Prediction of protein disorder

Zsuzsanna Dosztányi
MTA-ELTE Momentum Bioinformatics Group
Department of Biochemistry, Eotvos Lorand University,
Budapest, Hungary
dosztanyi@caeser.elte.hu

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

Bioinformatics of protein disorder

- ***Part 1*** Prediction of protein disorder
  - Databases
  - Prediction of protein disorder
- ***Part 2*** Biology of disordered proteins
  - Evolutionary and functional characteristics of IDPs
  - Prediction of functional regions within IDPs

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

JMB

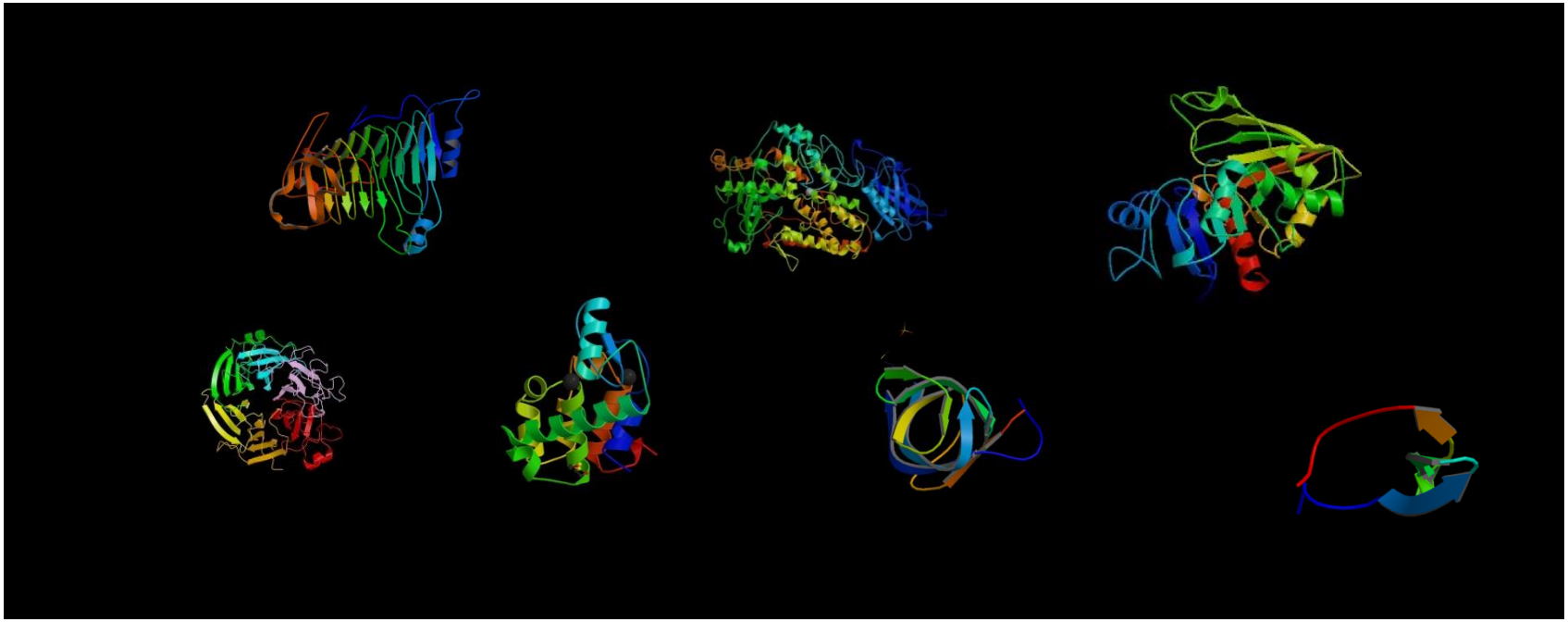
Intrinsically Unstructured Proteins: Re-assessing the Protein Structure-Function Paradigm

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

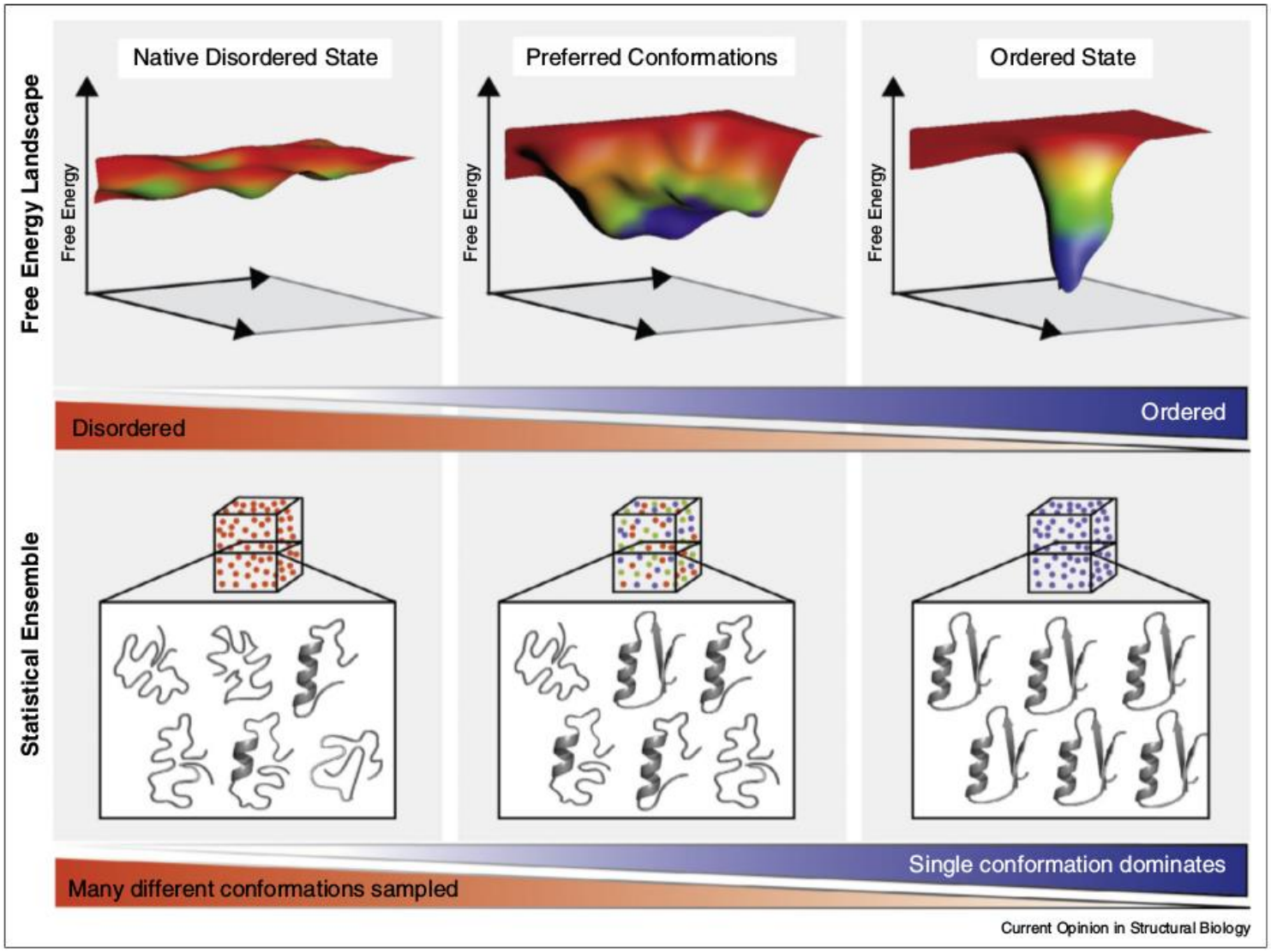
A major challenge in the post-genome era will be determination of the functions of the encoded protein sequences. Since it is generally assumed that the function of a protein is closely linked to its three-dimensional structure, prediction or experimental determination of the library of protein structures 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-

Protein Structure/Function Paradigm



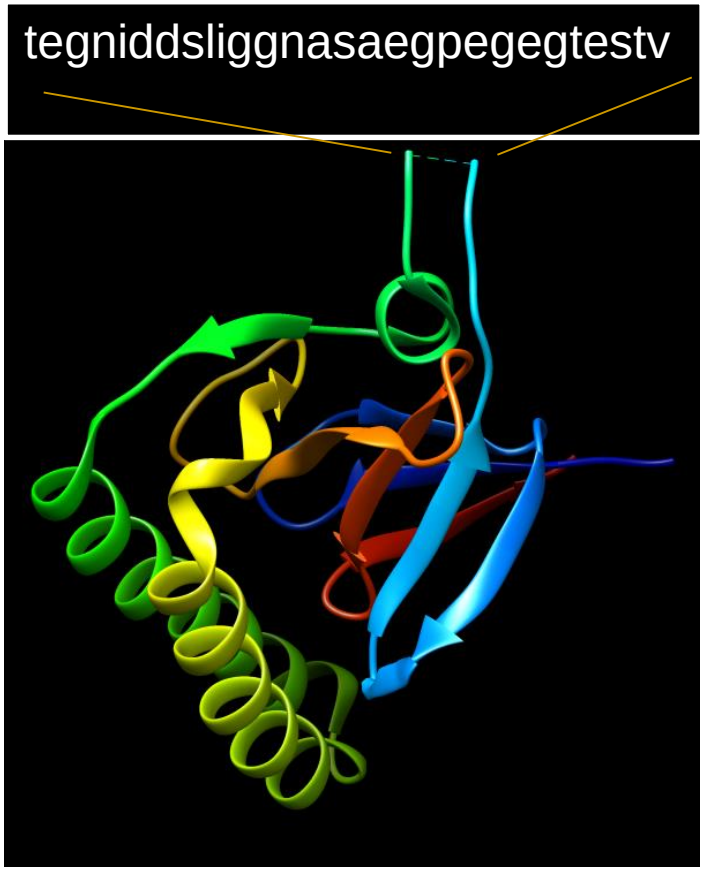
Dominant view: 3D structure is a prerequisite for protein function

Funnels

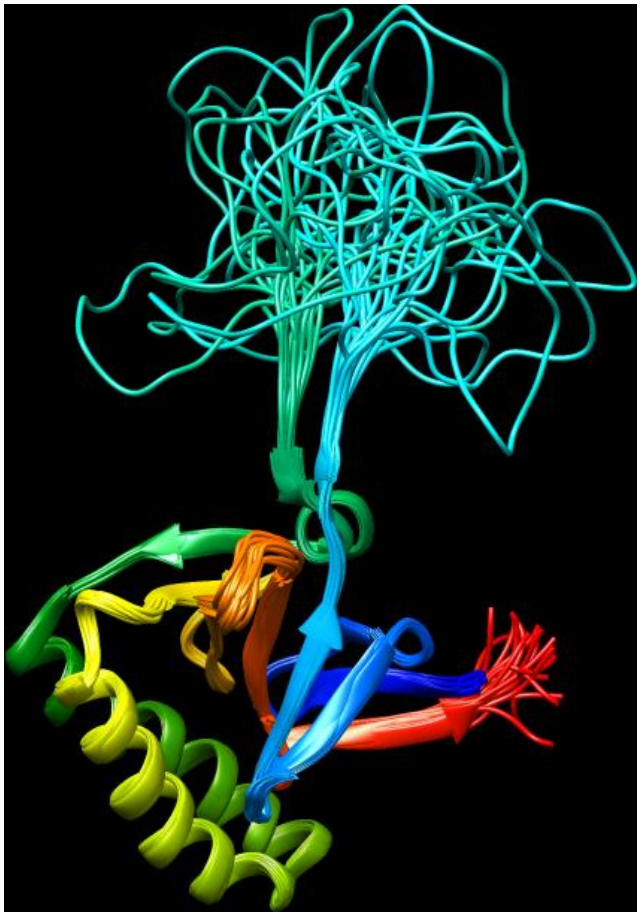

Free Energy Landscape
Native Disordered State
Preferred Conformations
Ordered State
Free Energy
Free Energy
Free Energy
Disordered
Ordered
Statistical Ensemble
Many different conformations sampled
Single conformation dominates
Current Opinion in Structural Biology

Where can we find disordered proteins?

In the PDB



Missing electron density regions from the PDB



NMR structures with large structural variations

Experimental detection of disorder

In the literature

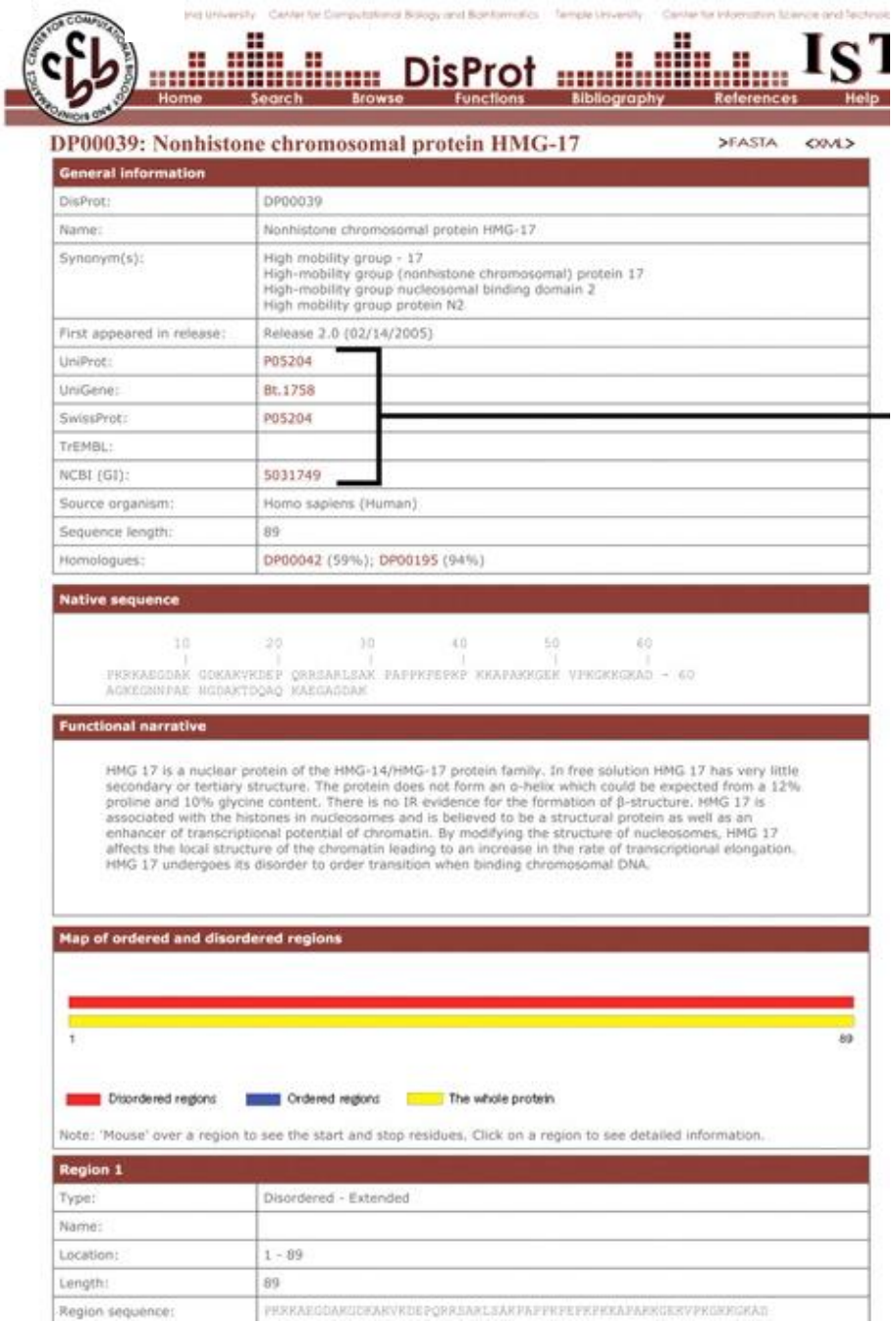
- Failed attempts to crystallize
- Lack of NMR signals
- Heat stability
- Protease sensitivity
- Increased molecular volume
- “Freaky” sequences …

Disprot

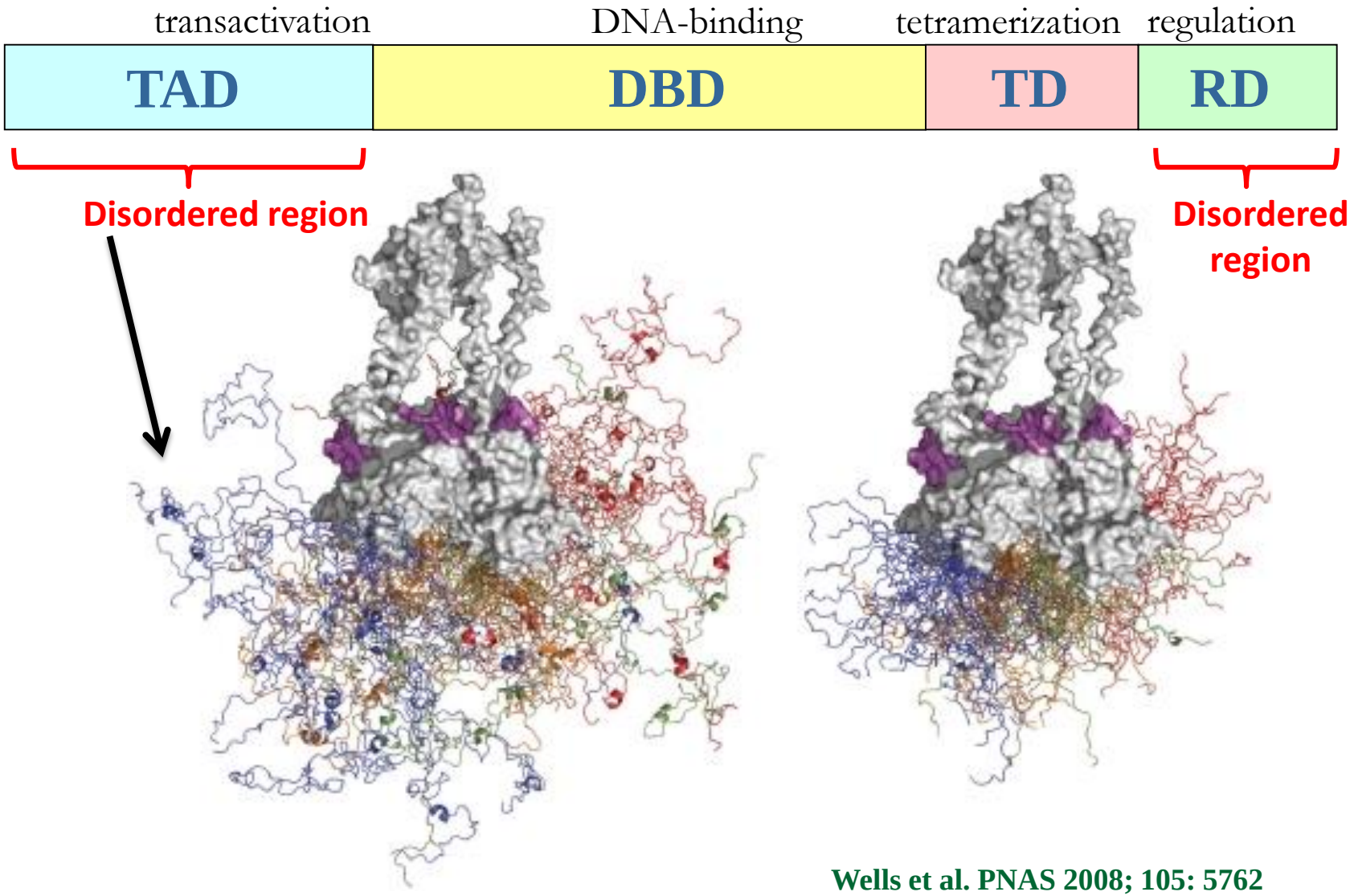
www.disprot.org

Current release: **6.02**
Release date: **05/24/2013**
Number of proteins: **694**
Number of disordered regions: **1539**

Experimentally verified disordered proteins collected from literature (X-ray, NMR, CD, proteolysis, SAXS, heat stability, gel filtration, …)



p53 tumor suppressor



Wells et al. PNAS 2008; 105: 5762

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 …

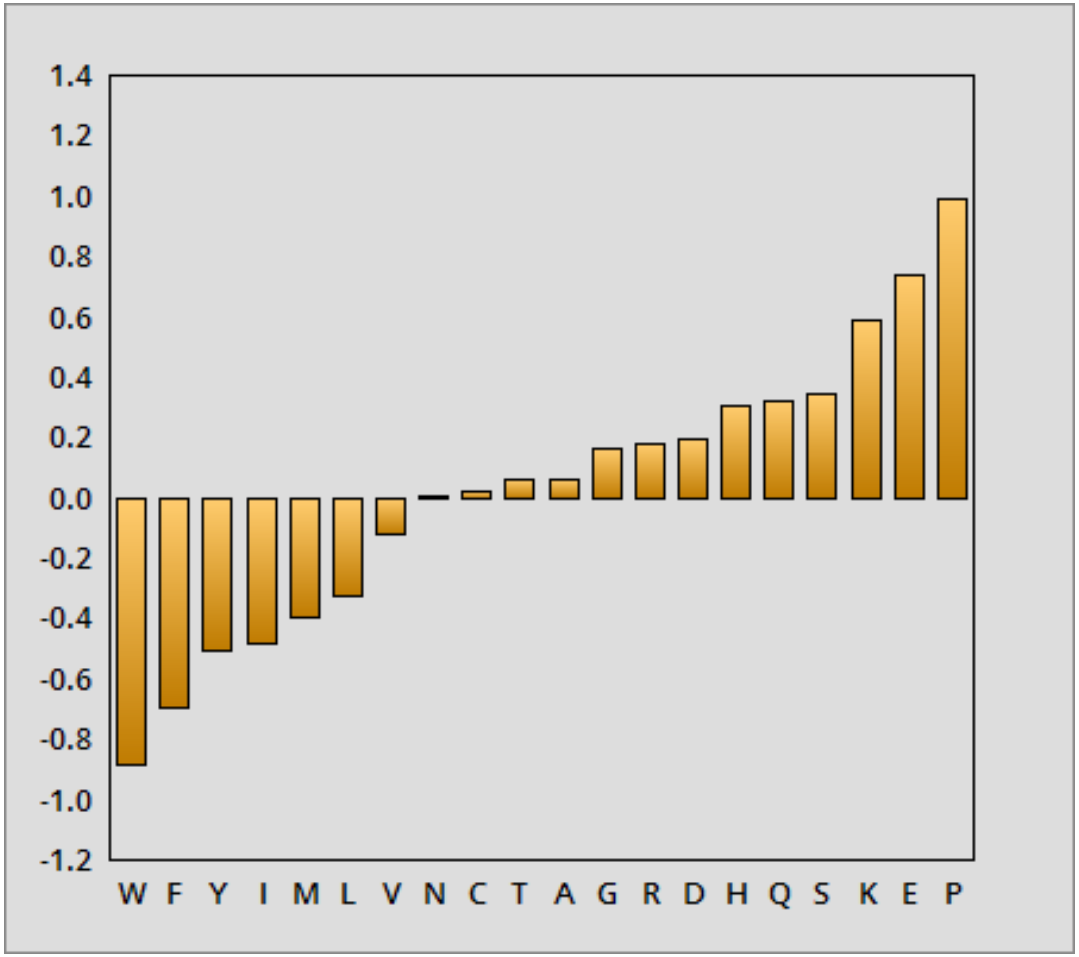
- Based on amino acid propensity scales or on simplified biophysical models
  - **GlobPlot**, FoldIndex, FoldUnfold, **IUPred**, UCON, **TOP-IDP**

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

TOP-IDPs

The amino acid propensity scale that discriminates ordered from disordered proteins

TOP-IDP

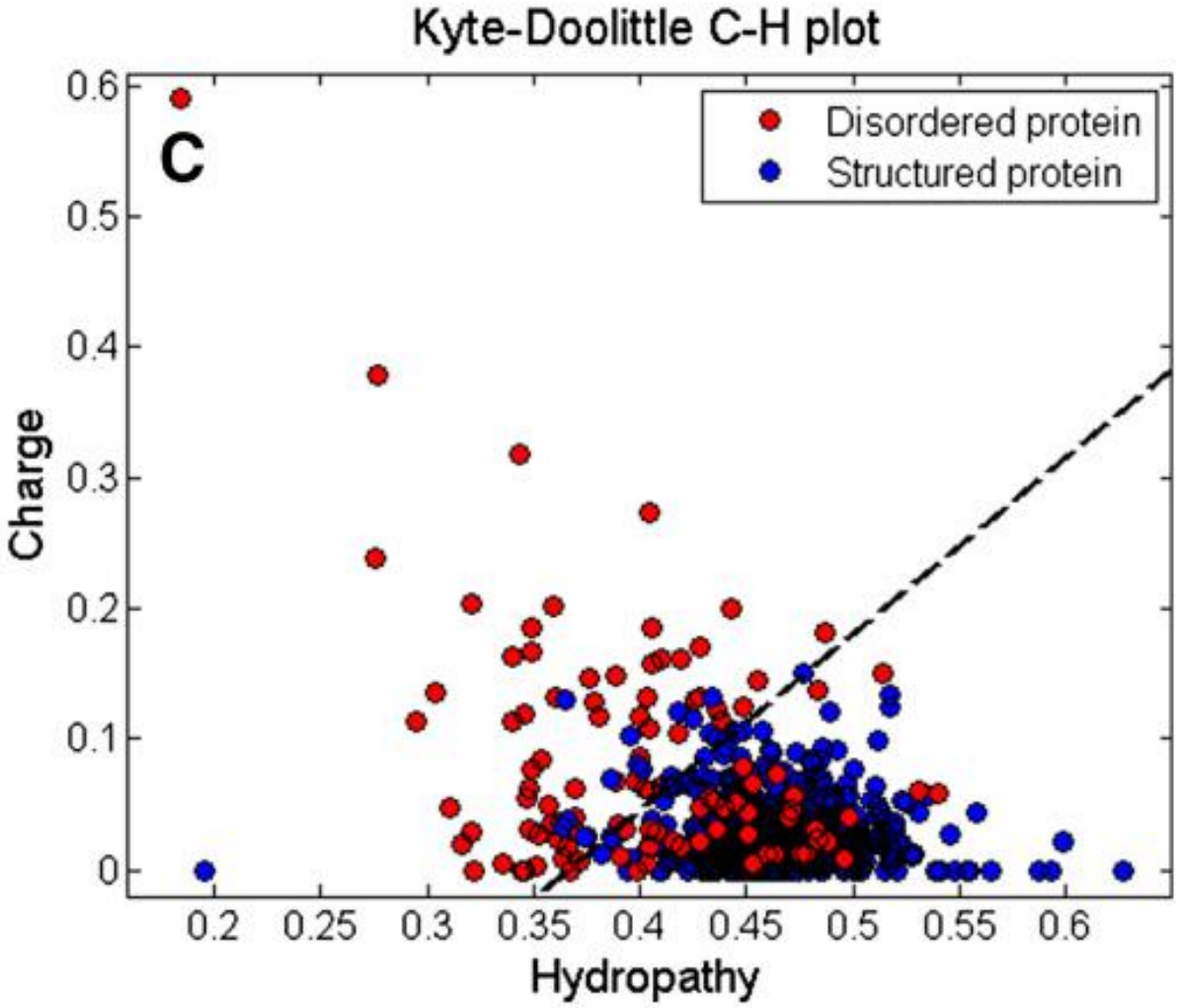


Campen et al. Protein Pept Lett. 2008; 15:956-963

Charge-hydropathy plot

Globular proteins have a hydrophobic core and charged residues are compensated by oppositely charged residues

Charge-hydropathy plot



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

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
```

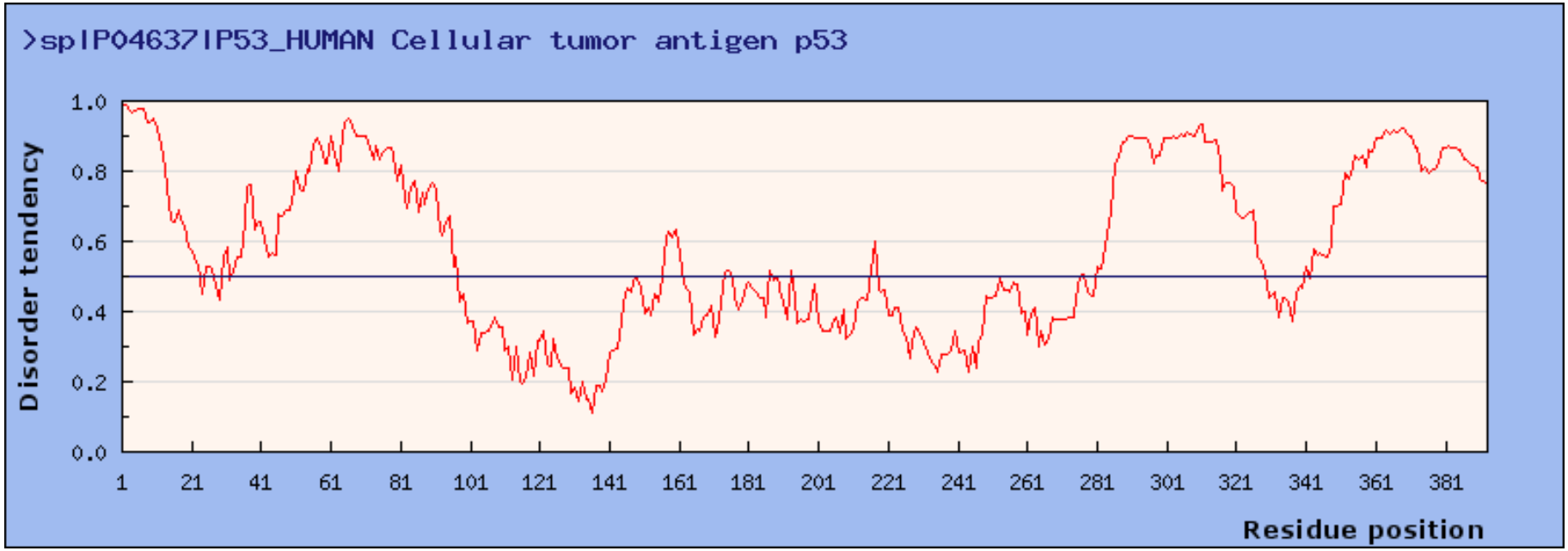
Amino acid composition (***n***)

```
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

$E\ (\ estimated\ \ \ \ )\ /\ L$

A typical output (IUPred)



Predictions are on a per residue basis

Dosztanyi (2005) ***JMB*** **347, 827**

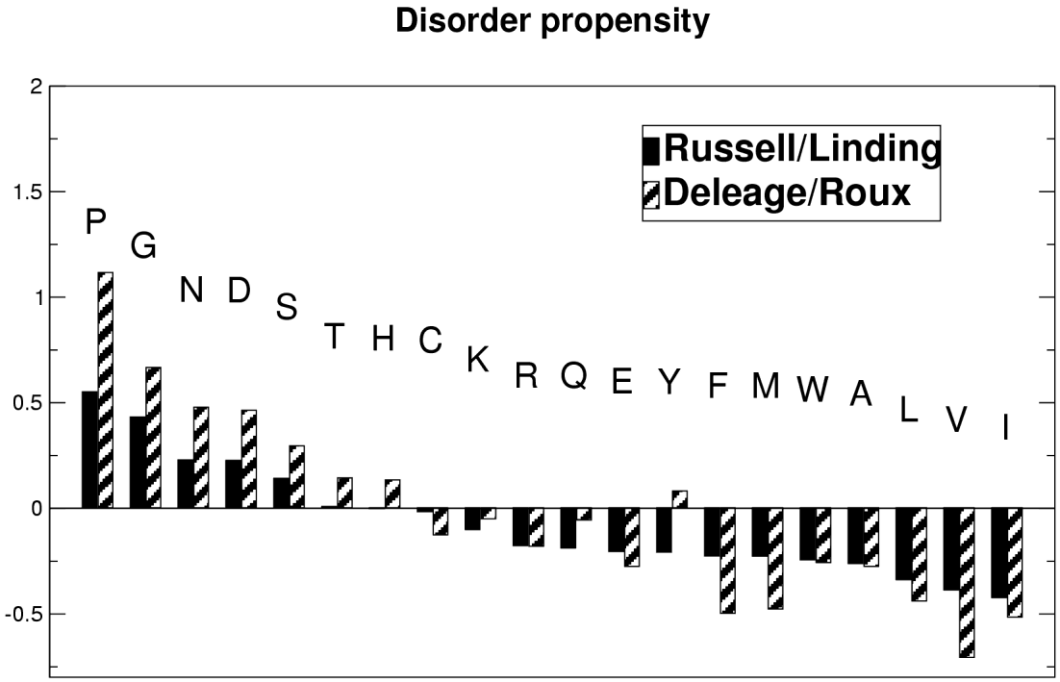
GlobPlot

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

GlobPlot

Compare the tendency of amino acids:

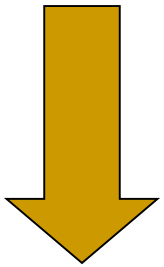
- to be in coil (irregular) structure.
- to be in regular secondary structure elements



Linding (2003) *NAR* 31, 3701

A non typical output (GlobPlot)

From position specific predictions

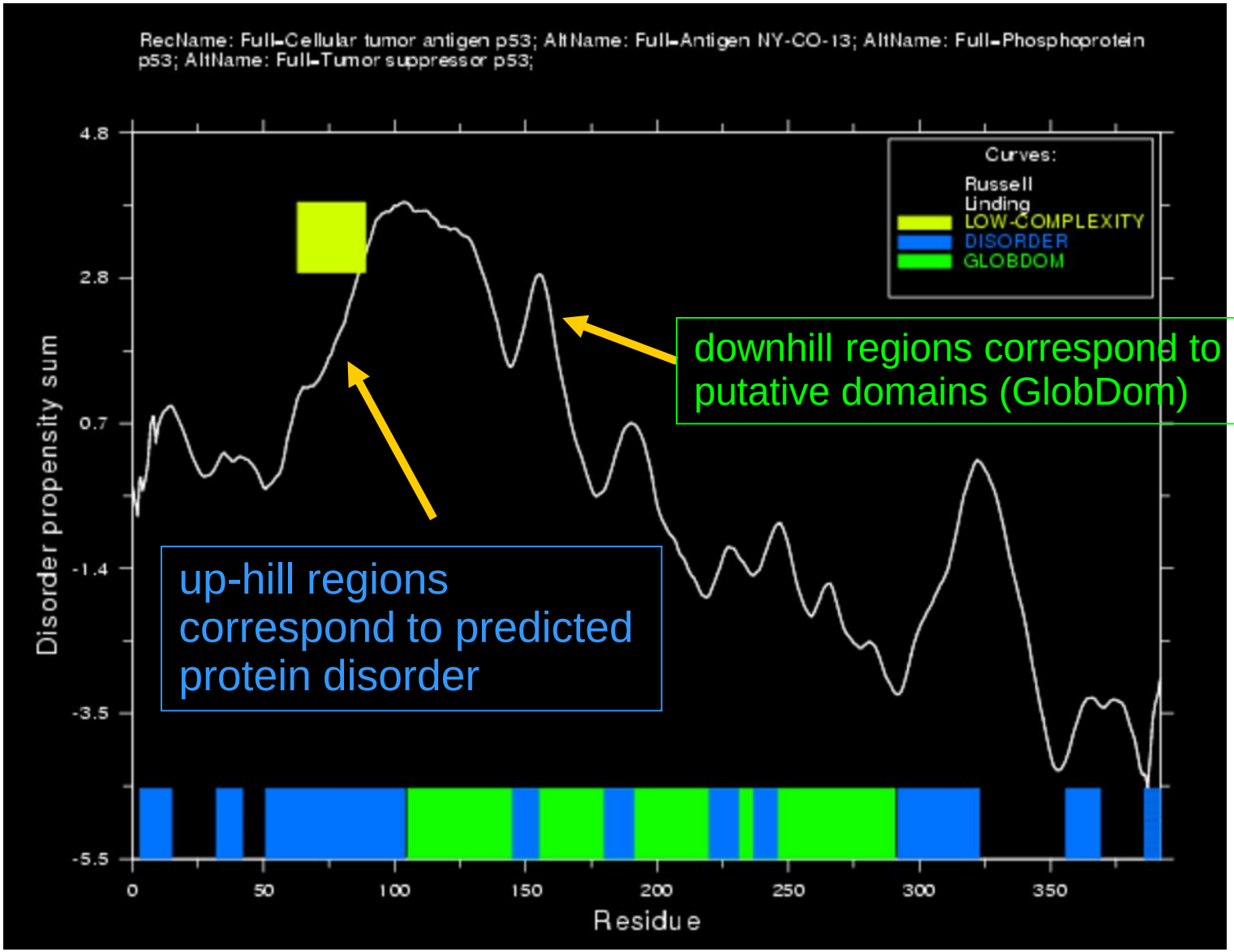


Where are the ordered domains?

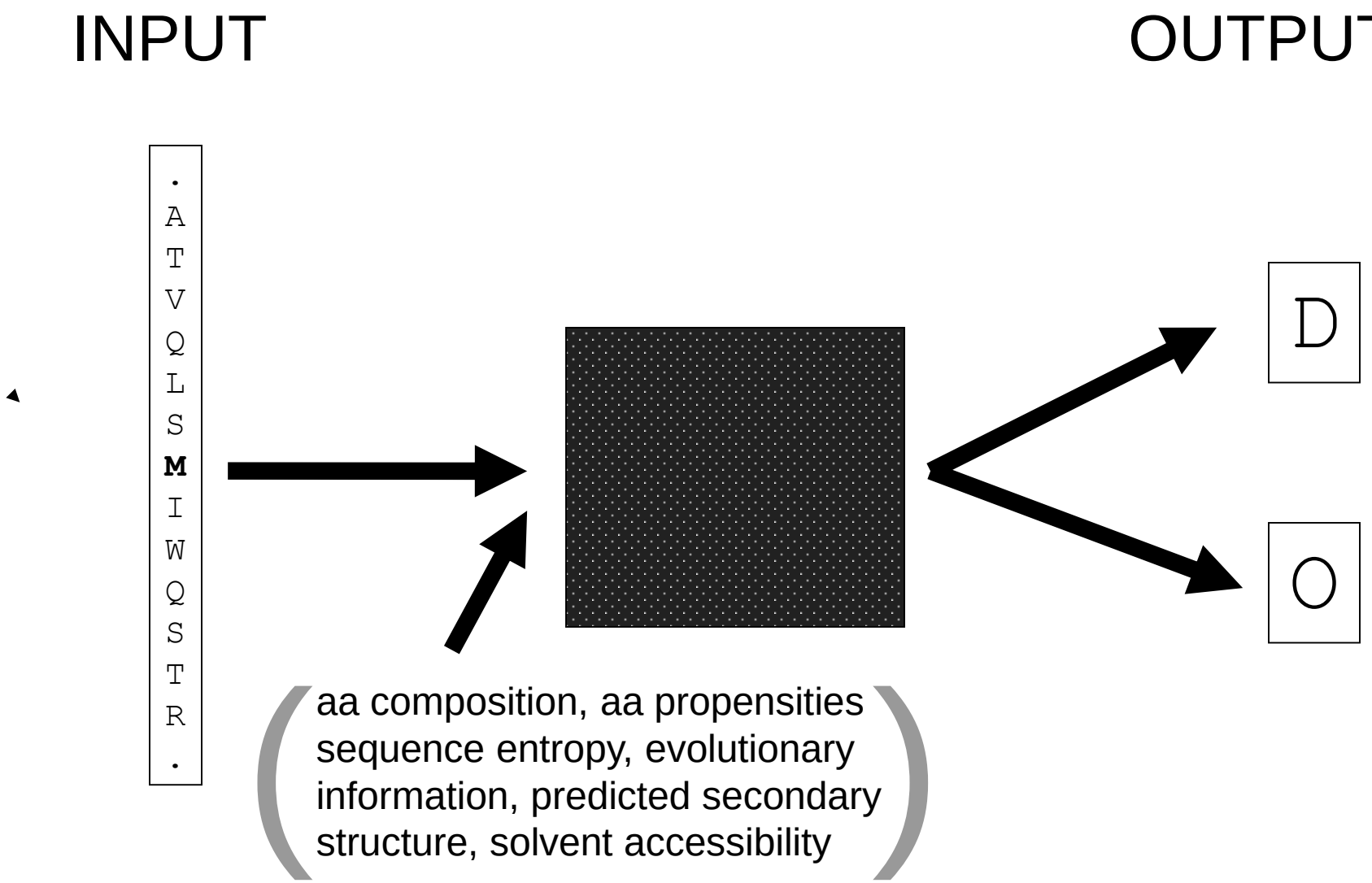
Longer disordered segments?

(Noise vs. real data)

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
Disorder propensity sum
4.8
2.8
0.7
-1.4
-3.5
-5.5
0
50
100
150
200
250
300
350
Residue
downhill regions correspond to putative domains (GlobDom)
up-hill regions correspond to predicted protein disorder

Machine learning approaches



PONDR VSL2

Differences in short and long disorder

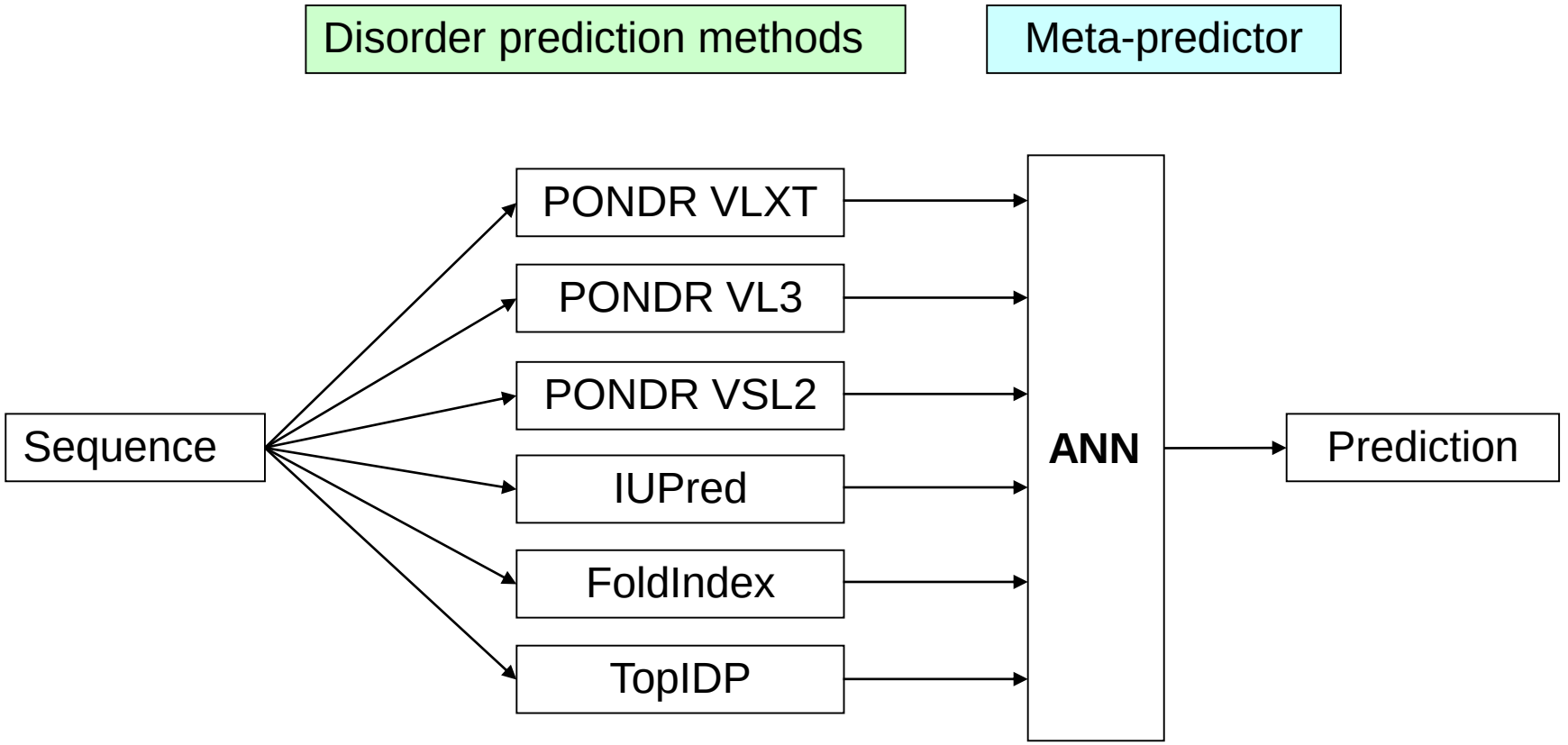
- amino acid composition
- methods trained on one type of dataset tested on other dataset resulted in lower efficiencies

PONDR VSL2: separate predictors for short and long disorder ***combined***

length independent predictions

Peng (2006) *BMC Bioinformatics* 7, 208

Metaservers:



Xue et al. Biochem Biophys Acta. 2010; 180: 996

Accuracy

- True positive: Disordered residues are predicted as disordered
- False positive: Ordered residues predicted as disordered
- True negative: Ordered residues predicted as ordered
- False negative: Disordered residues predicted as ordered

$$Sensitivity = \frac{TP}{TP + FN} = \frac{TP}{N_{disorder}}$$

$$Specificity = \frac{TN}{TN + FP} = \frac{TN}{N_{order}}$$

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

75-90%

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

Genome level annotations

- 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

DisProt	PDB	Predictors	Consensus
Disorder	Disorder	Any	Disorder
Disorder	Structure	Any	Ambiguous
Disorder	Ambiguous	Any	Ambiguous
Structure	Disorder	Any	Ambiguous
Structure	Structure	Any	Structure
Structure	Ambiguous	Any	Ambiguous
Ambiguous	Any	Any	Ambiguous
None	Disorder	Any	Disorder
None	Structure	Any	Structure
None	Ambiguous	Any	Ambiguous
None	None	Disorder	Disorder (LC)
None	None	Structure	Structure (LC)

Potenza et el. NAR 2014;

IDP prediction and other 1D prediction methods

- Secondary structure prediction methods
  - Coil is an ordered, irregular structural element
  - Disordered proteins usually do not contain stable secondary structural element (e.g. by CD)
  - They can contain transient secondary structure elements (by NMR)
  - Use secondary structure predictions methods for disordered proteins with extreme caution
  - Long segments without predicted secondary structure may indicate proteins disorder (NORsnet)
- Low complexity regions
- Signal sequences, transmembrane helix predictions
- Coiled coil

Practical