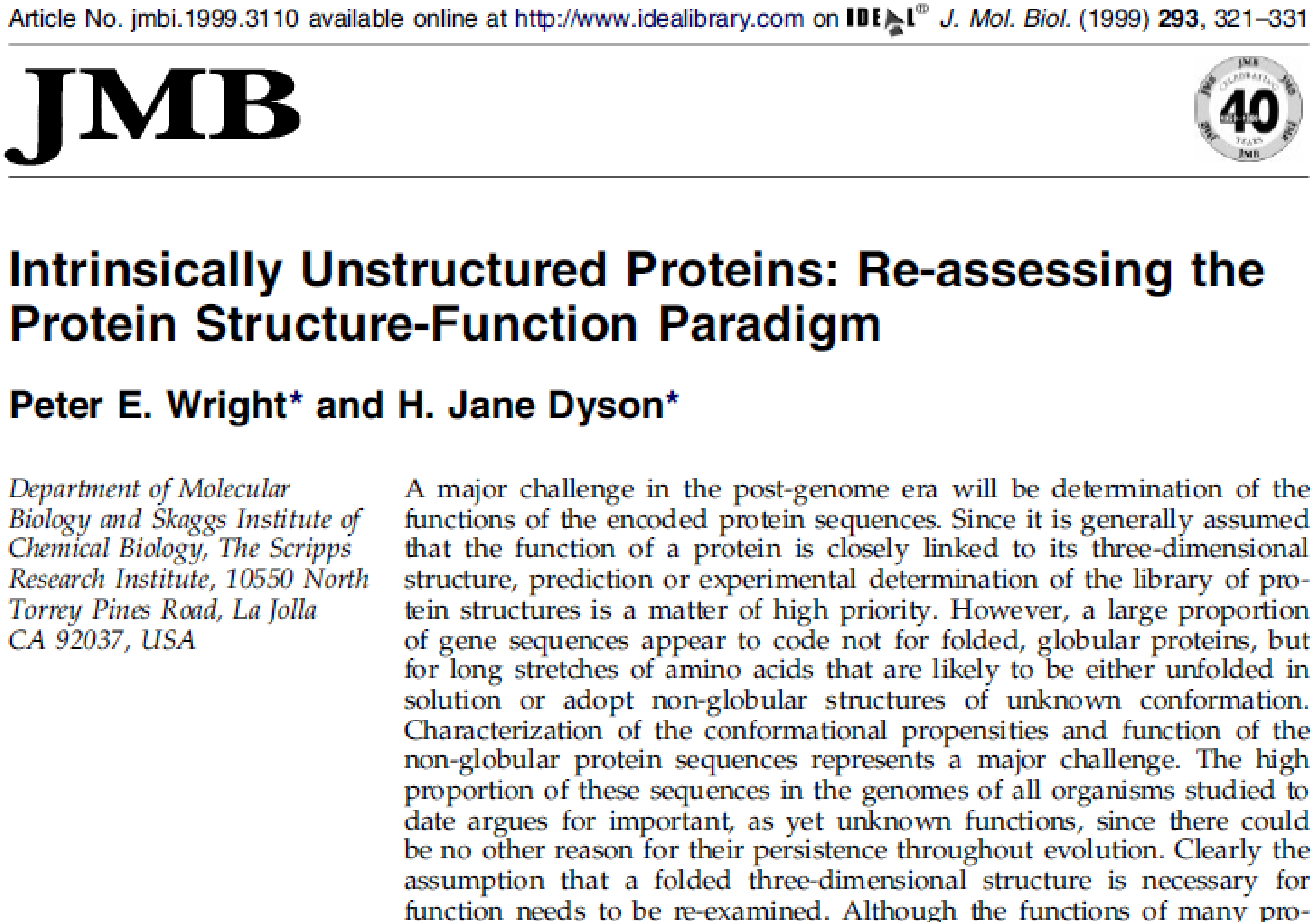
Intrinsically disordered proteins

Zsuzsanna Dosztányi

EMBO course
Budapest, 3 June 2016

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



Article No. jmbi.1999.3110 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (1999) **293**, 321–331

JMB

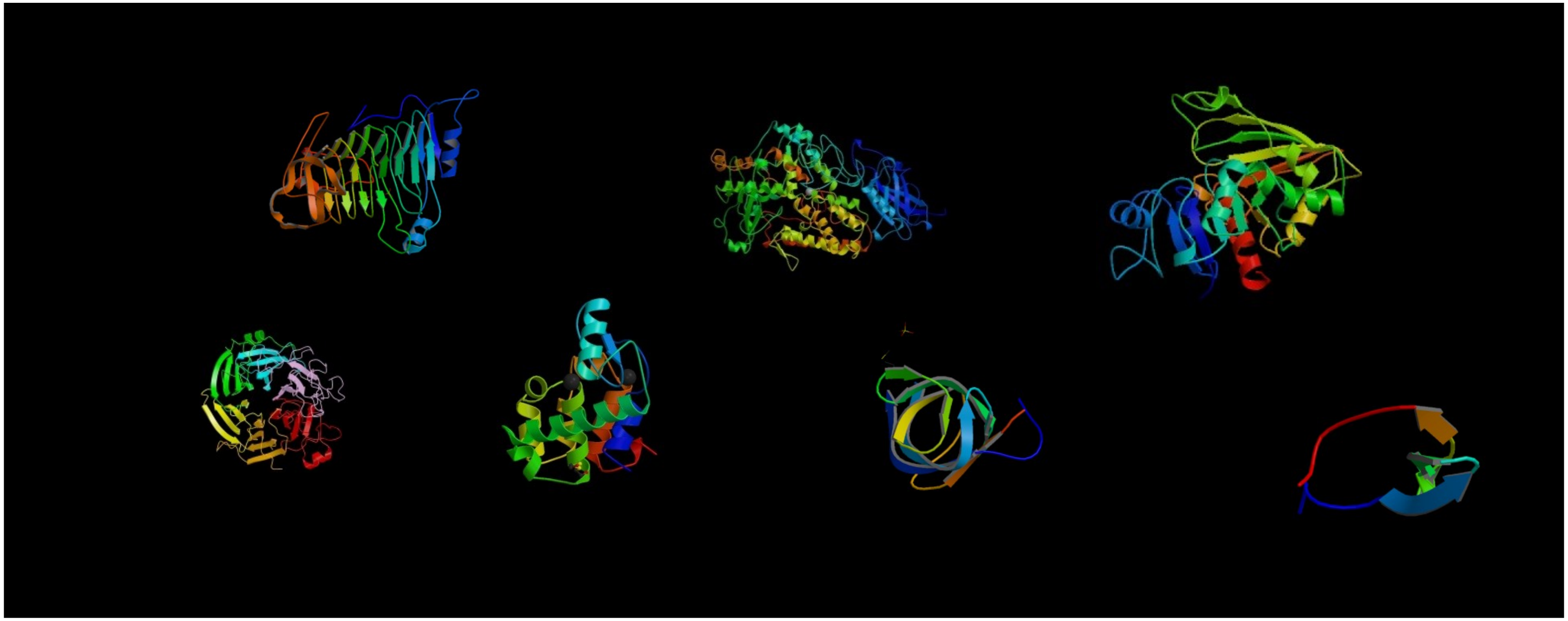
Intrinsically Unstructured Proteins: Re-assessing the Protein Structure-Function Paradigm

Peter E. Wright* and H. Jane Dyson*

Department of Molecular Biology and Skaggs Institute of Chemical Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla CA 92037, USA

A major challenge in the post-genome era will be determination of the functions of the encoded protein sequences. Since it is generally assumed that the function of a protein is closely linked to its three-dimensional structure, prediction or experimental determination of the library of protein structures is a matter of high priority. However, a large proportion of gene sequences appear to code not for folded, globular proteins, but for long stretches of amino acids that are likely to be either unfolded in solution or adopt non-globular structures of unknown conformation. Characterization of the conformational propensities and function of the non-globular protein sequences represents a major challenge. The high proportion of these sequences in the genomes of all organisms studied to date argues for important, as yet unknown functions, since there could be no other reason for their persistence throughout evolution. Clearly the assumption that a folded three-dimensional structure is necessary for function needs to be re-examined. Although the functions of many pro-

Ordered structures from the PDB



Over 100000 PDB structures

Not everything in the PDB is ordered

Cofactors, complex, DNA-RNA, crystal contacts

Where can we find disordered proteins?

In the literature

Failed attempts to crystallize

Lack of NMR signals

Heat stability

Protease sensitivity

Increased molecular volume

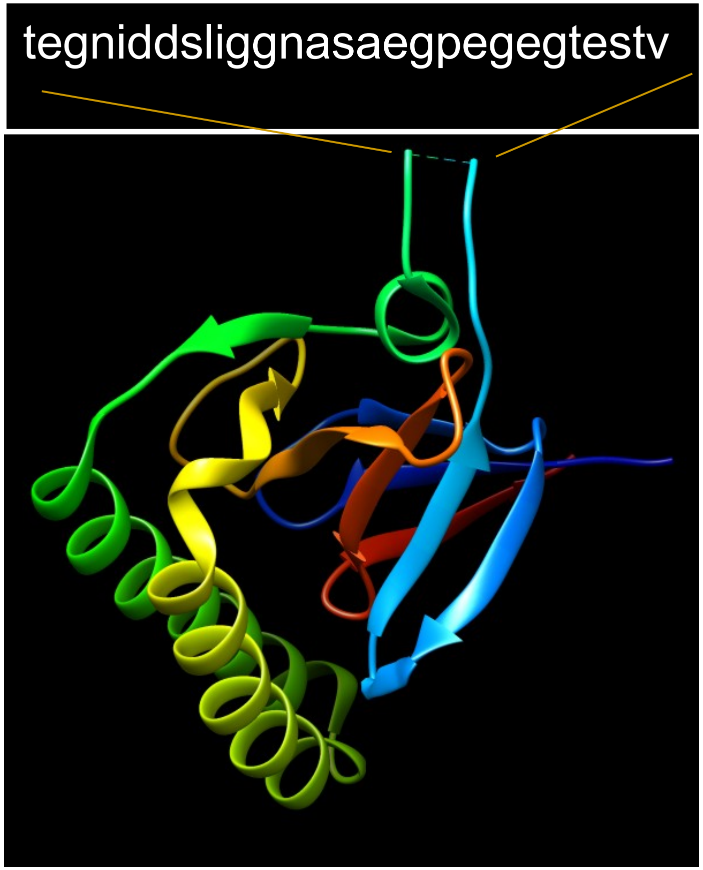
“Freaky” sequences …

Disprot database:

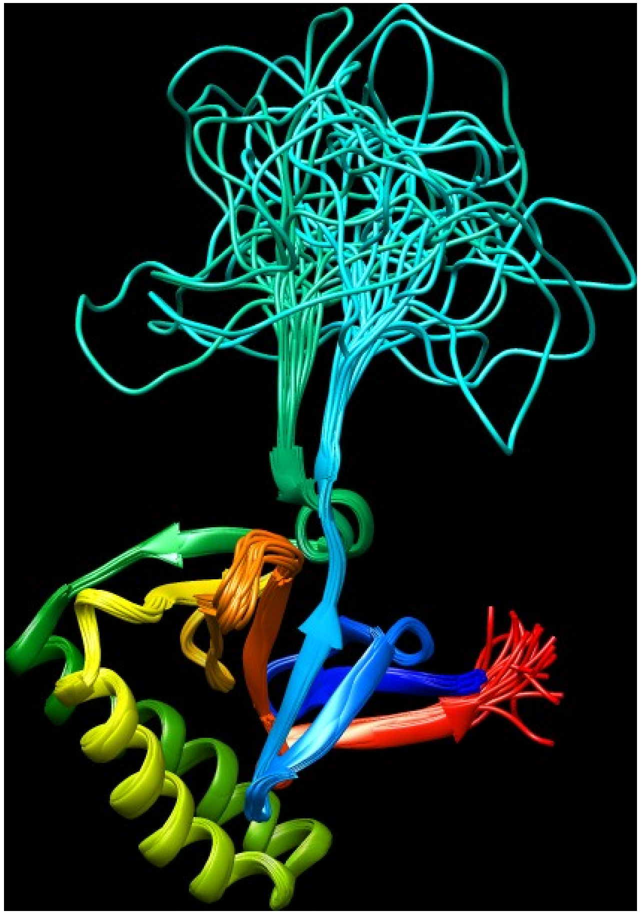
www.disprot.org

Where can we find disordered proteins?

In the PDB

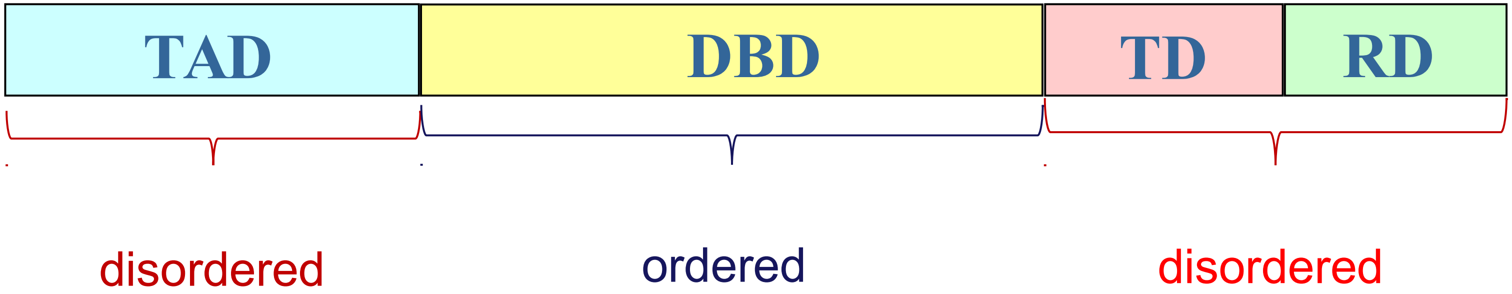


Missing electron density regions from the PDB



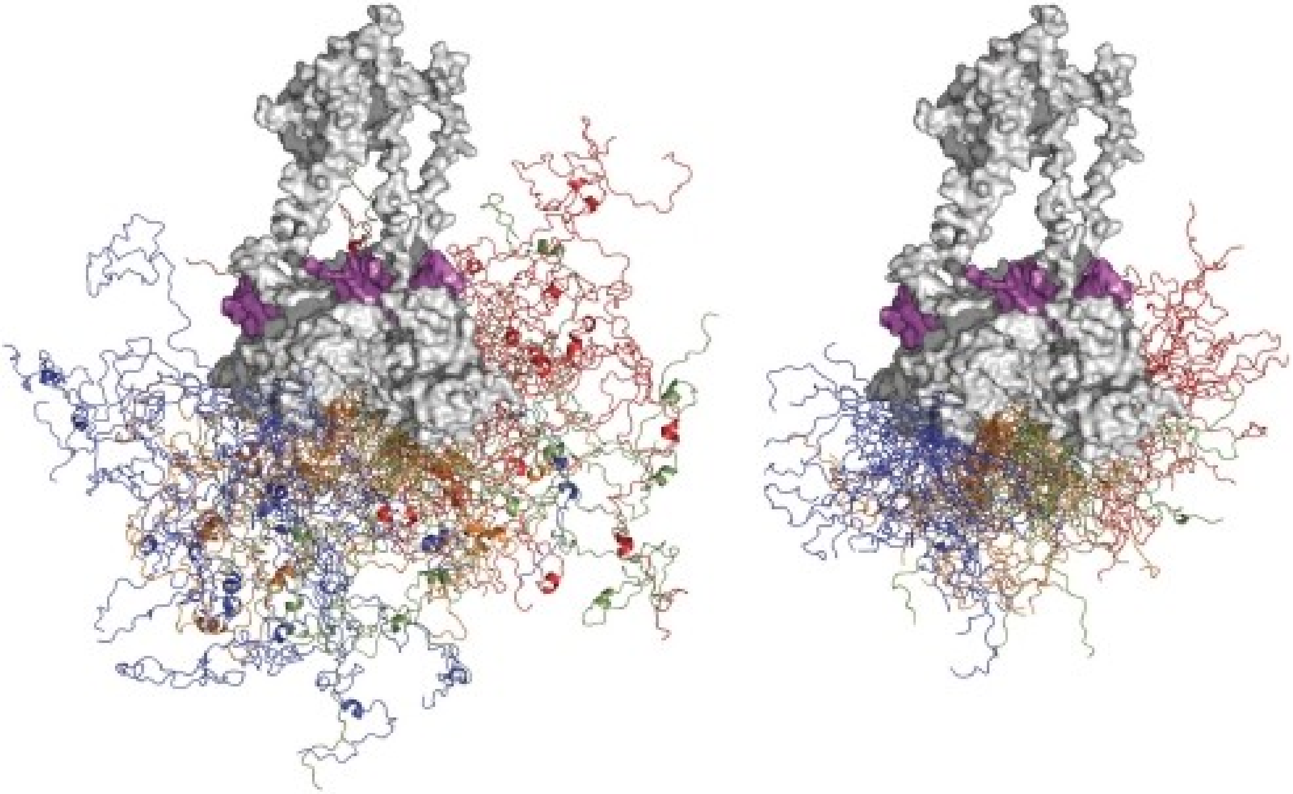
NMR structures with large structural variations

p53 tumor suppressor



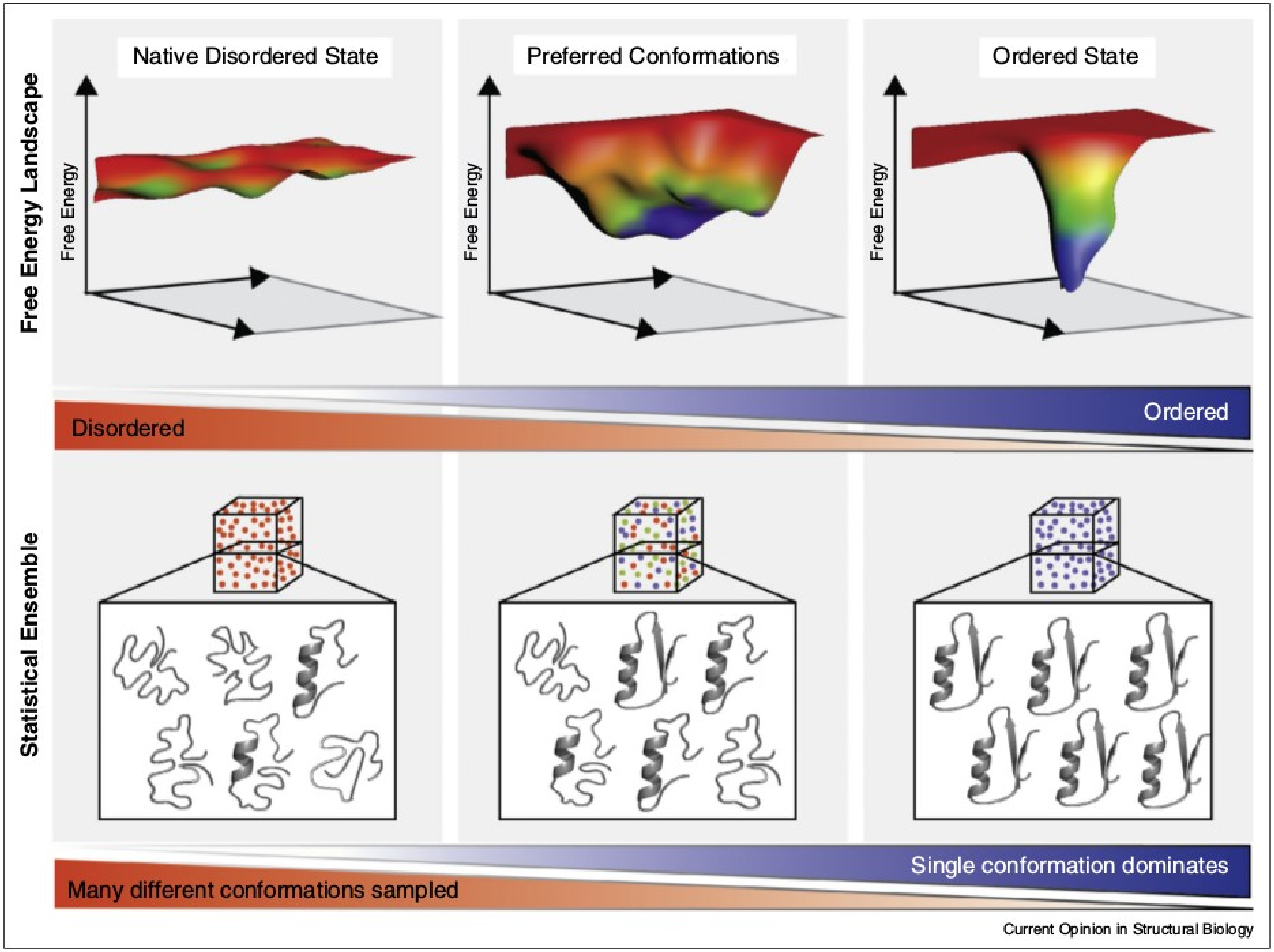
TAD DBD TD RD

disordered ordered disordered



Wells et al. PNAS 2008; 105: 5762

Funnels

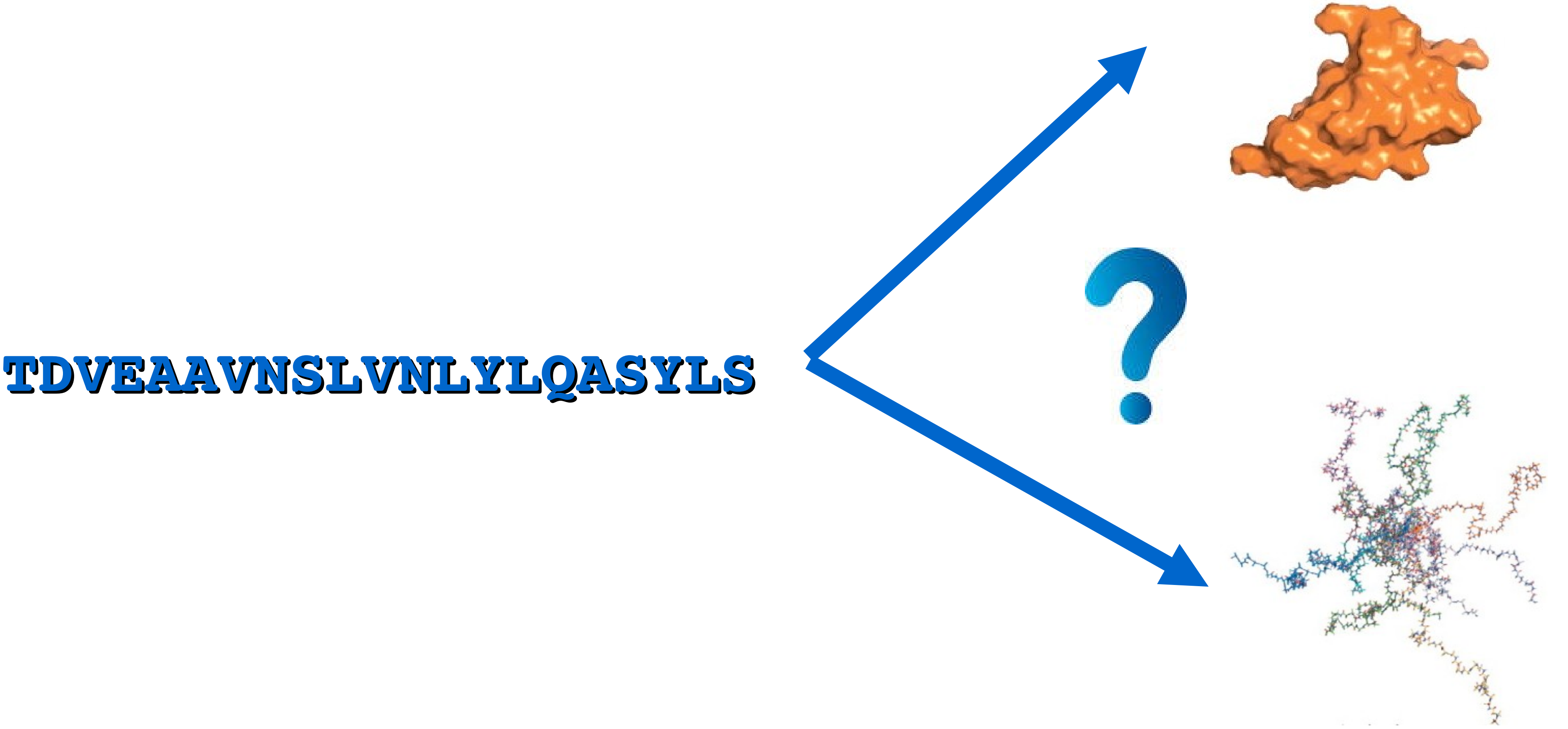


Flock et al Curr Opin Struct Biol. 2014; 26:62

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



TDVEAAVNSLVNLYLQASYLS

How can we discriminate ordered and disordered regions ?

Prediction: classification problem

Input

1. sequence
2. propensity vector
3. alignment (profile)
4. interaction energies

Method

1. statistical methods
2. machine learning
3. structural approach

Output (property)

1. binary
2. score

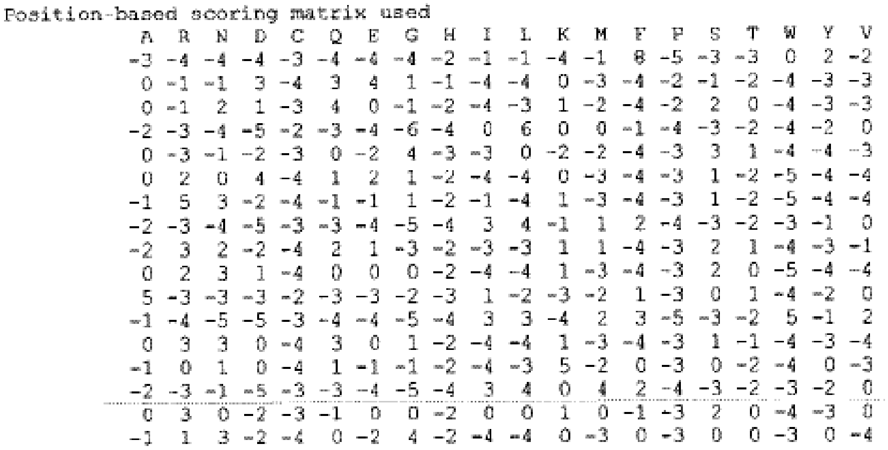
Training/Assessment

1. DisProt
2. PDB

DISOPRED2

…..AMDDLMLSPDDIEQWFTED…..

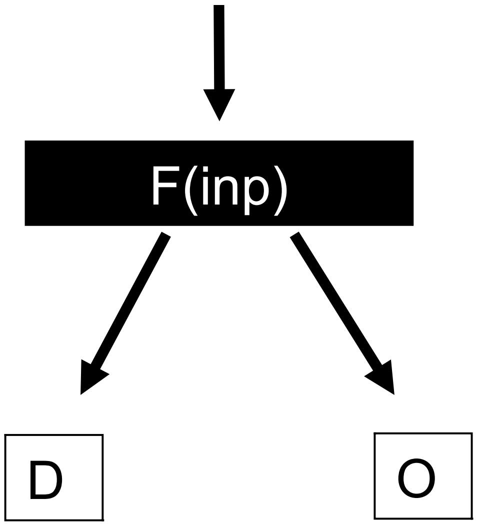
Raw profile from PSI-BLAST Log File



```
Position-based scoring matrix used
           A   R   N   D   C   Q   E   G   H   I   L   K   M   F   P   S   T   W   Y   V
          -3  -4  -4  -4  -3  -4  -4  -4  -2  -1  -1  -4  -1   8  -5  -3  -3   0   2  -2
           0  -1  -1   3  -4   3   4   1  -1  -4  -4   0  -3  -4  -2  -1  -2  -4  -3  -3
           0  -1   2   1  -3   4   0  -1  -2  -4  -3   1  -2  -4  -2   2   0  -4  -3  -3
          -2  -3  -4  -5  -2  -3  -4  -6  -4   0   6   0   0  -1  -4  -3  -2  -4  -2   0
           0  -3  -1  -2  -3   0  -2   4  -3  -3   0  -2  -2  -4  -3   3   1  -4  -4  -3
           0   2   0   4  -4   1   2   1  -2  -4  -4   0  -3  -4  -3   1  -2  -5  -4  -4
          -1   5   3  -2  -4  -1  -1   1  -2  -1  -4   1  -3  -4  -3   1  -2  -5  -4  -4
          -2  -3  -4  -5  -3  -3  -4  -5  -4   3   4  -1   1   2  -4  -3  -2  -3  -1   0
          -2   3   2  -2  -4   2   1  -3  -2  -3  -3   1   1  -4  -3   2   1  -4  -3  -1
           0   2   3   1  -4   0   0   0  -2  -4  -4   1  -3  -4  -3   2   0  -5  -4  -4
           5  -3  -3  -3  -2  -3  -3  -2  -3   1  -2  -3  -2   1  -3   0   1  -4  -2   0
          -1  -4  -5  -5  -3  -4  -4  -5  -4   3   3  -4   2   3  -5  -3  -2   5  -1   2
           0   3   3   0  -4   3   0   1  -2  -4  -4   1  -3  -4  -3   1  -1  -4  -3  -4
          -1   0   1   0  -4   1  -1  -1  -2  -4  -3   5  -2   0  -3   0  -2  -4   0  -3
          -2  -3  -1  -5  -3  -3  -4  -5  -4   3   4   0   4   2  -4  -3  -2  -3  -2   0
           0   3   0  -2  -3  -1   0   0  -2   0   0   1   0  -1  -3   2   0  -4  -3   0
          -1   1   3  -2  -4   0  -2   4  -2  -4  -4   0  -3   0  -3   0   0  -3   0  -4
```

Trained in missing residues from X-ray structures

SVM with linear kernel



F(inp)

D

O

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

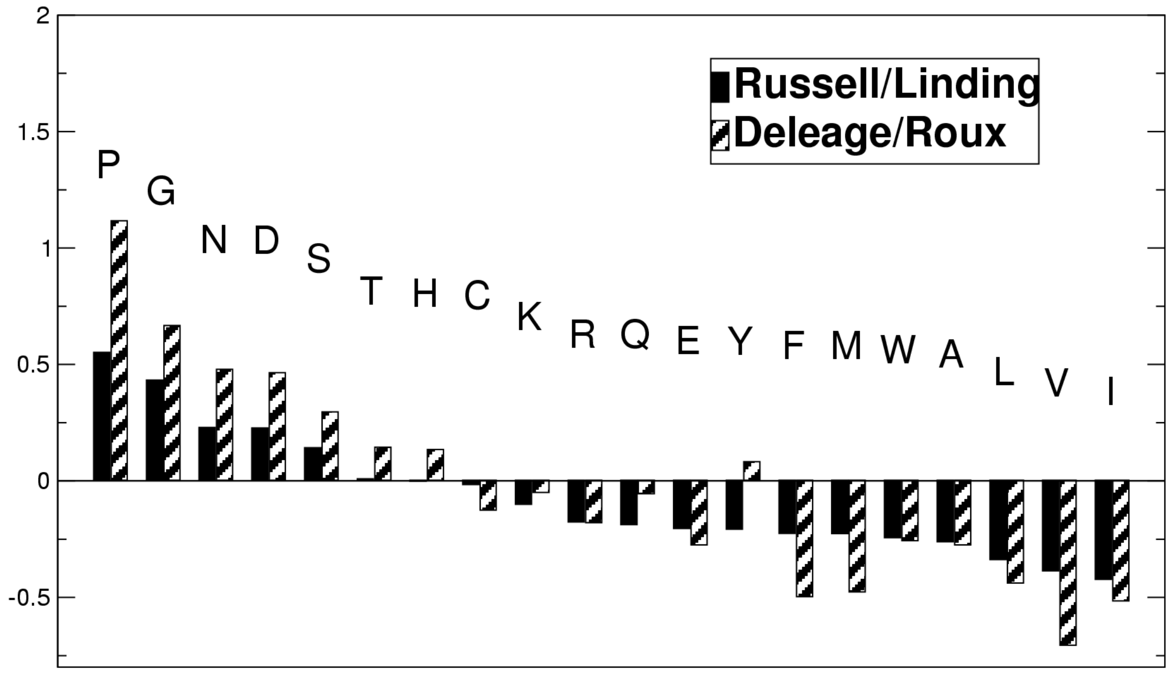
GlobPlot

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

Compare the tendency of amino acids:

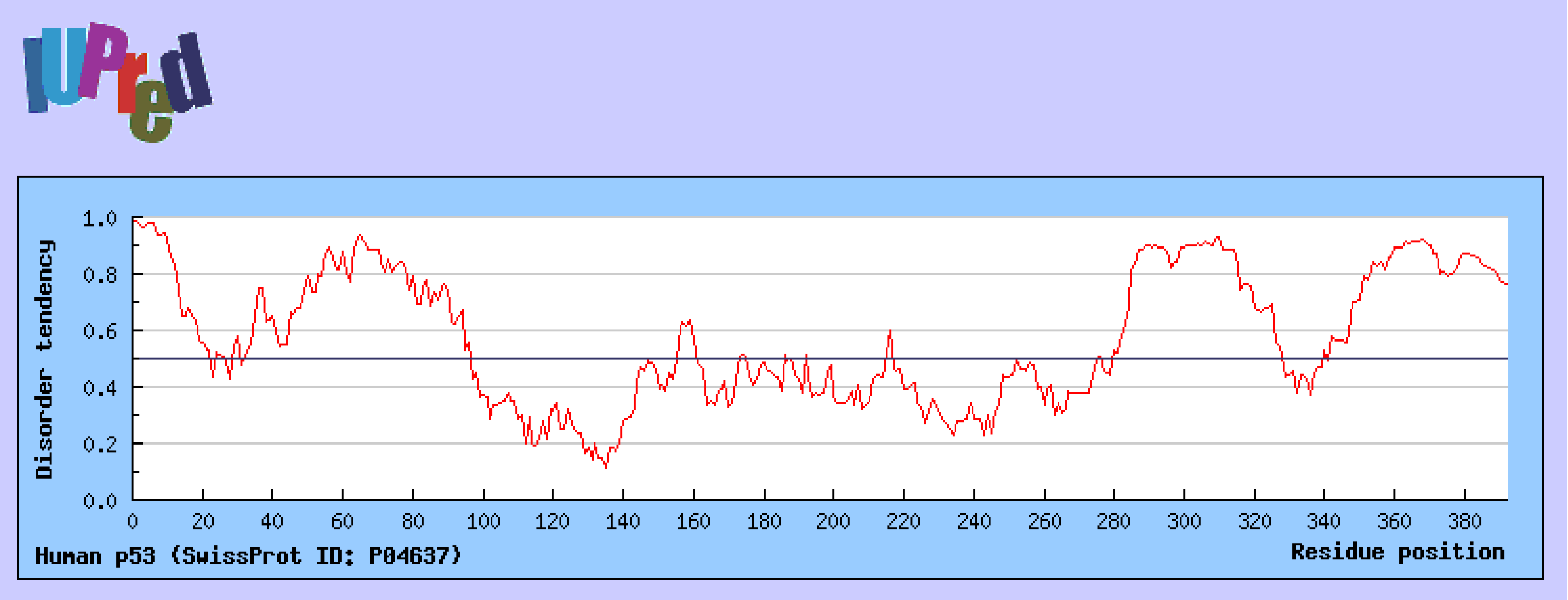
- to be in coil (irregular) structure.
- to be in regular secondary structure elements

Disorder propensity

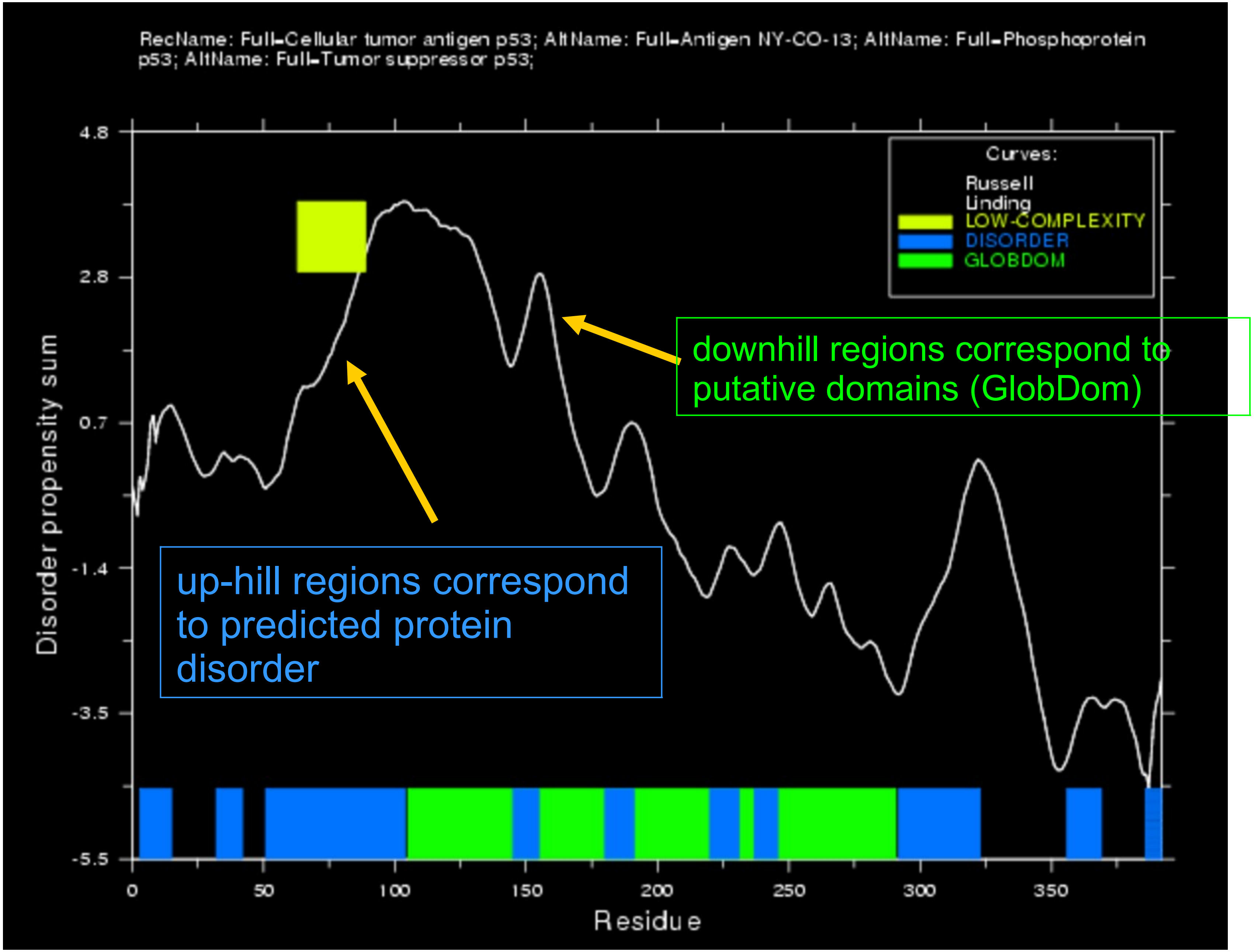


Linding (2003) *NAR* 31, 3701

Typical output



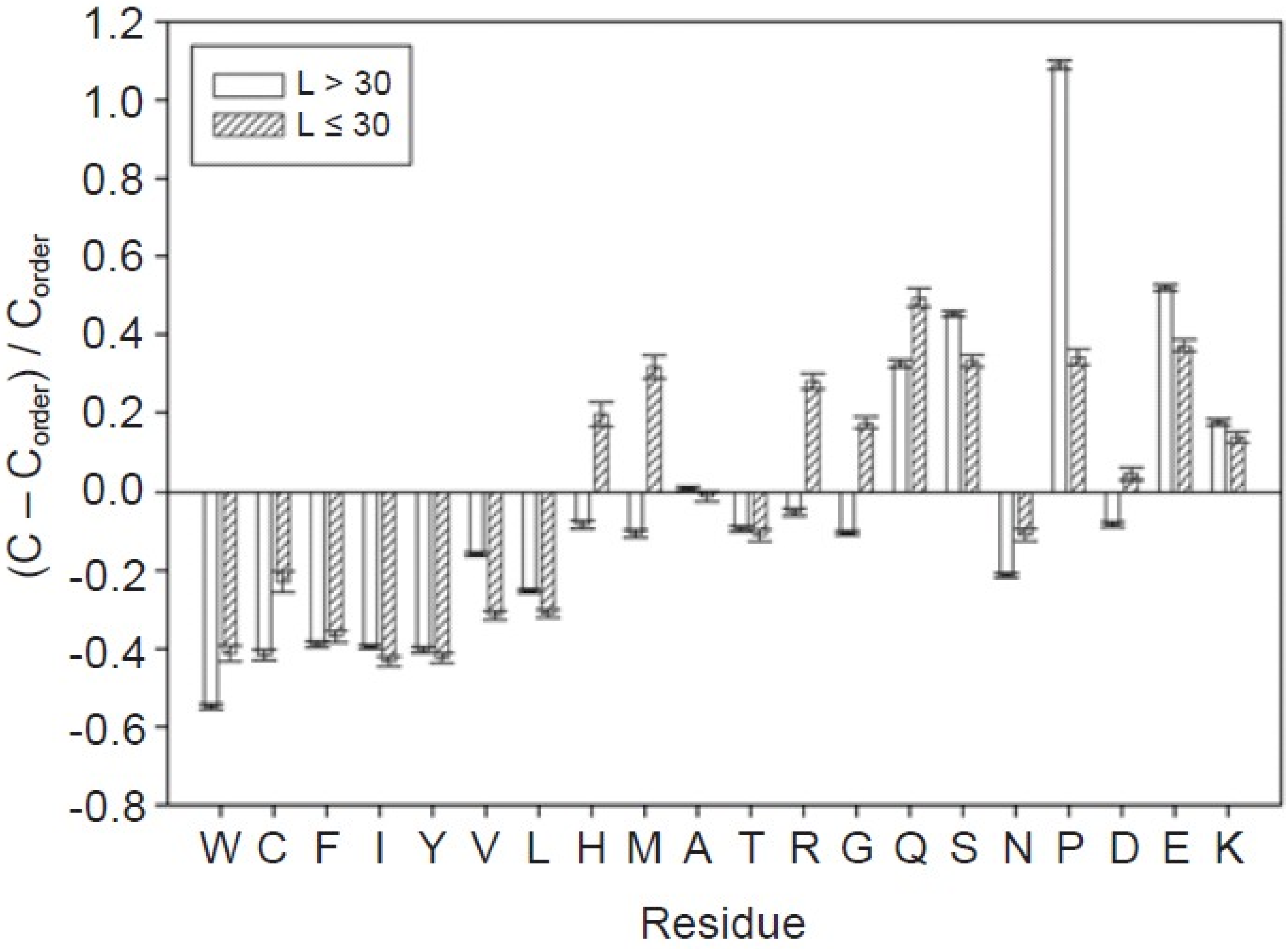
GlobPlot



downhill regions correspond to putative domains (GlobDom)

up-hill regions correspond to predicted protein disorder

Different flavors of disorder



Short and long disordered regions have different compositional biases

PONDR VSL2

Differences in short and long disorder

- amino acid composition
- Short disorder is often at the termini
- methods trained on one type of dataset tested on other dataset resulted in lower efficiencies

- Short version – Long version

- PONDR VSL2:

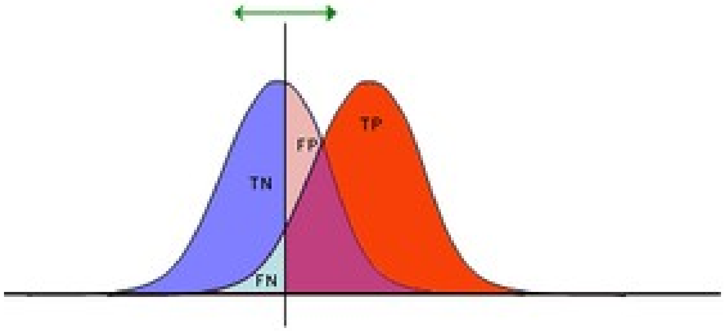
separate predictors for short and long disorder

combined

length independent predictions

Peng (2006) BMC Bioinformatics 7, 208

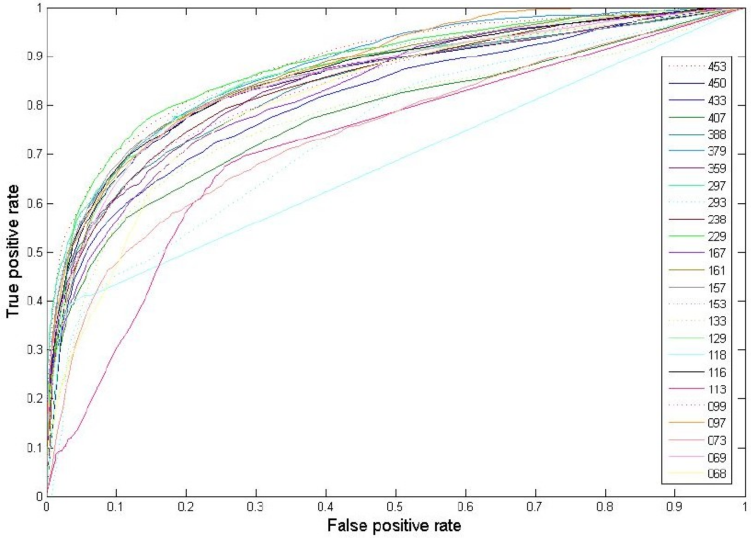
Evaluation



TP	FP
FN	TN
1	1

$$Acc = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right),$$

ROC curve



For each value of P in increments of 0.01 the TP-rate & the FP-rate are calculated, and the ‘Area Under Curve’ (AUC) score is calculated.

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

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

Functions of IDPs

I Entropic chains

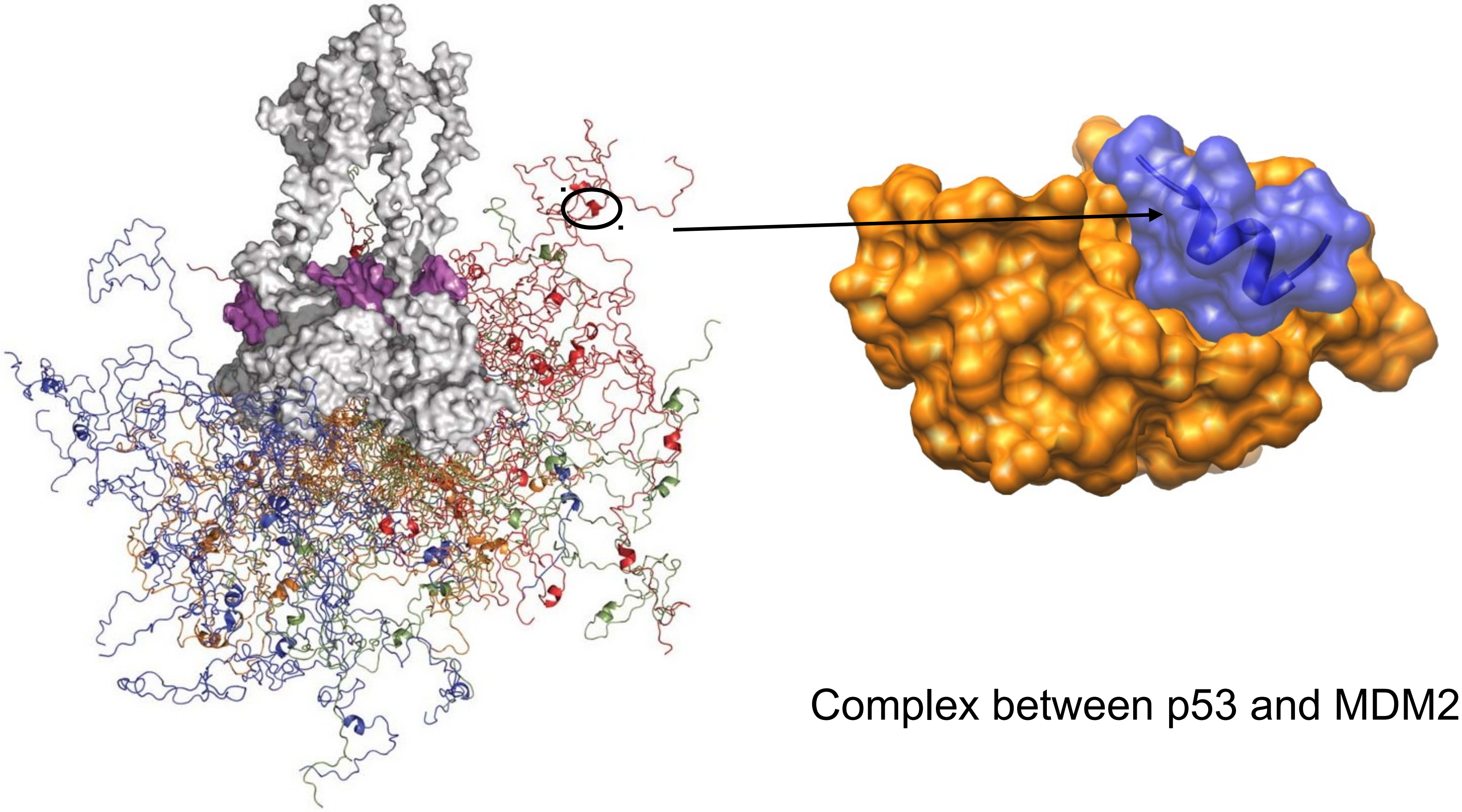
II Linkers

III Molecular recognition

IV Protein modifications (e.g. phosphorylation)

V Assembly of large multiprotein complexes

Protein interactions of IDPs



Complex between p53 and MDM2

Coupled folding and binding

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

Signaling
Regulation

Binding regions within 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
- Computational methods

Binding regions within IDPs

- **SLIMs: Short linear motifs**

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

  although enriched in IDRs, around 20% are in located within IDRs

- **Disordered binding regions, Morfs**

  undergo disorder to order transition upon binding

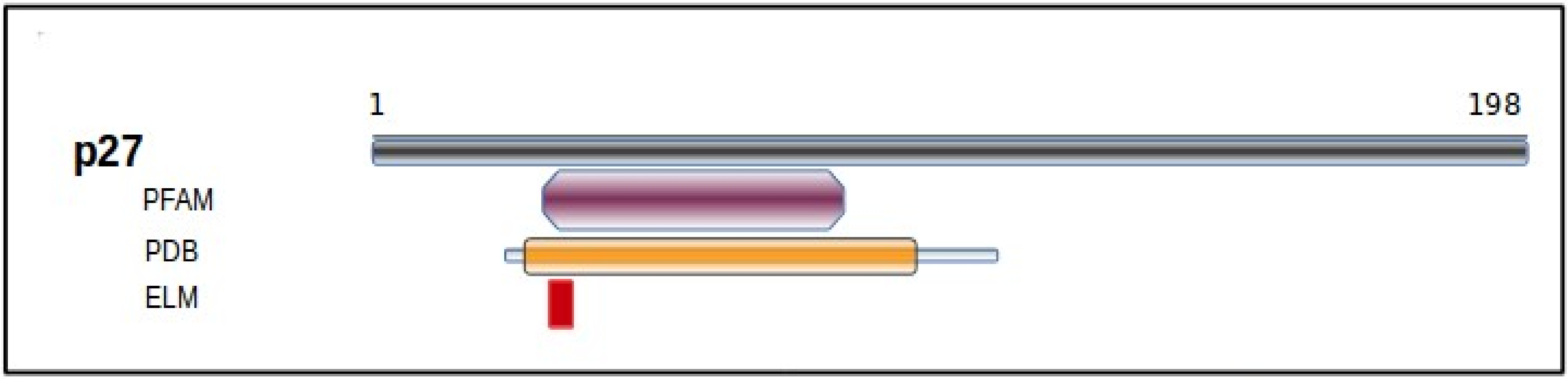
  usually less then 30 residues, can be up to 70

- **Intrinsically disordered domains**

  evolutionary conserved disordered segments

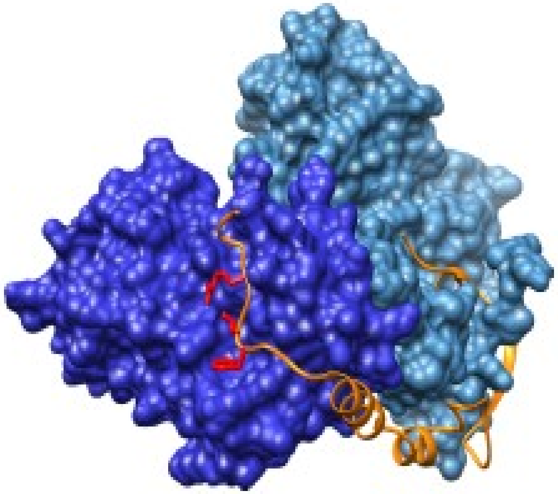
p27

Inhibitor of CDK2-CyclinA complex.



[RK].L.{0,1}[FYLIVMP]

CDN1B_HUMAN	30-33	HPKPSACRNLFGPVDHEEL
MPIP1_HUMAN	11-15	PEPPHRRRLLFACSPPPAS
CDC6_HUMAN	94-98	HSHTLKGRRLVFDNQLTIK
RB_HUMAN	873-877	SNPPKPLKKLRFDIEGSDE
P53_HUMAN	381-385	GQSTSRHKKLMFKTEGPDS
VE1_HPV18	127-130	SGQKKAKRRLFTISDSGYG



Bioinformatical approaches

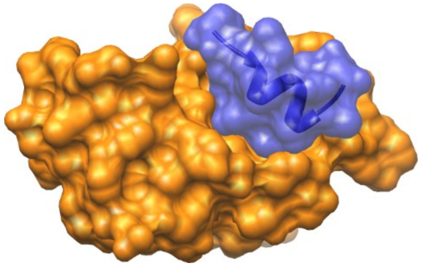
(~10, as opposed to the more than 50 disorder prediction methods)

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

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

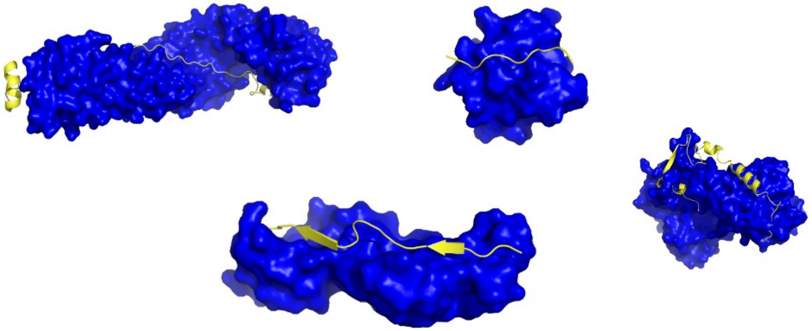
  (Regular Expression, PSSMs)
- □ Conservations patterns (**SlimPrints, PhyloHMM**)

Prediction of 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



Prediction of disordered binding regions – ANCHOR

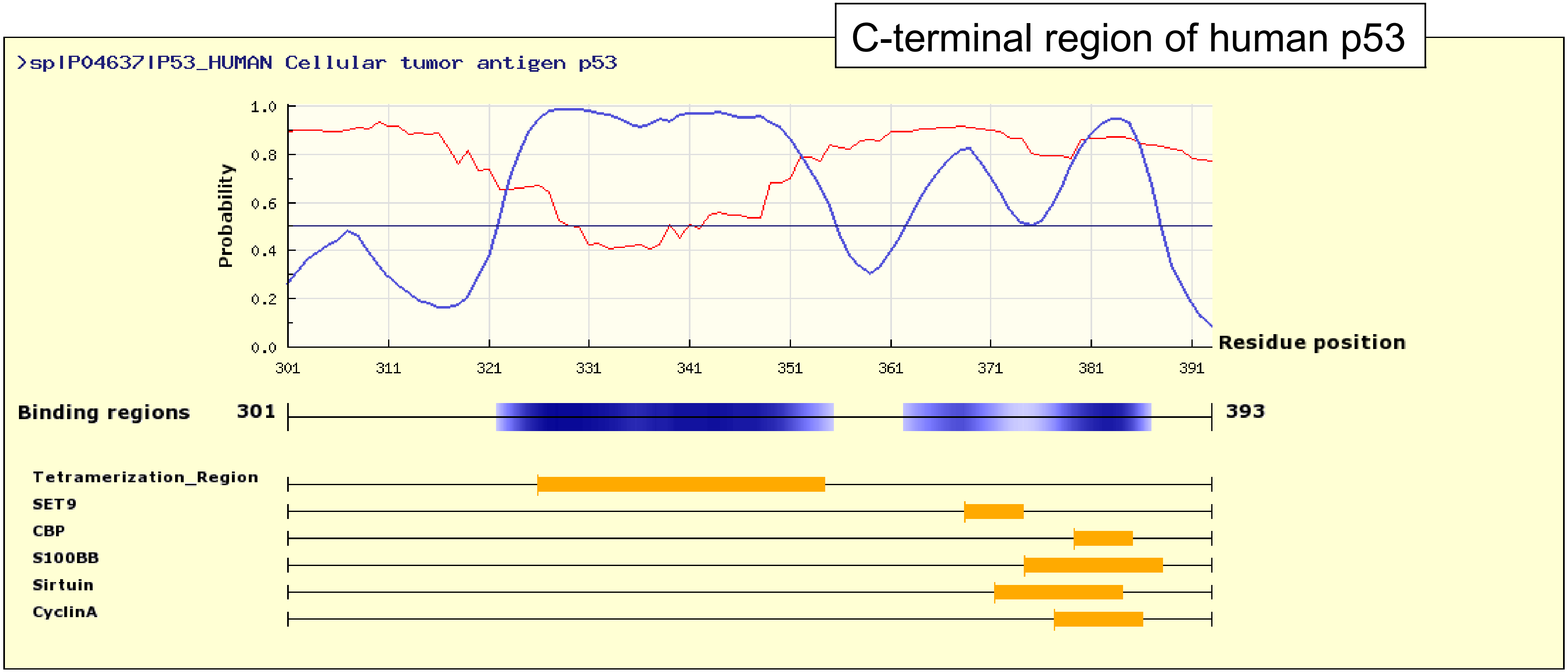
What discriminates disordered binding regions?

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

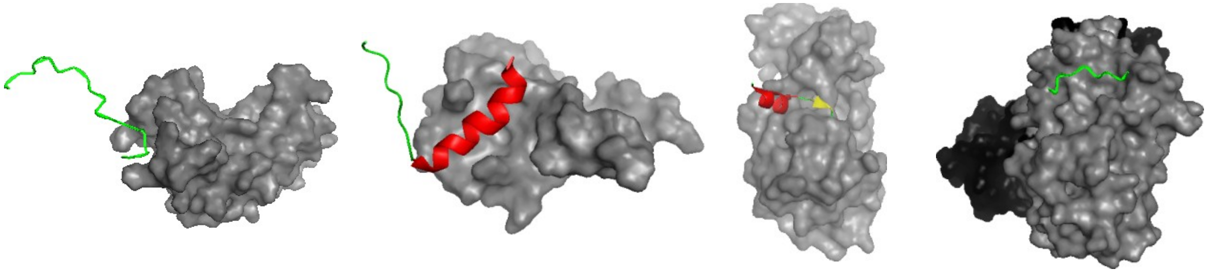
Based on simplified physical model

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

ANCHOR



C-terminal region of human p53



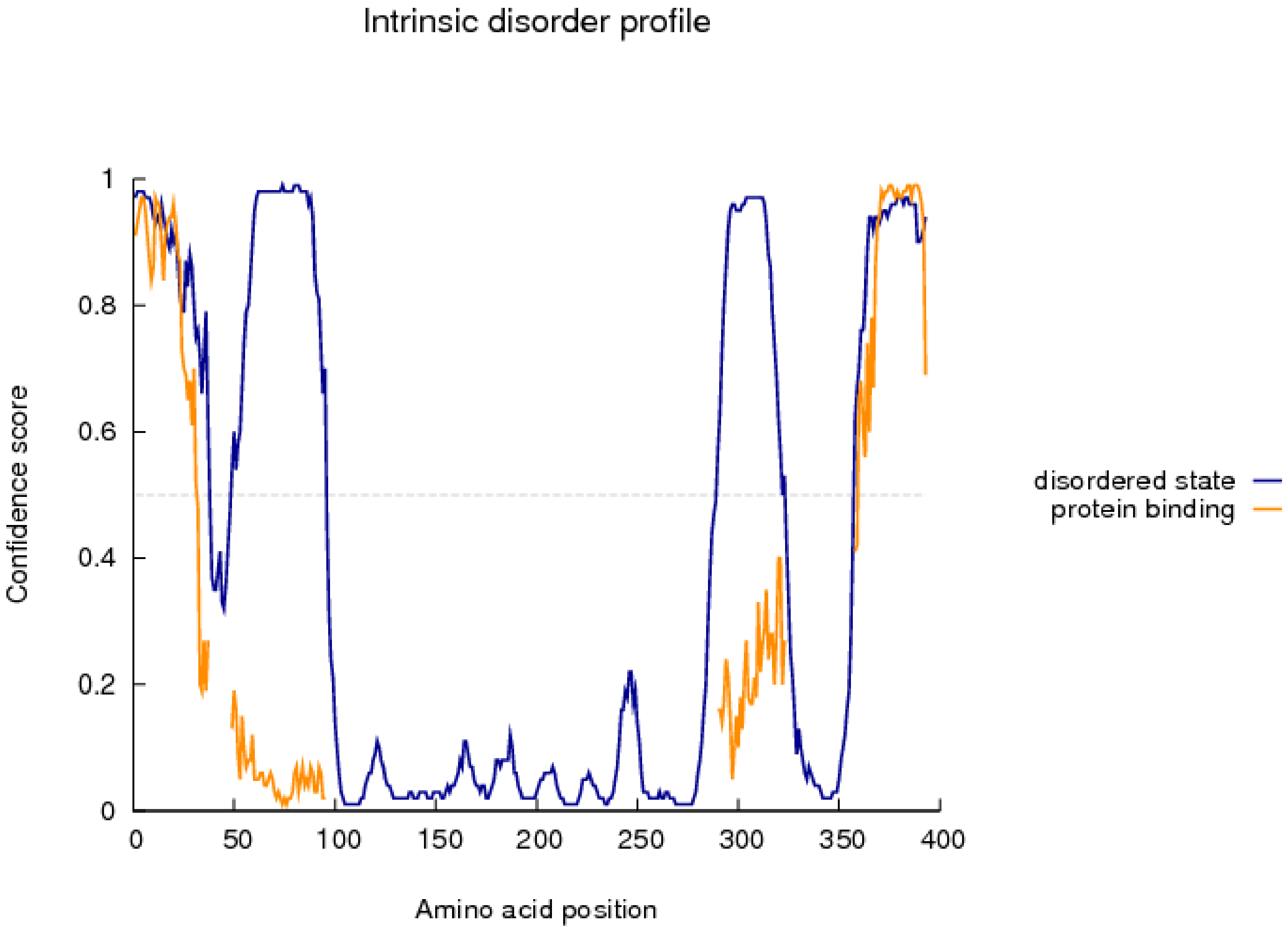
Dosztanyi et al. Bioinformatics. 2009;25:2745

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

Prediction of binding regions within IDPs

- Combined predictions provide more biologically meaningful predictions

- Lot of rooms for improvements...

- What is the binding partner?