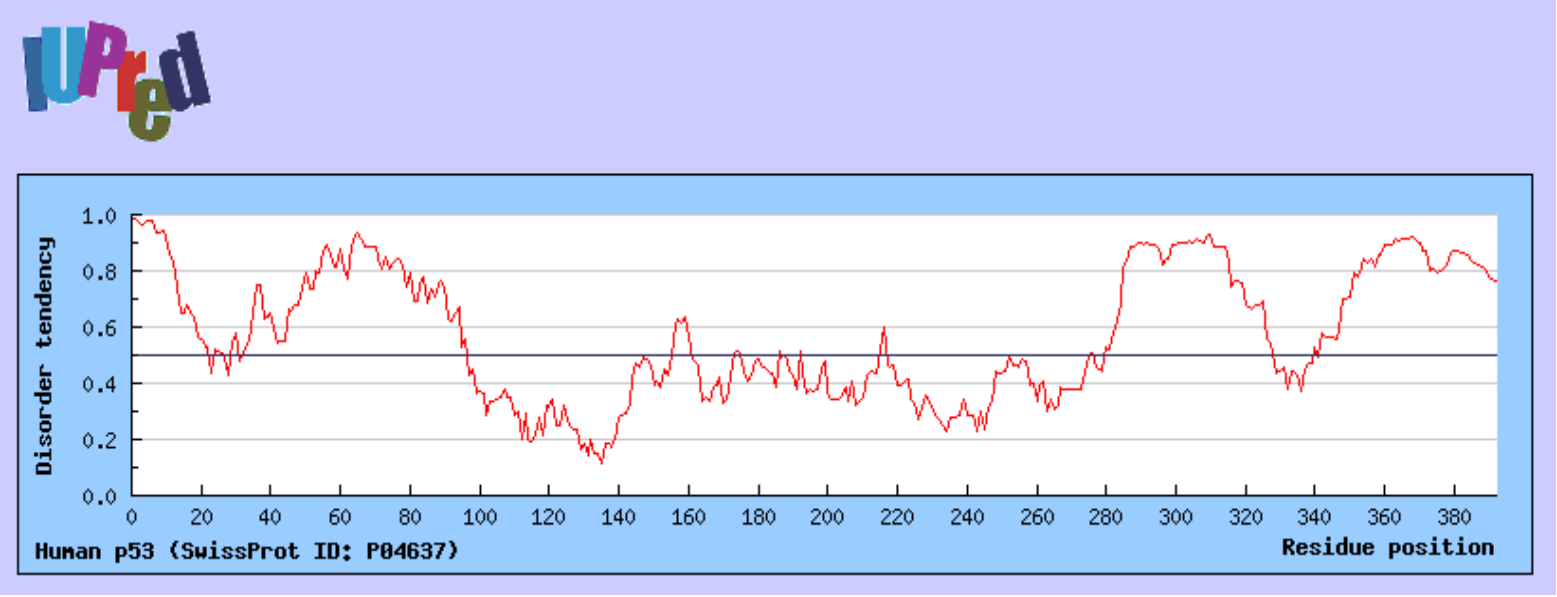
Amino acid composition of environ-ment:

A – 10%
C – 0%
D – 12 %
E – 10 %
F – 2 %
etc…

Estimate the interaction energy between the residue and its environment

Decide the probability of the residue being disordered based on this

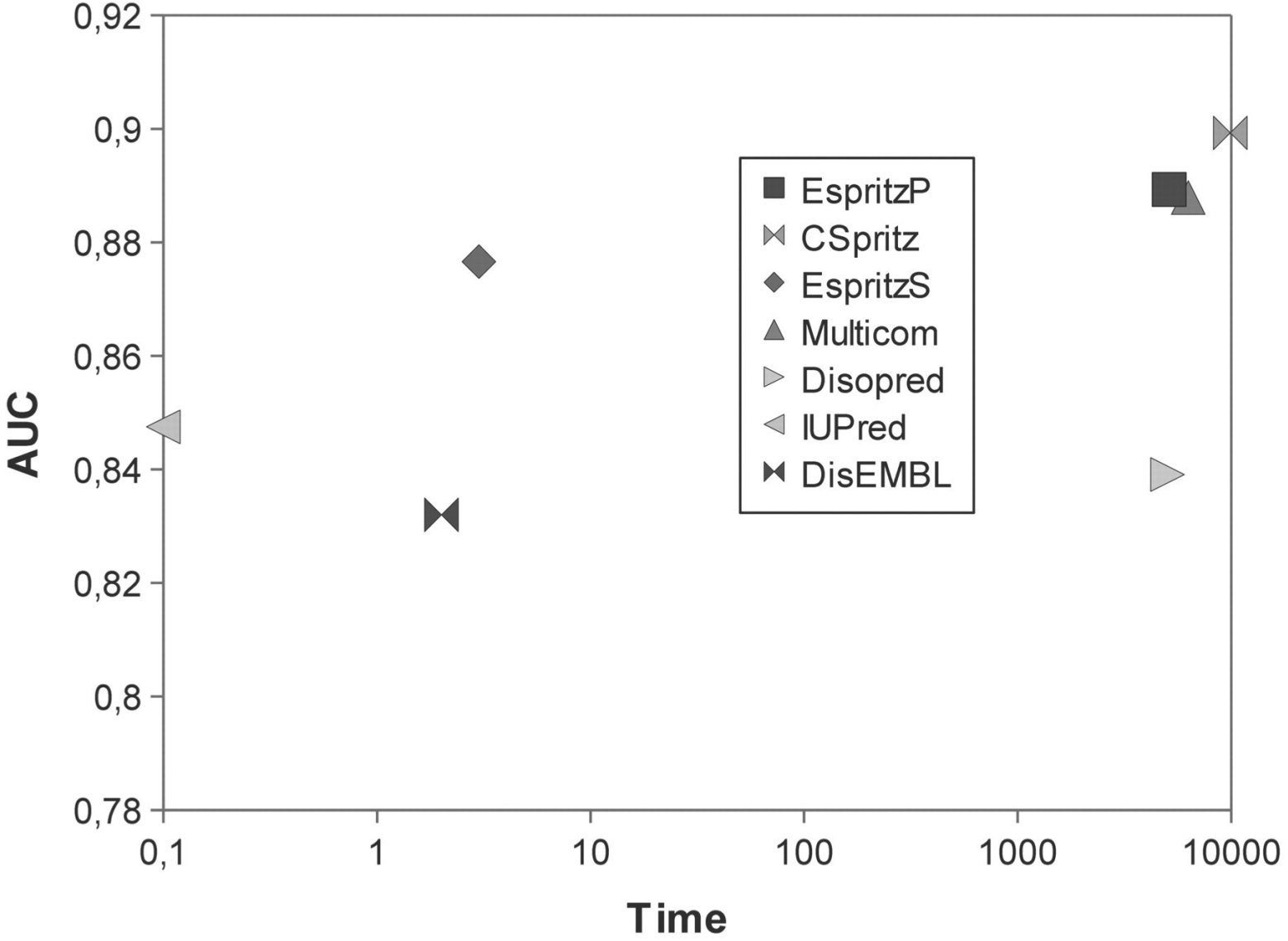
# IUPred: http://iupred.enzim.hu/



# Prediction of protein disorder

- Disordered is encoded in the amino acid sequence
- Can be predicted from the sequence
- ~80% accuracy
- Large-scale studies
  - Evolution
  - Function
- Binary classification

**Time versus performance plot for different predictors.**




**Ian Walsh et al. Bioinformatics 2012;28:503-509**



Bioinformatics

# Genome level annotations

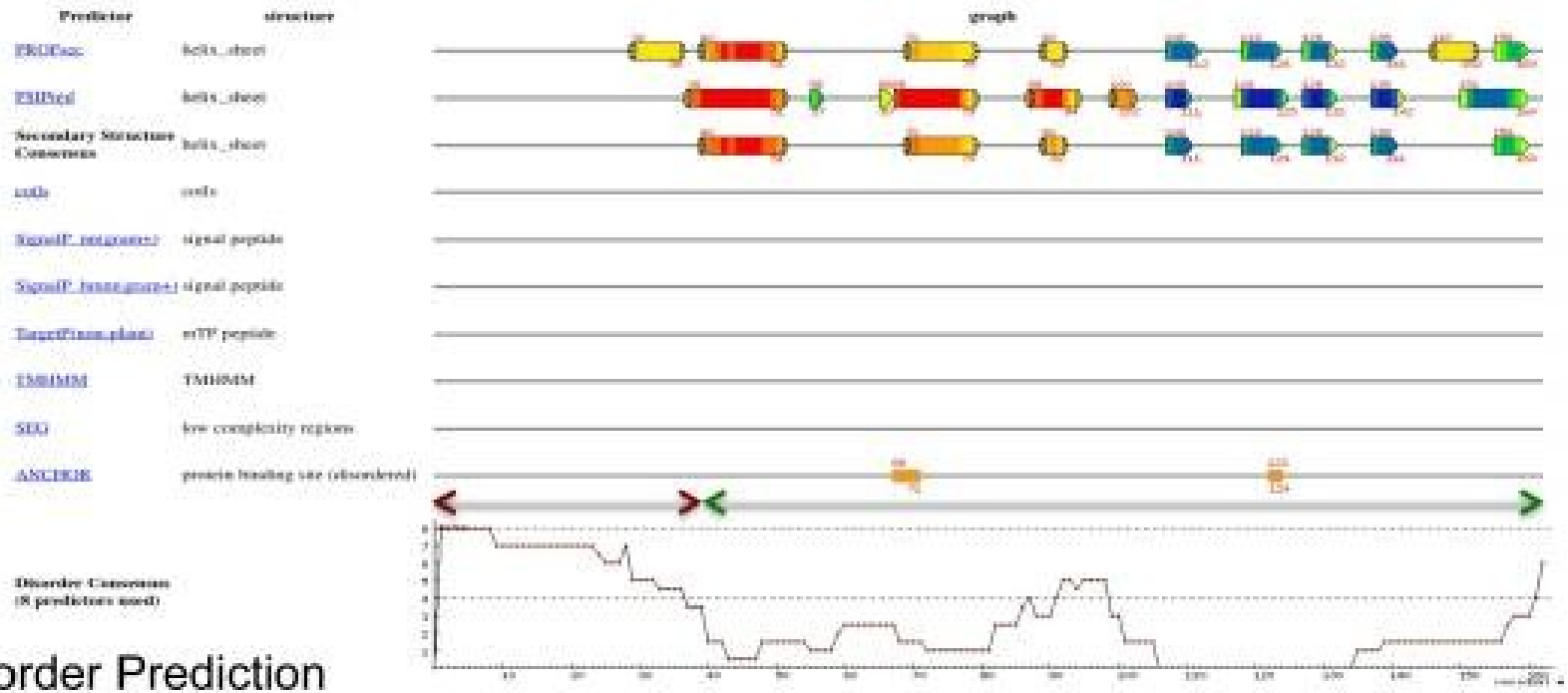
- Bridging over the large number of sequences and the small number of experimentally verified cases
- Combining experiments and predictions
  - MobiDB: http://mobidb.bio.unipd.it
  - D2P2: http://d2p2.pro
  - IDEAL: http://www.ideal.force.cs.is.nagoya-u.ac.jp/IDEAL/
- Multiple predictors
- How to resolve contradicting experiments/ predictions?
  - Majority rules

# MobiDB

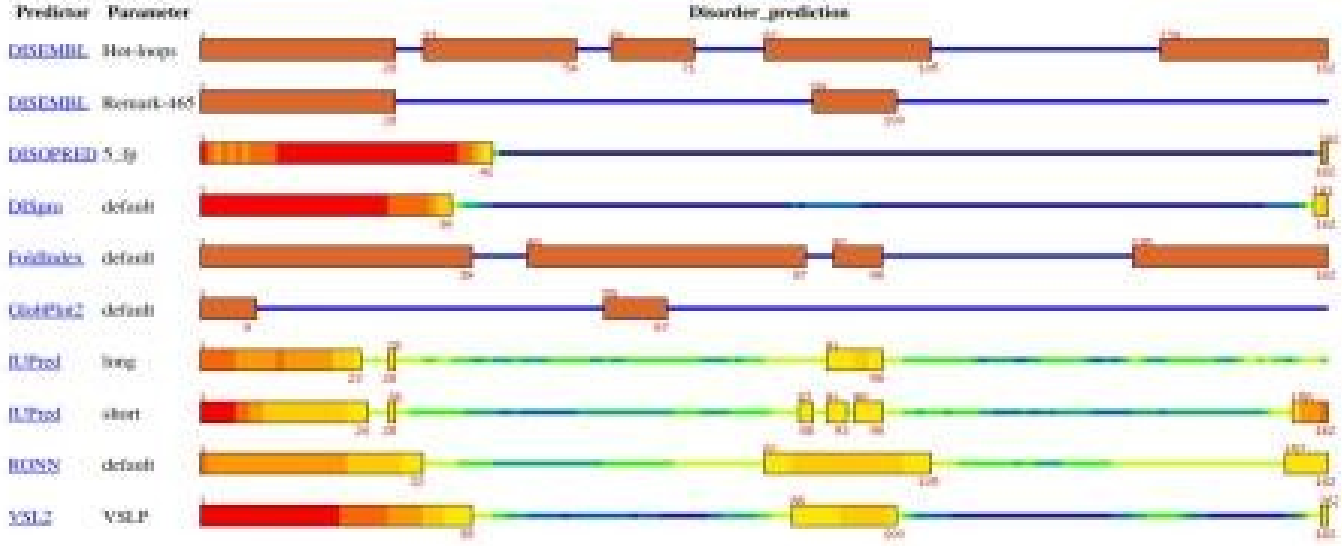

DP00070
100.0 %
PDB-NMR [-]
Consensus
PDB-NMR
35.71 %
2jn5
consensus [+]
0.000 %
3D view
2kkw
consensus [+]
42.14 %
3D view
2m55
consensus [+]
10.00 %
3D view
4bxl
consensus [+]
0.000 %
3D view
PDB-XRay [-]
Consensus
PDB-XRay
07.86 %
2x6m 1.62Å
consensus [+]
02.14 %
3D view
3q25 1.9Å
consensus [+]
0.000 %
3D view
3q26 1.54Å
consensus [+]
05.71 %
3D view
3q27 1.302Å
consensus [+]
04.29 %
3D view
3q28 1.6Å
consensus [+]
05.00 %
3D view
3q29 2.3Å
consensus [+]
03.57 %
3D view
Predictors [-]
Consensus
Predictors
41.43 %
DisEMBL-465
30.71 %
DisEMBL-HL
66.43 %
ESpritz-DisProt
100.0 %
ESpritz-NMR
30.00 %
ESpritz-XRay
34.29 %
GlobPlot
16.43 %
IUPred-long
40.00 %
IUPred-short
33.57 %
JRONN
40.00 %
VSL2b
90.71 %

# Construct optimization

A. Disorder Prediction

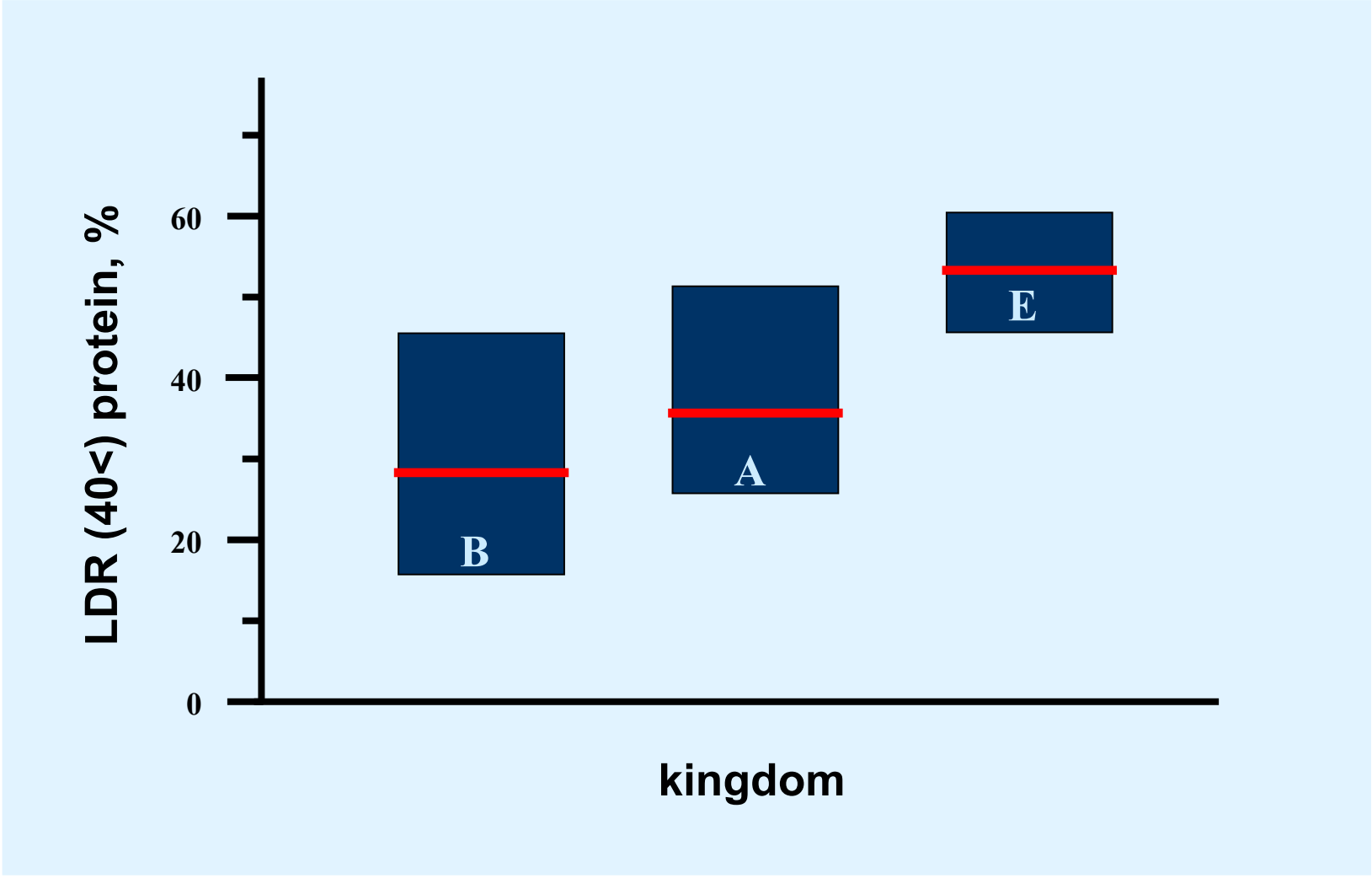



B. Disorder Prediction



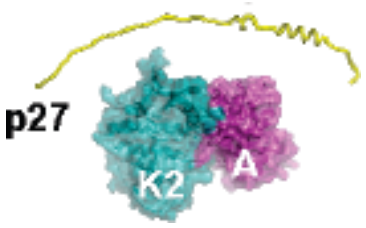
# How common is protein disorder?

Disorder content increases with evolutionary complexity

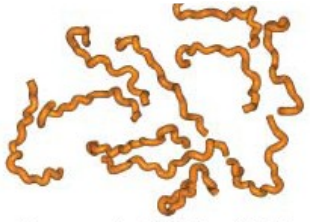


# Disorder is heterogeneous

Transient structures

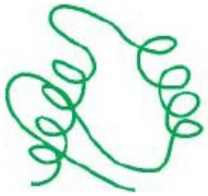



RC-like



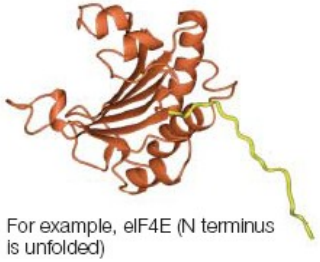
For example, ACTR (no NCBD)

Compact



For example, NCBD (no ACTR)

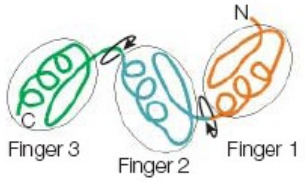
**Mostly folded, local disorder**



For example, eIF4E (N terminus is unfolded)


Flexible loop

**Linked folded domains**
(beads on a string)




For example, zinc fingers (no DNA)

# Structural ensemble PEDB database

**The following 2 entries have been returned for your query:**

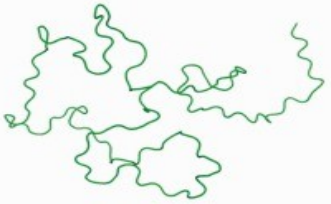
Select all ☐

Download selected

## ■ Ensemble description of K18 domain of Tau protein using NMR techniques

| Accesion ID | Correspondent | Release date | SAXS data | NMR data |
|---|---|---|---|---|
| 6AAC | Martin Blackledge | 2013-06-10 | No | Yes |

- Download complete entry (compressed)
- Download structure archive (.pdb)
- Download sequences (.fasta)
- Download experimental data



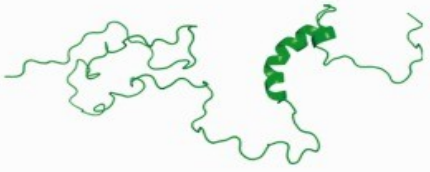
**Authors:** Markus Zweckstetter; Martin Blackledge; Valery Ozenne; Robert Schneider; Mingxi Yao; Jie-rong Huang; Loic Salmon; Malene Ringkjobing Jensen;

**Keywords:** asteroids; flexible-meccano; intrinsically disordered; NMR; single residue resolution;

## ■ Ensemble of the free form N-TAIL Measles nucleoprotein

| Accesion ID | Correspondent | Release date | SAXS data | NMR data |
|---|---|---|---|---|
| 7AAC | Martin Blackledge | 2013-06-13 | No | Yes |

- Download complete entry (compressed)
- Download structure archive (.pdb)
- Download sequences (.fasta)
- Download experimental data



**Authors:** Markus Zweckstetter; Martin Blackledge; Valery Ozenne; Robert Schneider; Mingxi Yao; Jie-rong Huang; Loic Salmon; Malene Ringkjobing Jensen;

**Keywords:** intrinsically disordered; NMR;

# How IDPs carry out their functions?

- Entropic chains

  Function directly results from disordered state

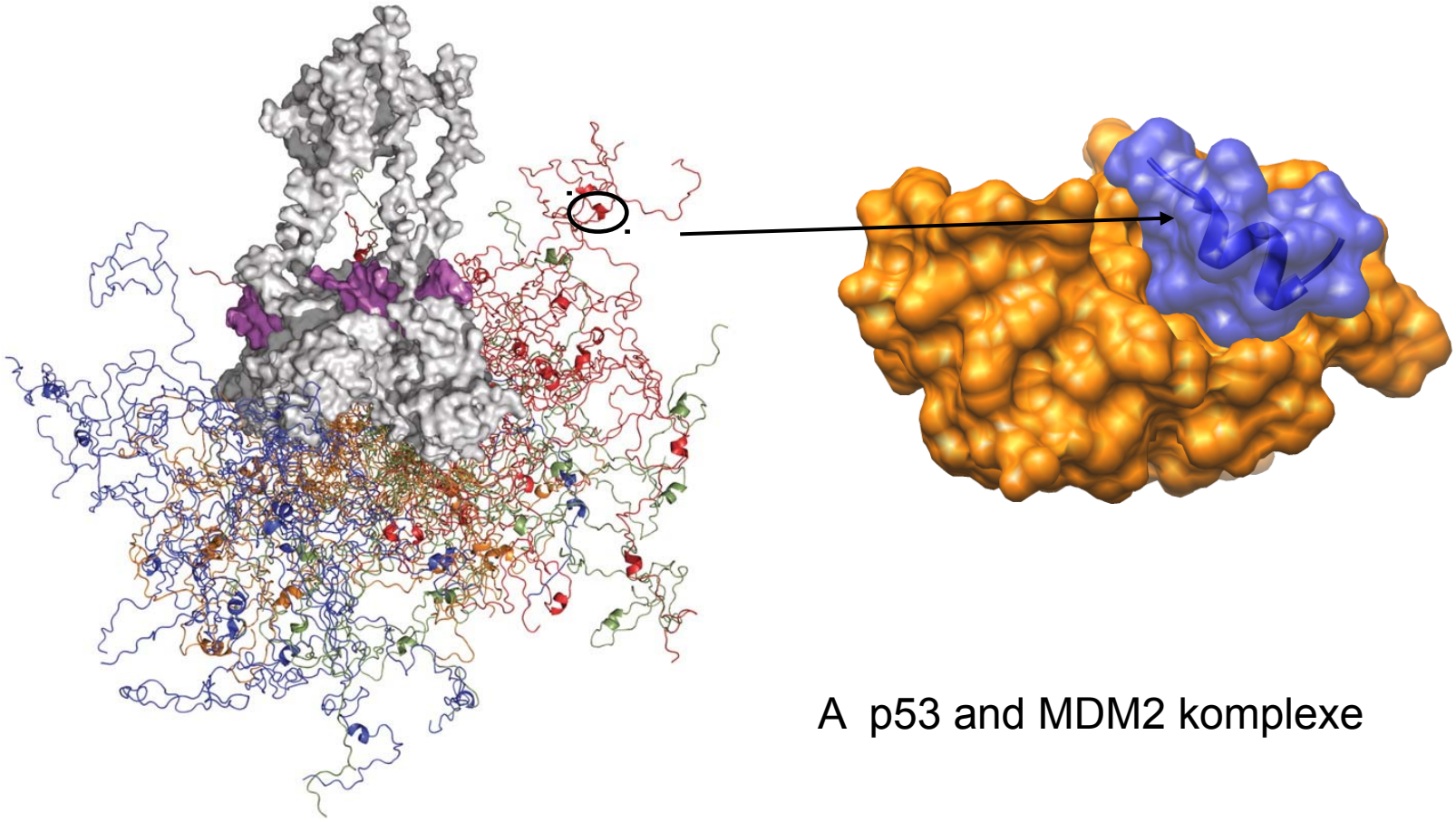
- Molecular recognition

  Coupled folding and binding

- “Assemblages”

  Functional sites formed by phase separation

# Protein interactions of IDPs



A  p53 and MDM2 komplexe

# Coupled folding and binding

- Entropic penalty
- Functional advantages
  - Weak transient, yet specific interactions
  - Post-translational modifications
  - Flexible binding regions that can overlap
  - Evolutionary plasticity

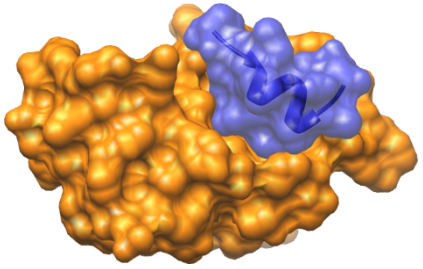
# Interactions of IDPs

- Complexes of IDPs in the PDB: ~ 200
- Known instances: ~ 2 000
- Estimated number of such interactions in the human proteome: ~ 1 000 000

- Experimental characterization is very difficult
  - Low expression level
  - Sensitive to proteolysis
  - Experimental methods are tailored for globular proteins

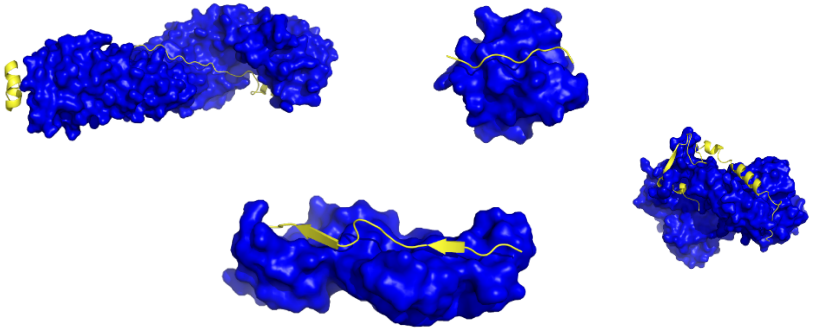
- Computational methods

# Prediction of binding sites located within IDPs

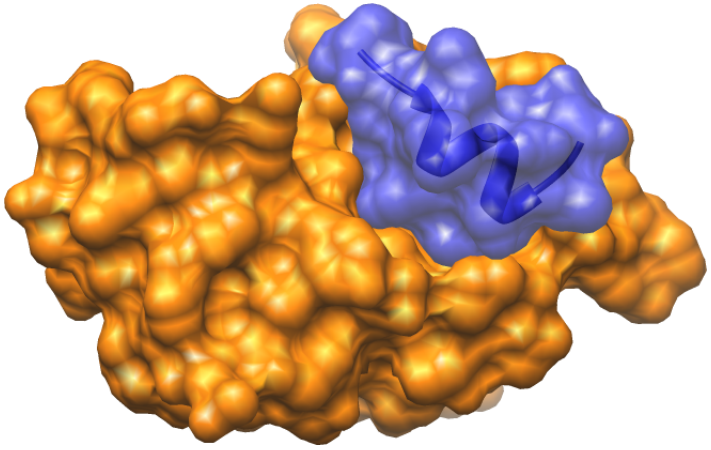


- Interaction sites are usually linear (consist of only 1 part)
- enrichment of interaction prone amino acids
- can be predicted from sequence without predicting the structure

- **Heterogeneity**
  - adopted secondary structure elements
  - size of the binding regions
  - flexibility in the bound form



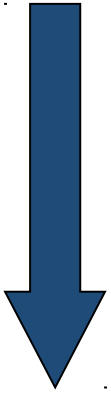
# Disordered protein complexes



Complex between p53 and MDM2

- Interaction sites are usually ***linear*** (consist of only 1 part)
- enrichment of interaction prone amino acids

**Sequence**



**Binding sites**

No need for structure, binding sites can be predicted from sequence alone

# Prediction of disordered binding regions – ANCHOR

What discriminates disordered binding regions?

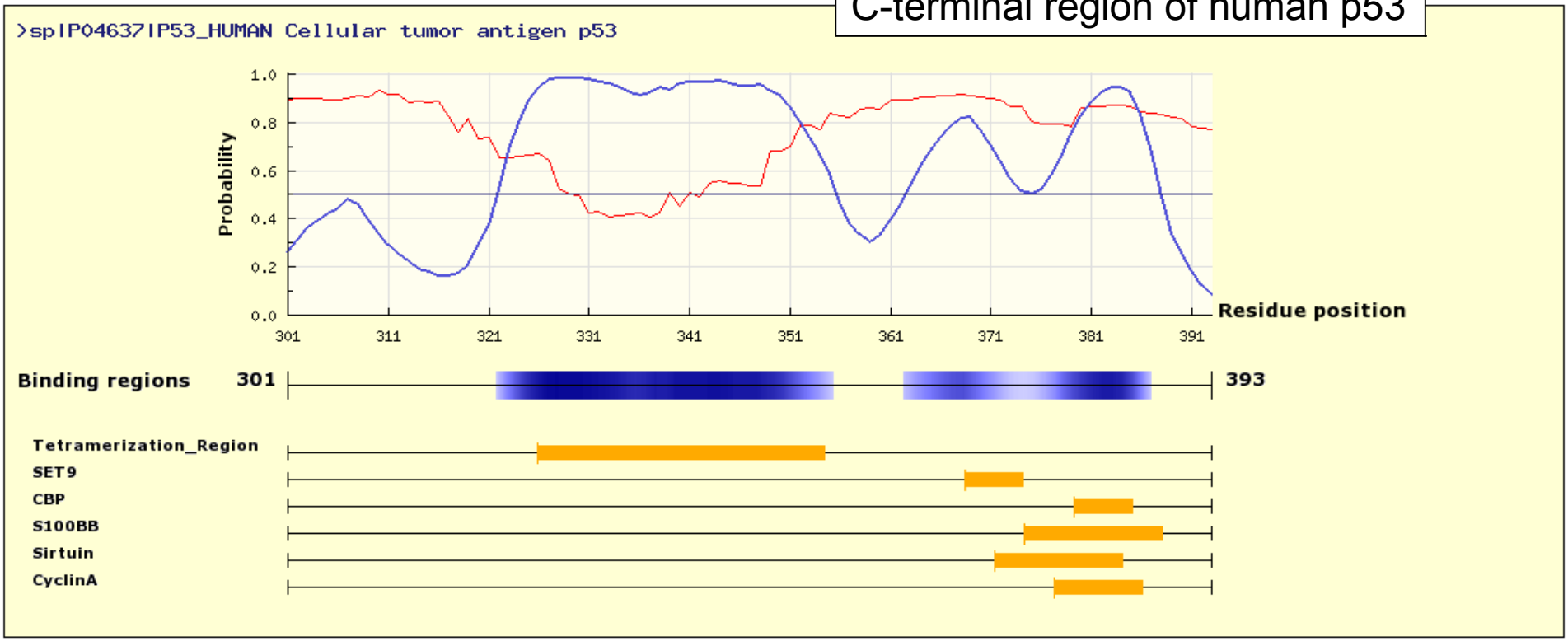
- A cannot form enough favorable interactions with their sequential environment
- It is favorable for them to interact with a globular protein

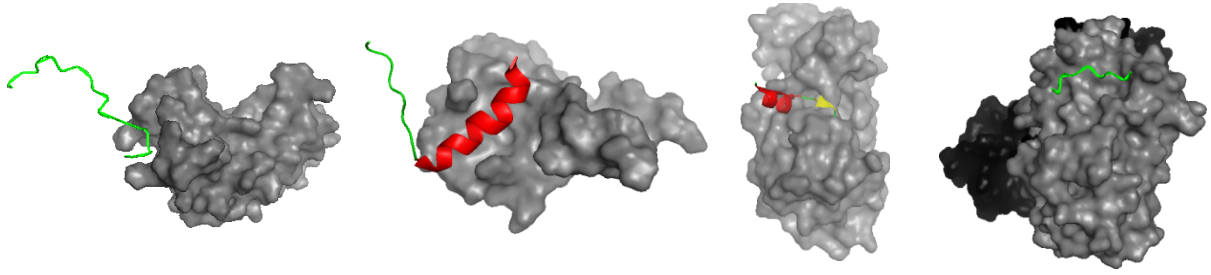
Based on simplified physical model

- Based on an energy estimation method using statistical potentials
- Captures sequential context

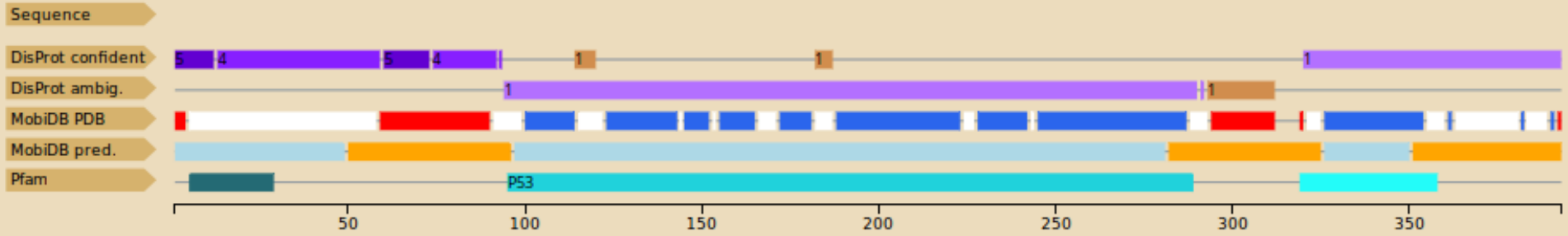
# Prediction of disordered binding regions - ANCHOR

C-terminal region of human p53



# Disorder Overview




DisProt Disorder
PDB Disorder
Predicted Disorder
DisProt Context-dependent
PDB Structure
Predicted Structure
PDB Ambiguous

# Disorder Region Details

Color by Evidences | Molecular function | Type of molecular transitions | Molecular partner

Hide ambiguous evidences