Intrinsic disorder proteins

Peter Tompa

Institute of Enzymology
Hungarian Academy of Sciences
Budapest, Hungary

Keith Dunker

CCBB & IUSM
Indiana University
Indianapolis, USA

Why do we study disorder proteins?

1) Gap in knowledge (1600 vs. thousands of IUPs)
2) Structural genomics initiatives
3) Bioinformatics studies
4) Single protein studies

Analysis of Signaling Interactions

- Examined each interaction on Pawson's website.
- Almost all of the interactions involved ordered regions binding to disordered partners.
- Conclusion: if Pawson's examples are typical, then a very significant proportion of protein-protein signaling interactions use disordered regions.

Parallel Paradigms

Catalysis

AA seq → **3-D Structure** → Function

Signaling

AA seq → **Disordered** → Function

Ensemble

Disorder and Function

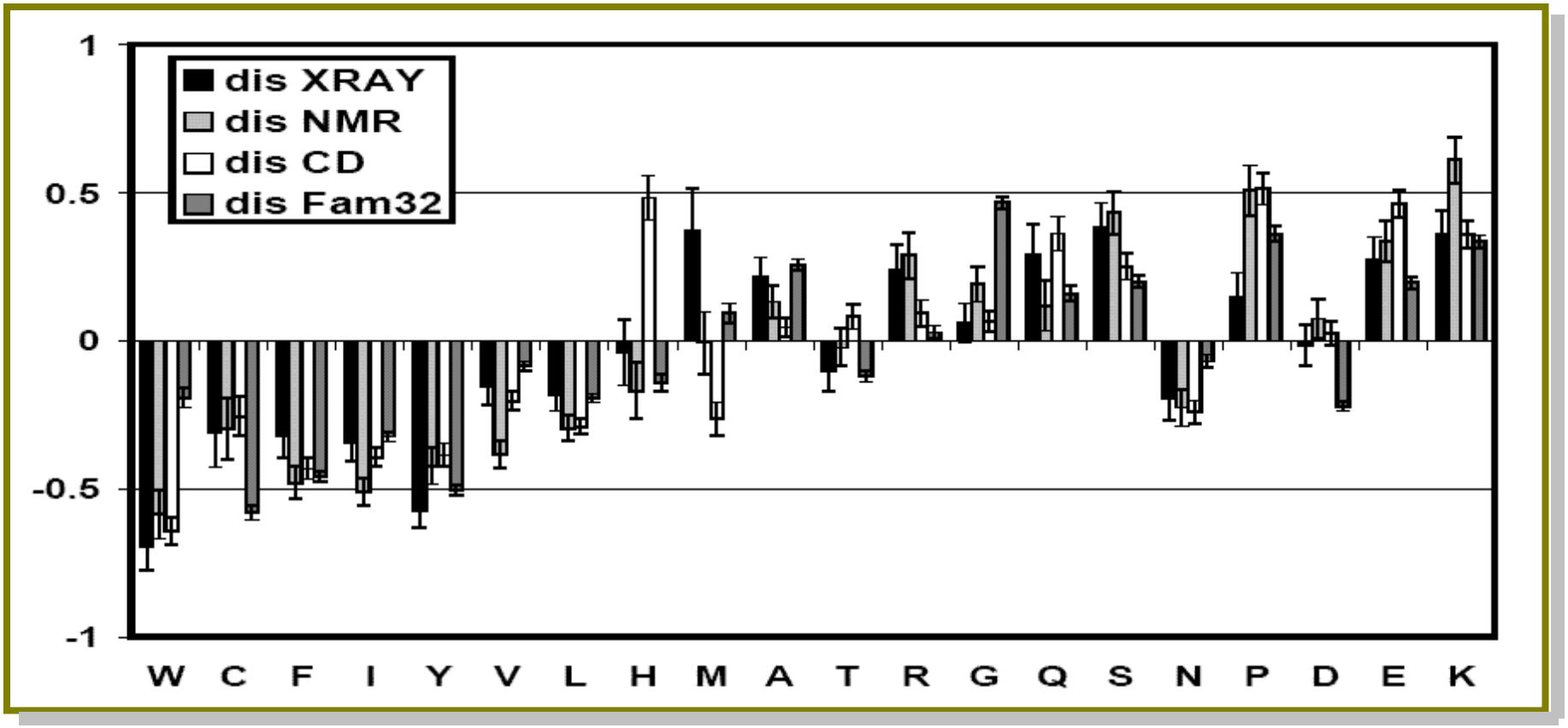
Category	Change	Examples	Descriptions
Molecular Recognition	**D → O**	**113**	Inter- and Intra-protein, ssDNA, dsDNA, tRNA, rRNA, mRNA, nRNA, bilayers, ligands, co-factors, metals
Protein Modification	**Variable**	**36**	Acetylation, fatty acylation, glycosylation, methylation, phosphorylation, ADP-ribosylation, ubiquitination, proteolytic digestion
Entropic Chains	**Variable**	**17**	Linkers, spacers, bristles, clocks, springs, detergents, self-transport

Dunker AK et al., Adv Protein Chem 62: 25-49 (2002)

Basic approaches to predict disorder

1) Machine learning
2) Structural approach

The unusual AA composition of IDPs



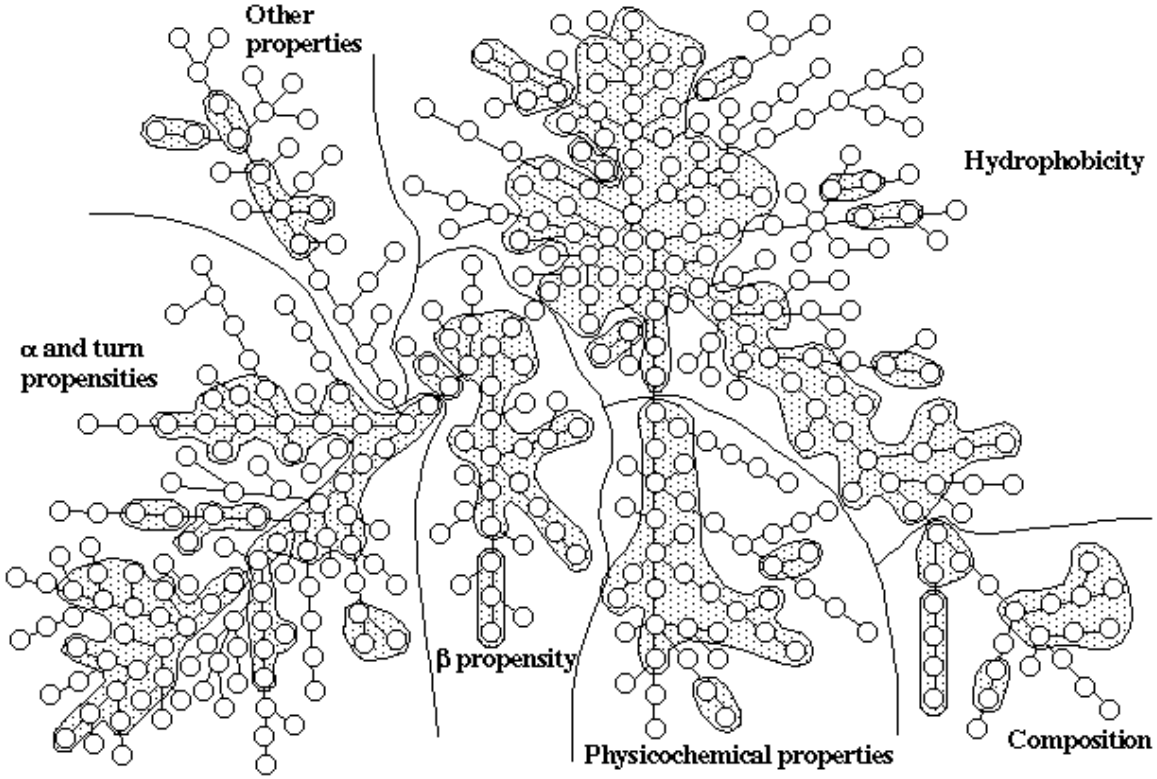
order-promoting

disorder-promoting

Dunker et al. (2001) *J. Mol. Graph. Model.* 19, 26

AA feature space: AAindex database
http://www.genome.jp/aaindex

A number is associated with every amino acid, which quantitatively describes how characteristic the given feature is to the AA (has 566 different scales at present)



Based on AA compositions, two things you might not want to do…

1) low complexity regions
2) regions w/o secondary structure

Low complexity regions

Sequence databases contain a lot of regions in which only a few amino acids occur (simple), or some dominate (biased), this can be described by an entropy function - Wootton (1993)

Shannon`s entropy

$$K_2 = -\sum_{i=1}^{N} \frac{n_i}{L}\left(\log_2 \frac{n_i}{L}\right)$$

L > 20; window size,
N alphabet size (20),
n_i: az i. aminosav száma

…may correspond to disorder

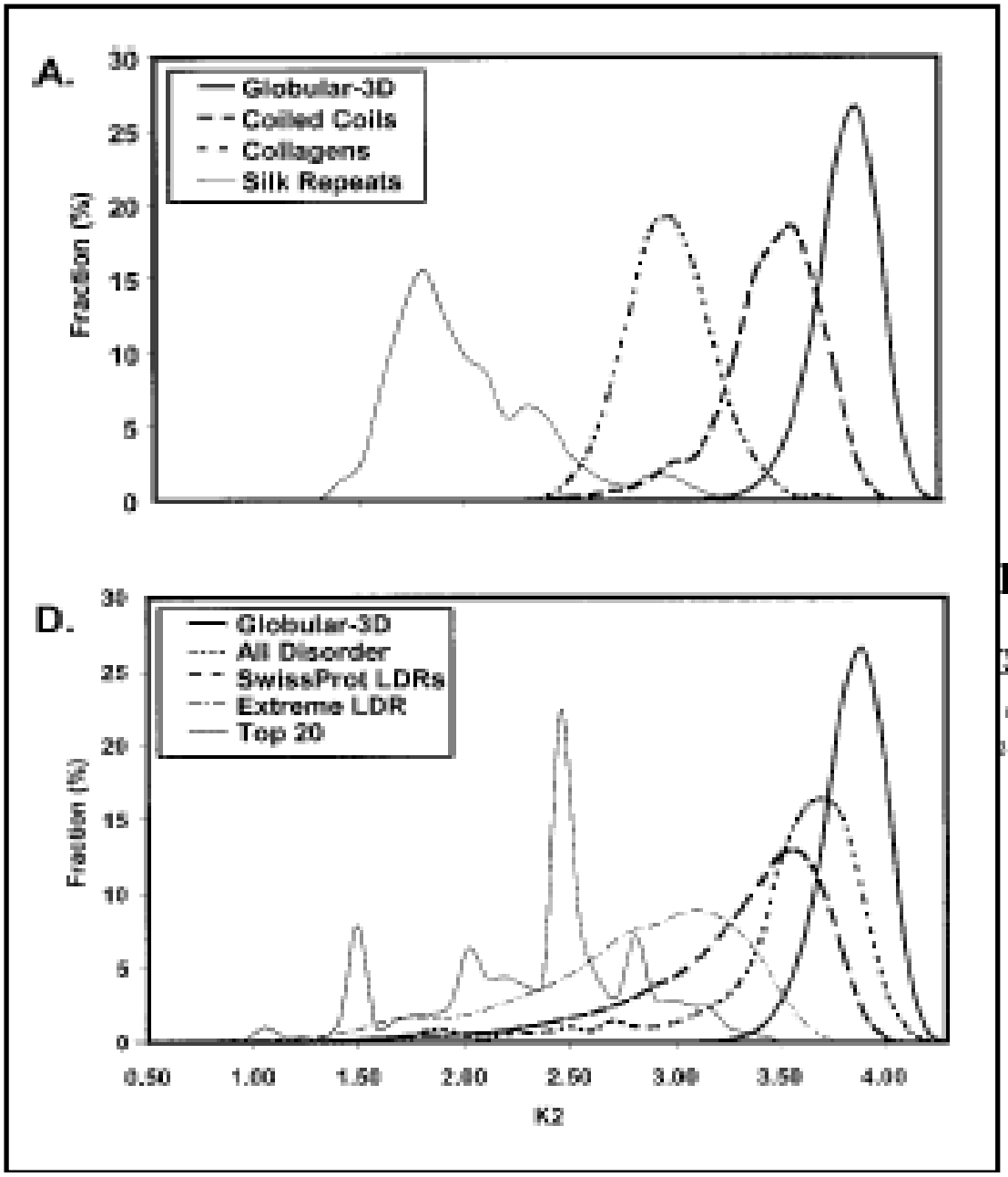
>SRY_MOUSE TRANSCRIPTIONAL ACTIVATOR

1-143 MEGHVKRPMNAFMVWSRGERHKLAQQNPSM
QNTEISKQLGCRWKSLTEAEKRPFFQEAQR
LKILHREKYPNYKYQPHRRAKVSQRSGILQ
PAVASTKLYNLLQWDRNPHAITYRQDWSRA
AHLYSKNQQSFYWQPVDIPTGHL

qqqqqqqqqqqfhnhhqqqqqfydhhqqqq 144-366
qqqqqqqqfhdhhqqkqqfhdhhqqqqqfh
dhhhhhqeqqfhdhhqqqqqfhdhqqqqqq
qqqqqfhdhhqqkqqfhdhhhhqqqqqfhd
hqqqqqqfhdhqqqqhqfhdhpqqkqqfhd
hpqqqqqfhdhhhqqqqkqqfhdhhqqkqq
fhdhhqqkqqfhdhhqqqqqfhdhhqqqqq
qqqqqqqqfhdqq

367-395
LTYLLTADITGEHTPYQEHLSTALWLAVS

The relationship of low complexity and disorder



PROTEINS: Structure, Function, and Genetics 42:38–48 (2001)

red Protein

Garner,[2*] Celeste J. Brown,[2] and A. Keith Dunker[2*]
State University, Pullman, Washington
ian, Washington

Based on AA compositions, two things you might not want to do…

1) low complexity regions
2) regions w/o secondary structure

…but there are globular proteins without Secondary structure and IDPs with secondary structure

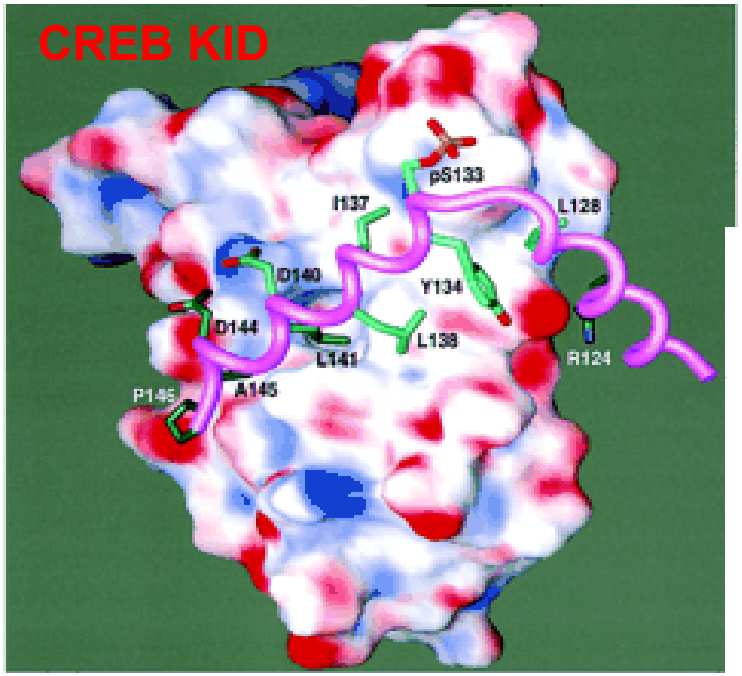
doi:10.1016/S0022-2836(02)00736-2 available online at http://www.idealibrary.com on IDEAL®

J. Mol. Biol. (2002) **322**, 53–64

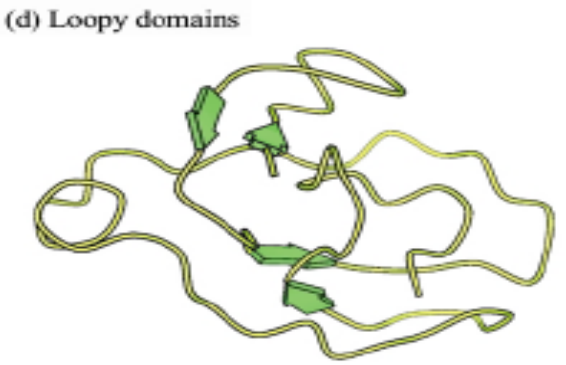
JMB

Loopy Proteins Appear Conserved in Evolution

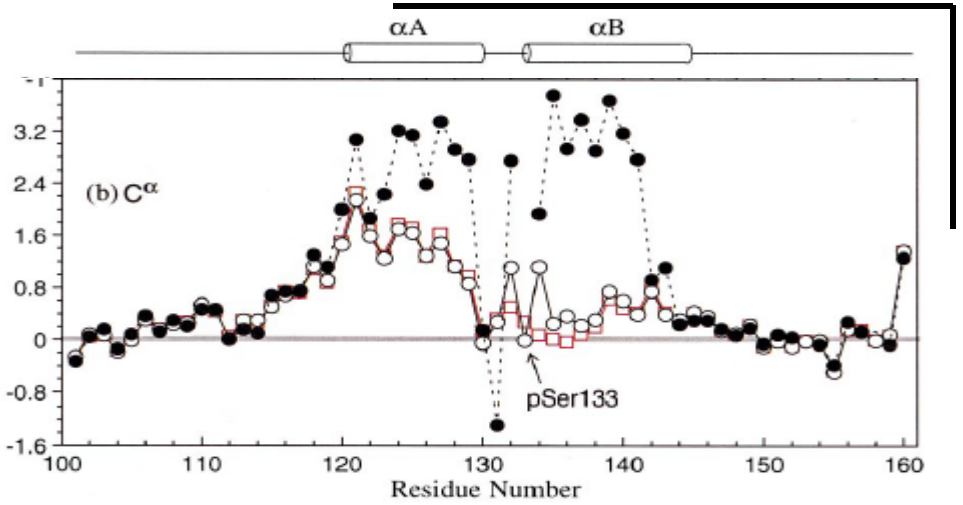
Jinfeng Liu[1,2,3], Hepan Tan[2] and Burkhard Rost[2,3,4]*



Radhakrishnan (1997) *Cell* 91, 741



1tbi

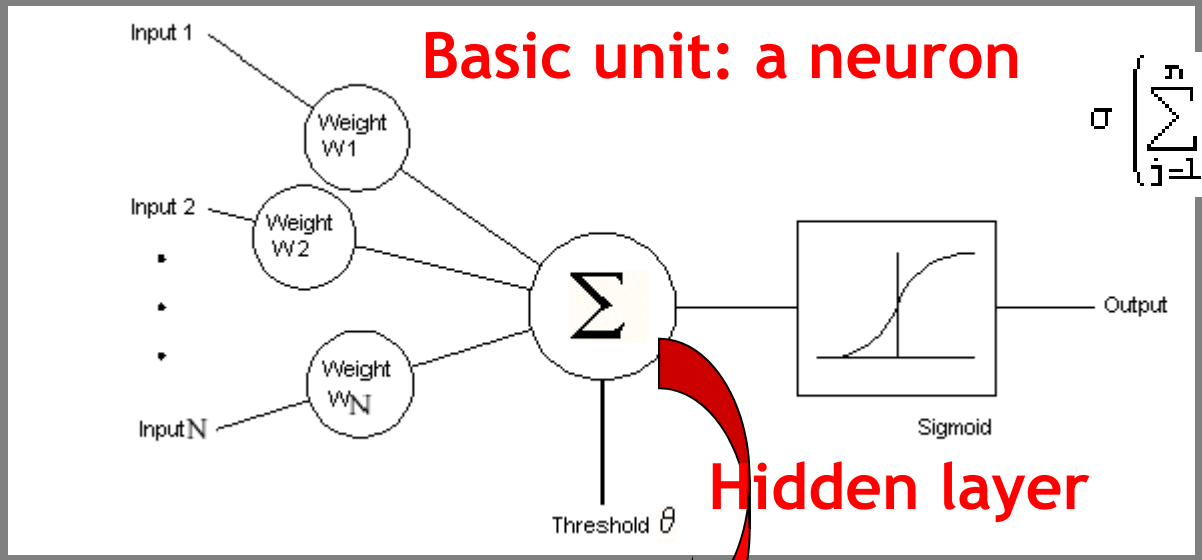


Radhakrishnan (1998) *FEBS Lett. 430, 317*

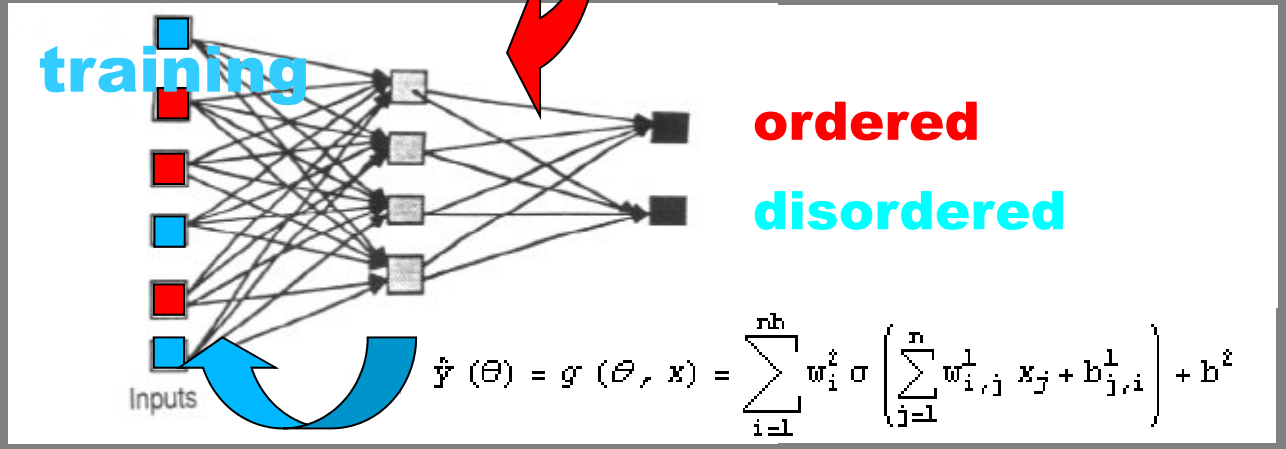
1) Machine learning

Artificial neural network (NN)

Artificial neural networks

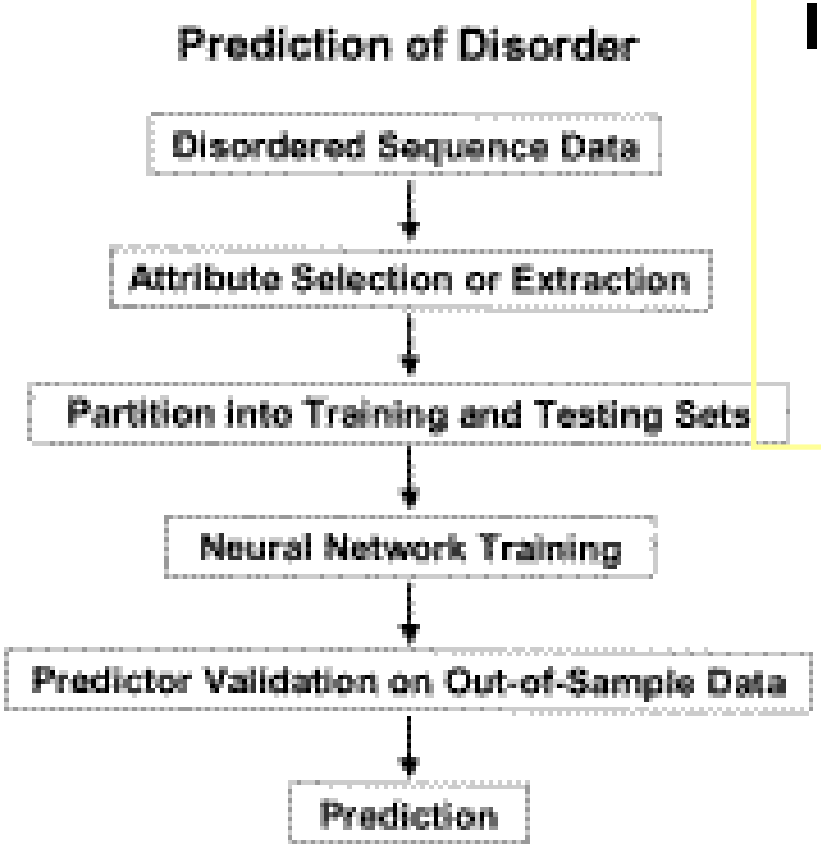


$$\sigma\left(\sum_{j=1}^{n} w_j x_j + b_j\right)$$



$$\hat{y}(\theta) = g(\theta, x) = \sum_{i=1}^{nh} w_i^2 \sigma\left(\sum_{j=1}^{n} w_{i,j}^1 x_j + b_{j,i}^1\right) + b^2$$

Predictor of naturally disordered regions (PONDR®)



Input

18 amino acids

Hydrophobicity

Sequence complexity

VL2, VL3

VLXT

VSL2

Dunker, 1998

Predicting Disordered Regions from Amino Acid Sequence: Common Themes Despite Differing Structural Characterization

Ethan Garner [1] Paul Cannon [2] Pedro Romero [2]

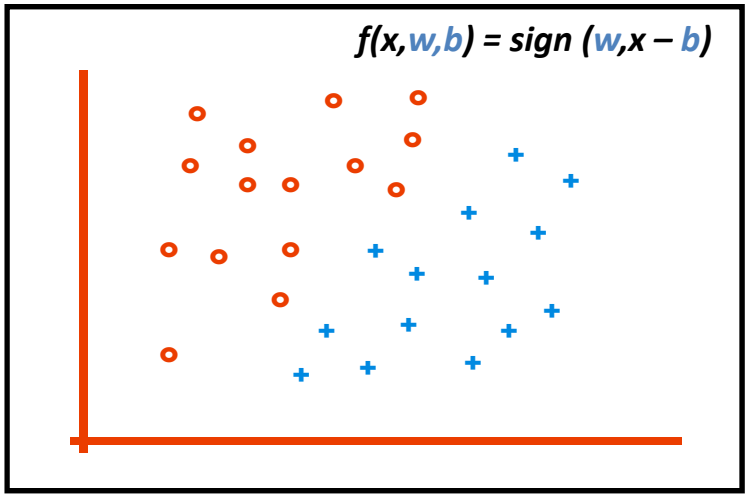
THOUSANDS OF PROTEINS LIKELY TO HAVE LONG DISORDERED REGIONS

PEDRO ROMERO, ZORAN OBRADOVIC
School of Electrical Engineering and Computer Science
Washington State University, Pullman, WA 99164

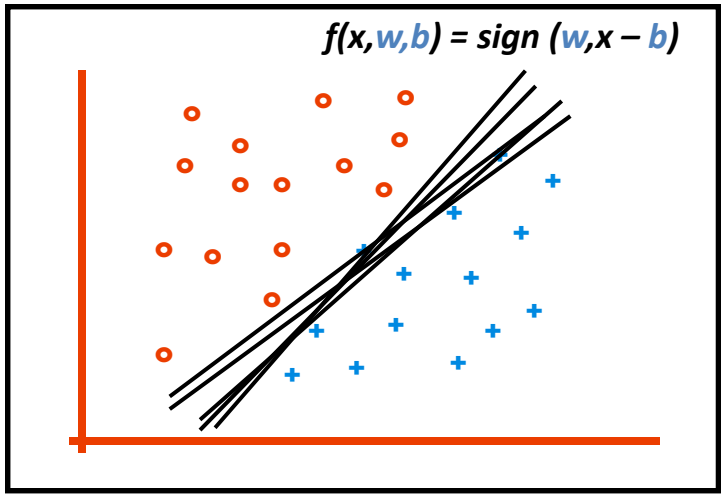
PROTEIN DISORDER AND THE EVOLUTION OF MOLECULAR RECOGNITION: THEORY, PREDICTIONS AND OBSERVATIONS

A. K. DUNKER, E. GARNER, S. GUILLIOT
dunker@mail.wsu.edu
Department of Biochemistry & Biophysics,
Washington State University, Pullman, WA 99164-4660

Support vector machine (SVM)

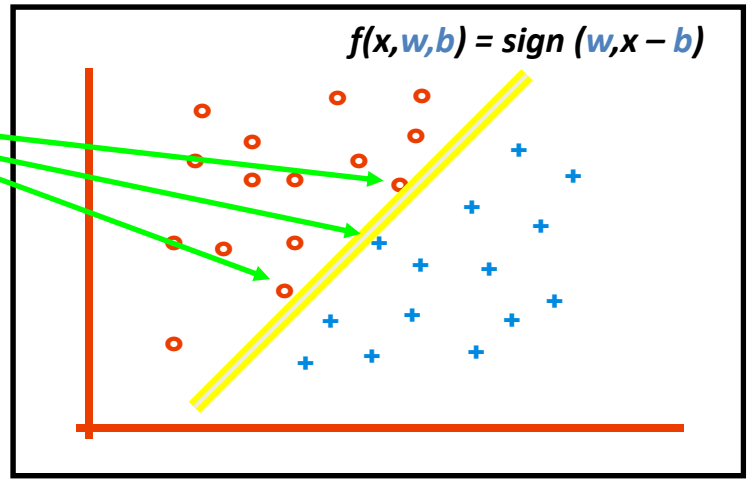


How would you classify this data?



But which is best?

Support vectors against which the margin pushes up



This is the simplest SVM, the linear SVM (LSVM)

The maximum margin linear classifier is considered best

2) Structural approach (interaction potential)

The protein non-folding problem

- **Protein folding problem**

  How does amino acid sequence determine protein structure ?

- **Protein non-folding problem**

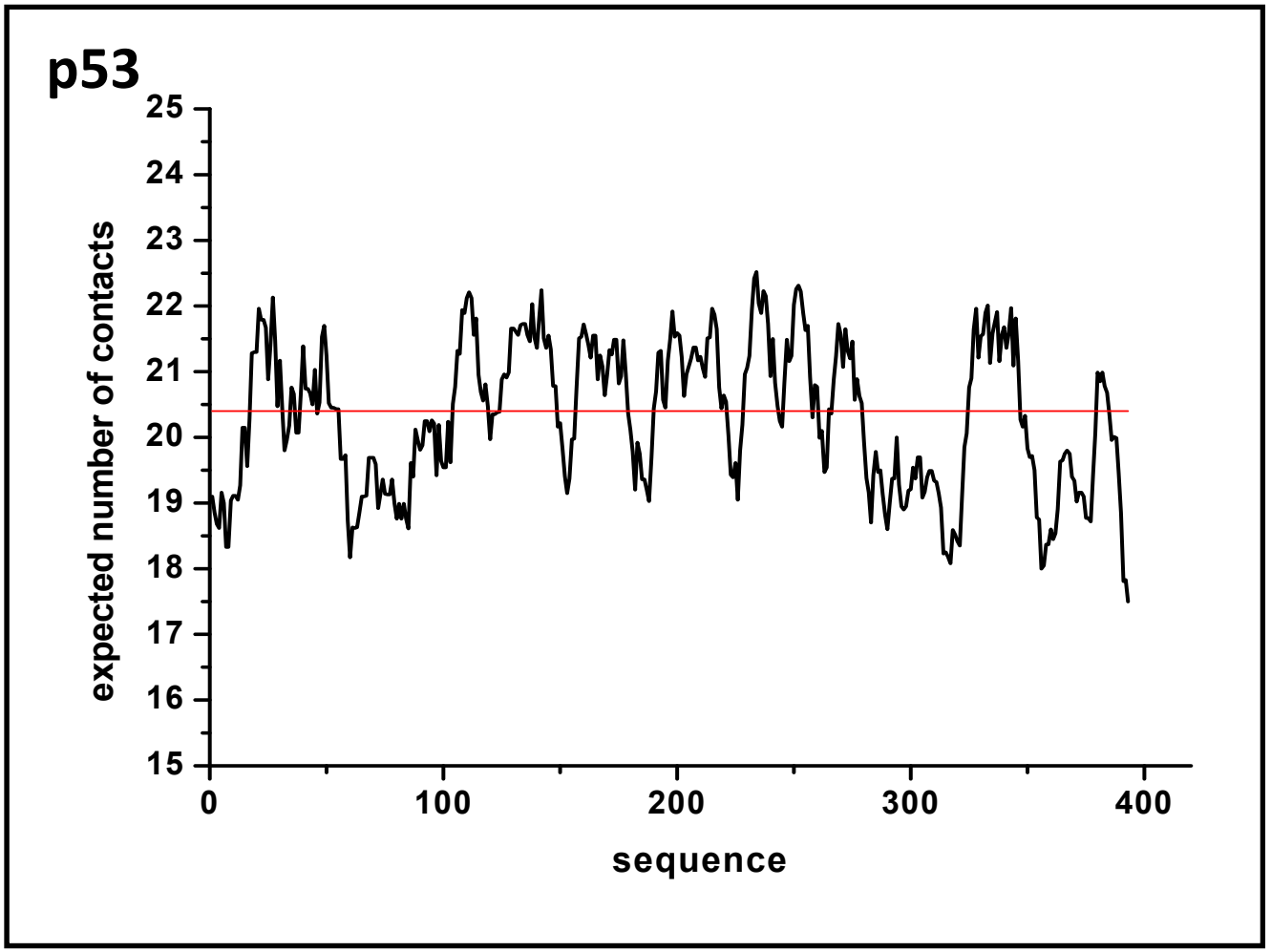
  How does amino acid sequence determine the lack of protein structure ?

The protein non-folding porblem

- Globular proteins have special sequences that enable the formation of a large number of favorable interactions
- IDPs contain (disorder-promoting) amino acids, which tend to avoid interacting with each other.
- An IDP thus cannot fold into a low-energy conformation.

A simple implementation, FoldUnfold

Calculates the contact number of amino acids



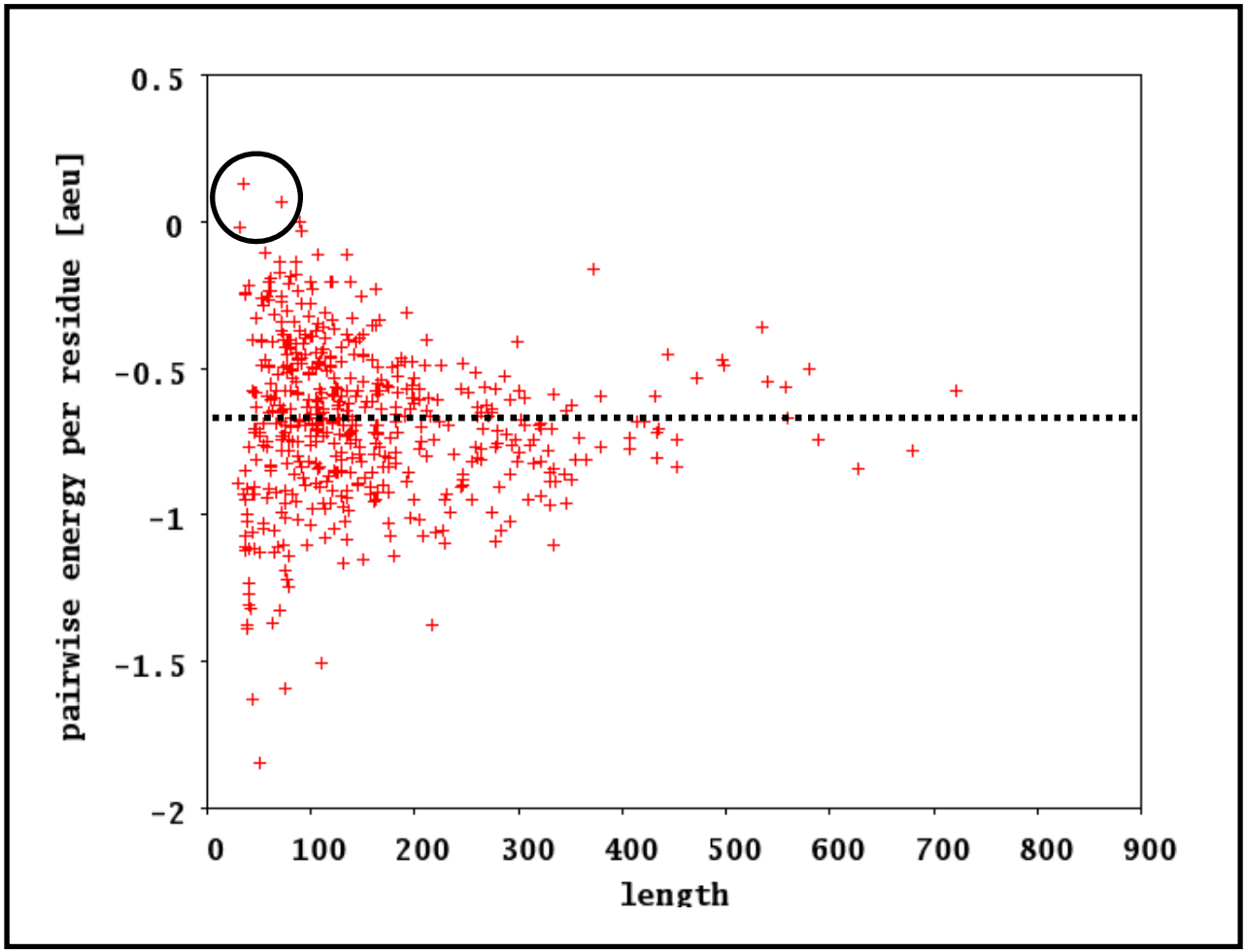
Estimating the total pairwise interresidue interaction energy of a sequence: IUPred

1) Calculate interresidue interaction energies from structure

2) Try to estimate the energy without knowing the structure

3) Apply the estimation to sequences w/o structure (e.g. to IDPs, which have no structure)

Interresidue interaction energy calculated for known structures



Dosztanyi (2005) J. Mol. Biol. 347, 827

How to estimate the interresidue interaction energy of a protein of unknown structure or w/o structure?

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
```

Calculated energy per residue

Sequence

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

?

Estimated energy per residue

Estimation of pairwise interresidue interaction energy from sequence

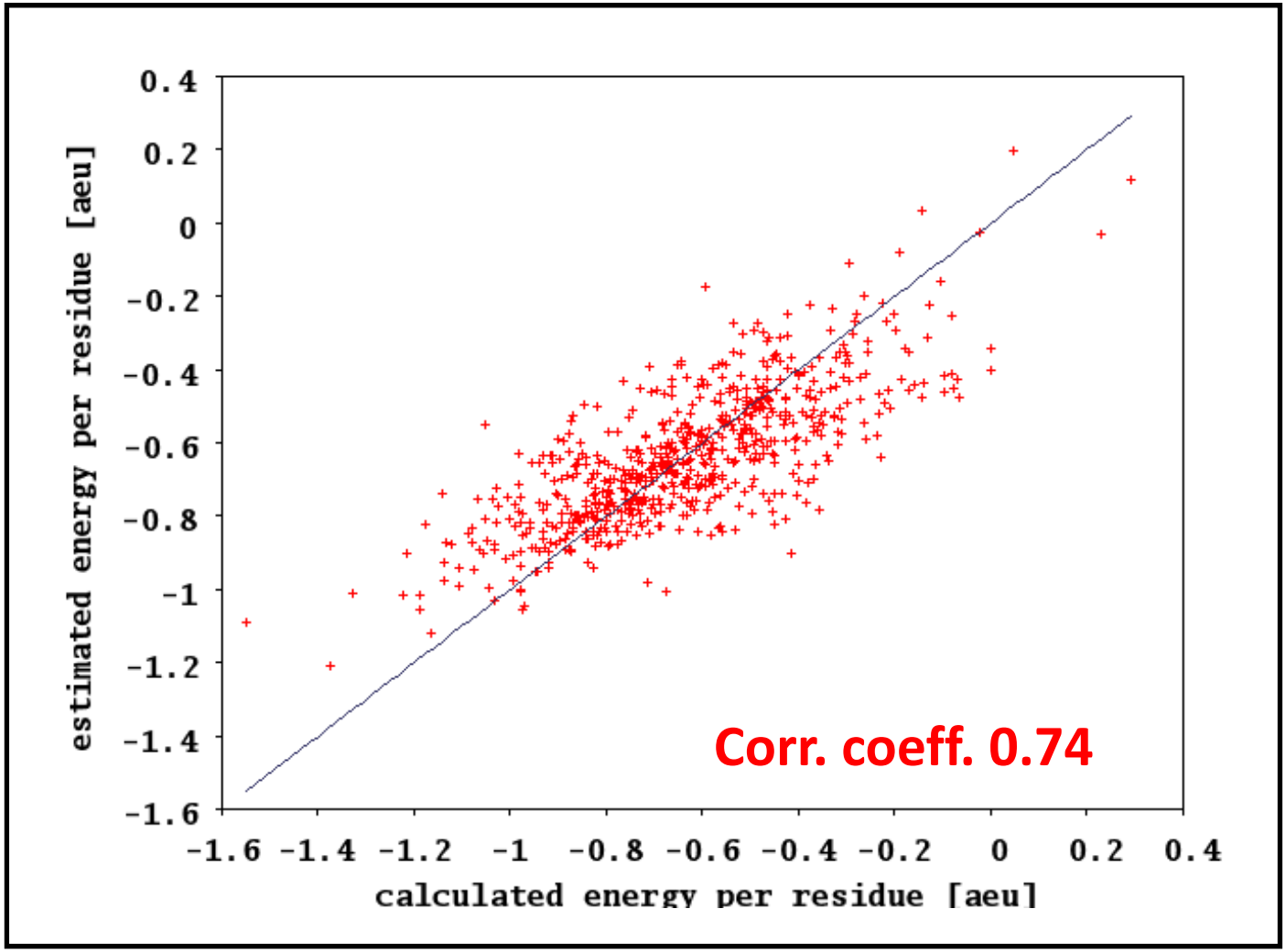
The contribution of individual AAs depends on its potential partners, i.e. its neighborhood. A quadratic formula is needed to take this into consideration.

$$E(estimated)/L = \begin{pmatrix} n_A & n_C & \cdots & n_Y \end{pmatrix} \begin{pmatrix} P_{AA} & P_{AC} & \cdots & P_{AY} \\ P_{CA} & P_{CC} & & \\ \vdots & & \ddots & \\ P_{YA} & \cdots & \cdots & P_{YY} \end{pmatrix} \begin{pmatrix} n_A \\ n_C \\ \vdots \\ n_Y \end{pmatrix}$$

The relationship between AA composition and energy is given by an optimized 20x20 energy predictor matrix, P_{ij}

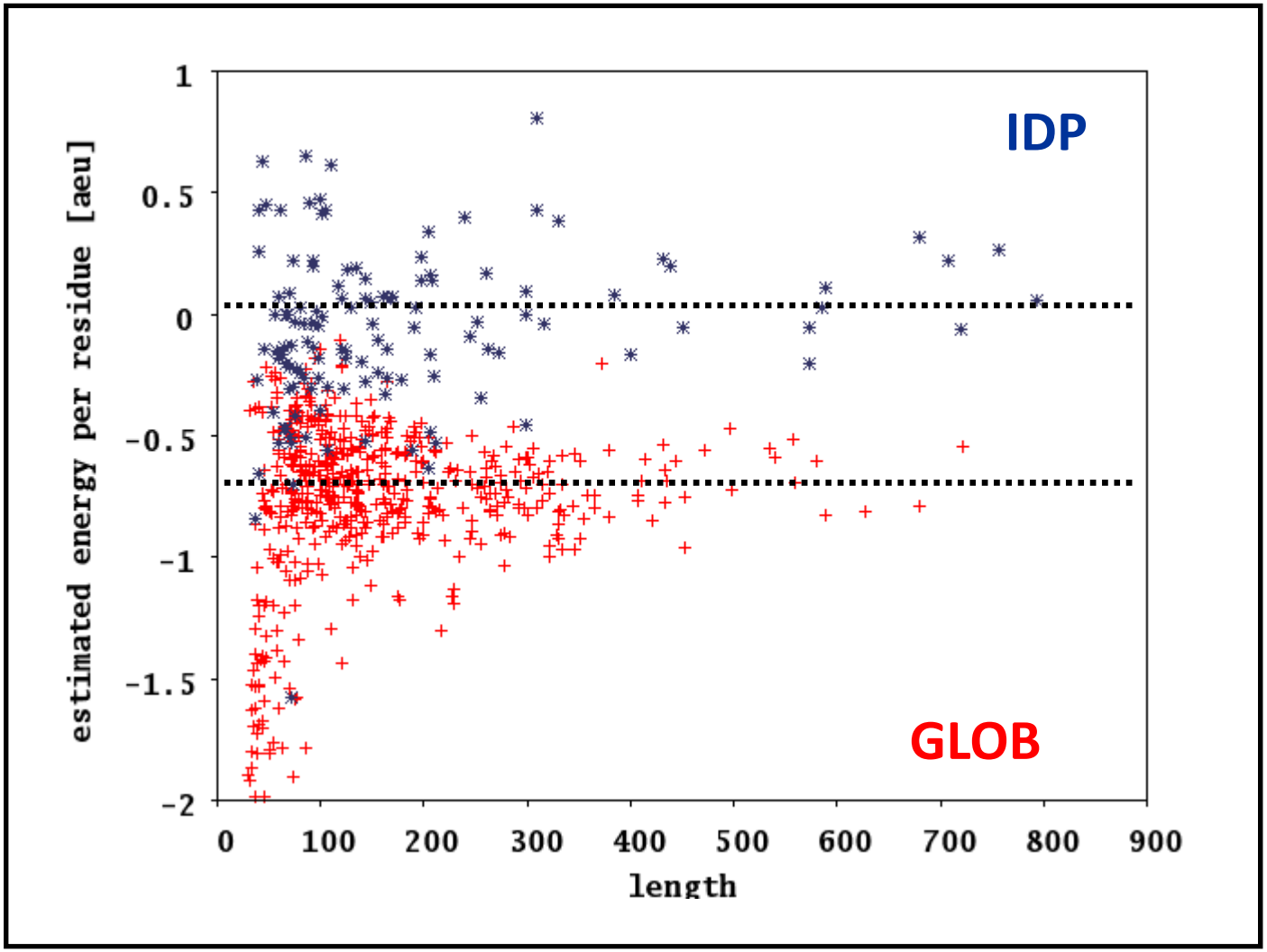
$$P_{ij}: \quad \sum_{k}^{\text{Globular proteins}} \left(E_k(calc) - E_k(est)\right)^2 \rightarrow \min$$

Correlation of calculated and estimated interaction energies



Dosztanyi (2005) *J. Mol. Biol.* 347, 827

The estimated energy for globular proteins and IDPs

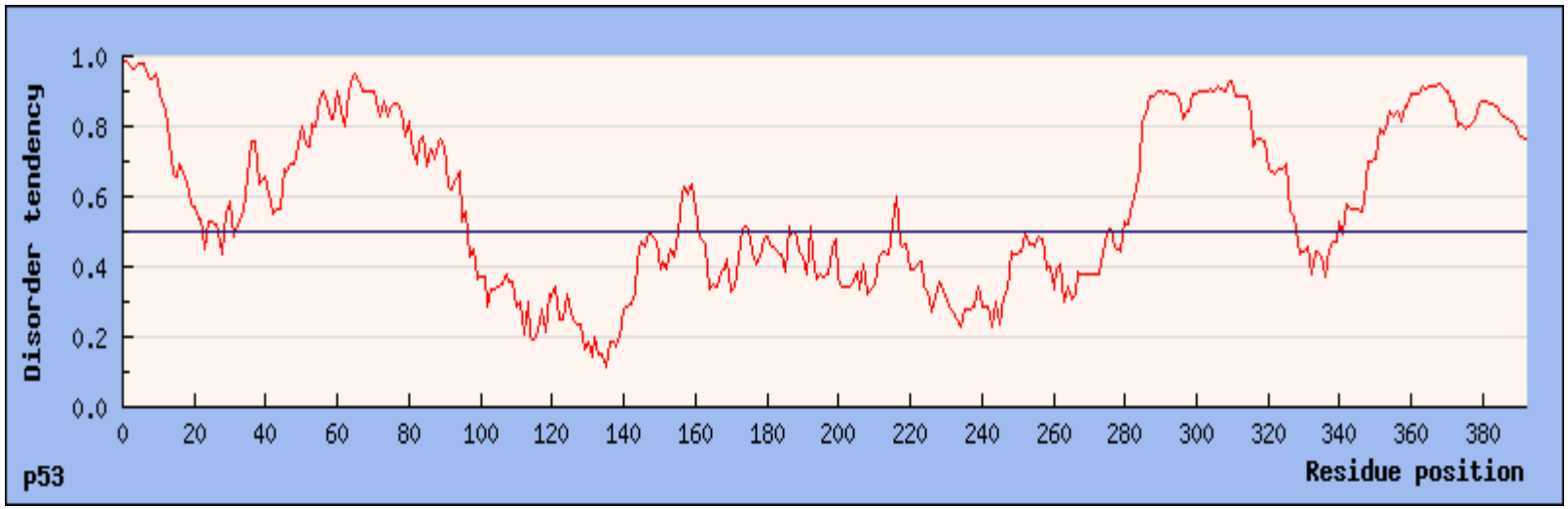


Dosztanyi (2005) *J. Mol. Biol.* 347, 827

Making it position-specific: the IUPred algorithm

IUPred: http://iupred.enzim.hu

IUPred, p53


Disorder tendency
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100
120
140
160
180
200
220
240
260
280
300
320
340
360
380
p53
Residue position

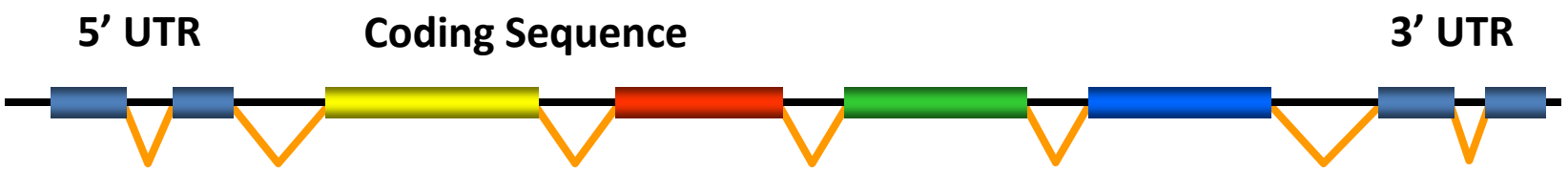
www.disprot.org

DisEMBL™	Intrinsic Protein Disorder Prediction
DISOPRED2	Disorder Prediction Server
DRIPPRED	Web based predictor for disordered regions in proteins
FoldIndex©	Estimate the fold probability of a protein
GlobPlot 2	Intrinsic Protein Disorder, Domain & Globularity Prediction
IUPred	Prediction of Intrinsically Unstructured Proteins
PONDR®	Predictors of Natural Disordered Regions
PreLink	Prediction of unfolded segments in a protein sequence based on amino acid composition
RONN	Regional Order Neural Network
VL2	DisProt Predictor of Intrinsically Disordered Regions
VL3, VL3H, VL3E	DisProt Predictor of Intrinsically Disordered Regions

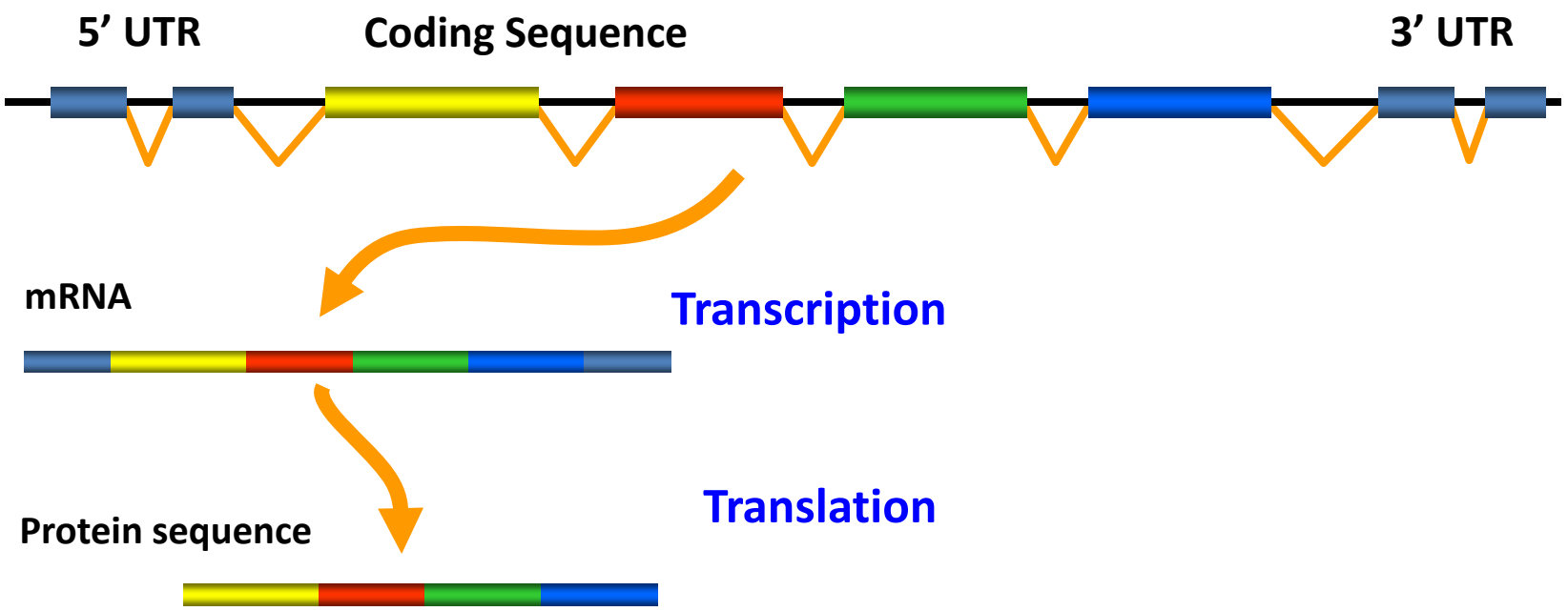
Alternative Splicing and Intrinsic Disorder

- Find proteins with both ordered and disordered regions.
- Find mRNA alternative splicing information for these proteins and map to the ordered and disordered regions.
- For alternatively spliced regions of mRNA, do they code for ordered protein more often or do they code for disordered protein more often?

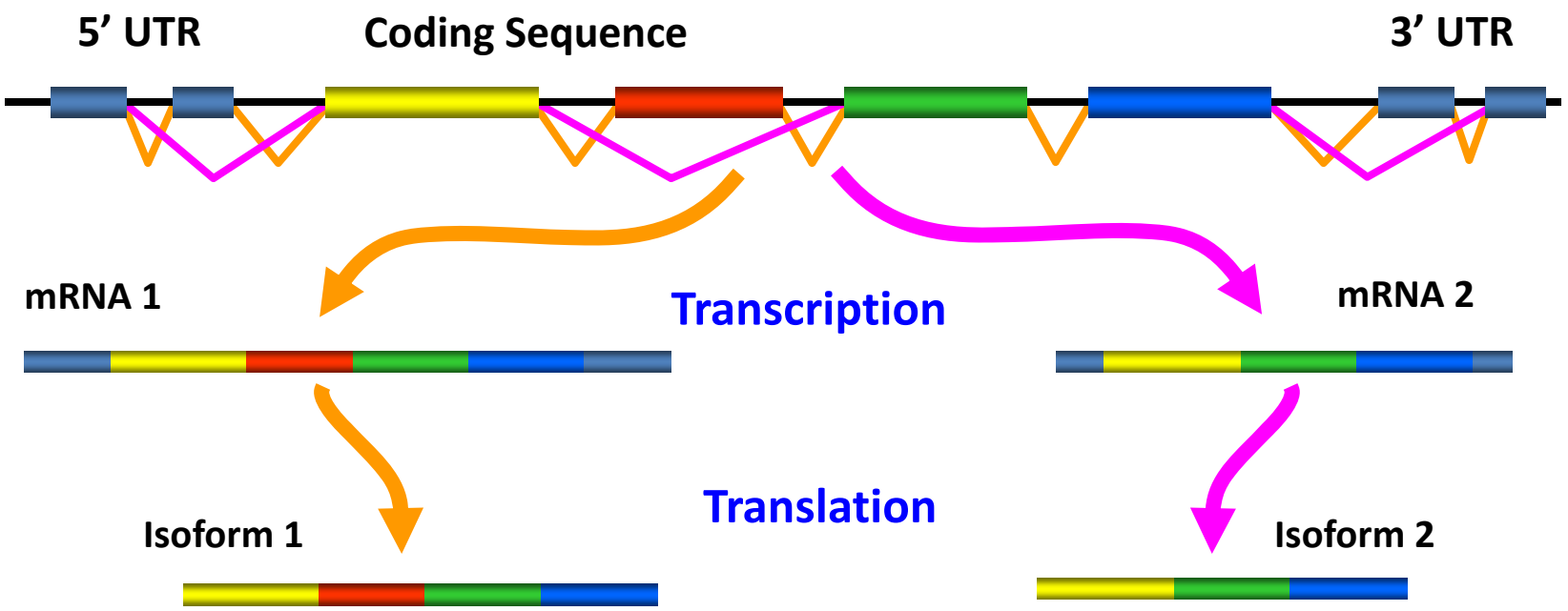
Alternative Splicing



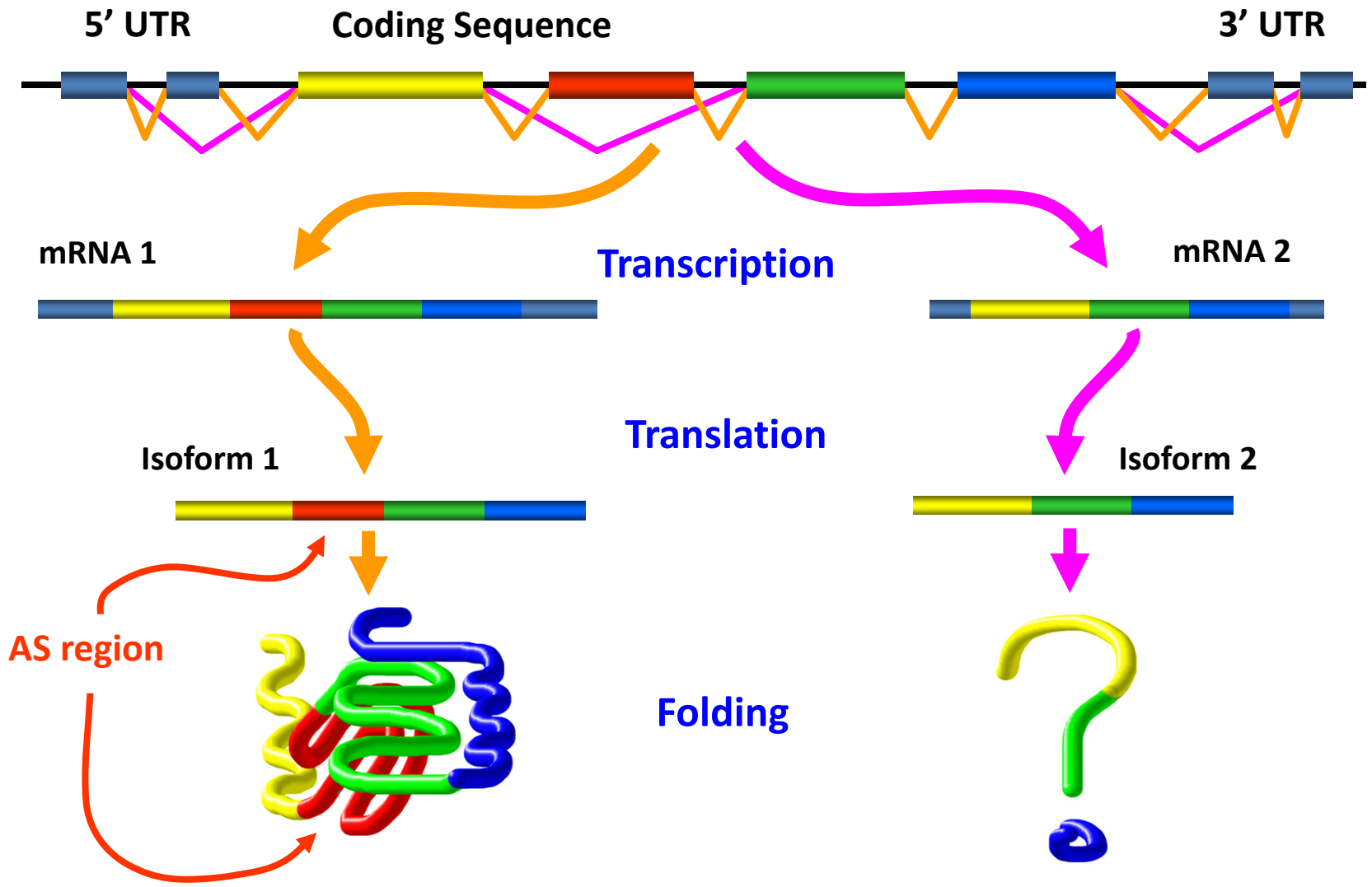
Alternative Splicing



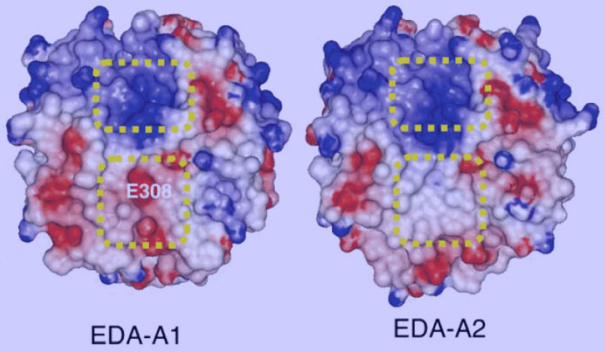
Alternative Splicing



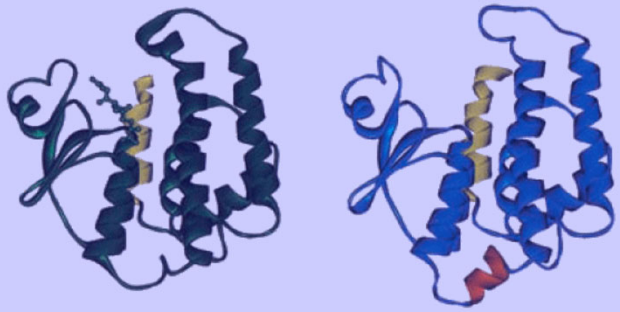
Alternative Splicing



Structural Studies of AS

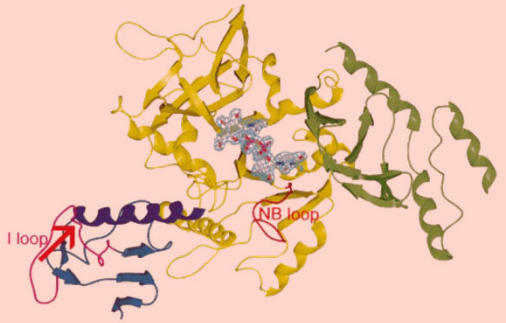


Structured AS regions

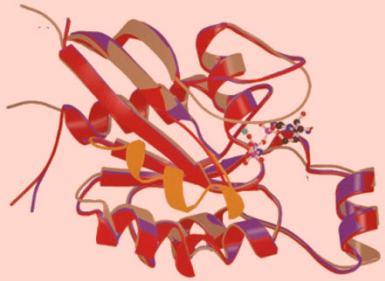


Glutathione S-transferase

Disordered AS regions



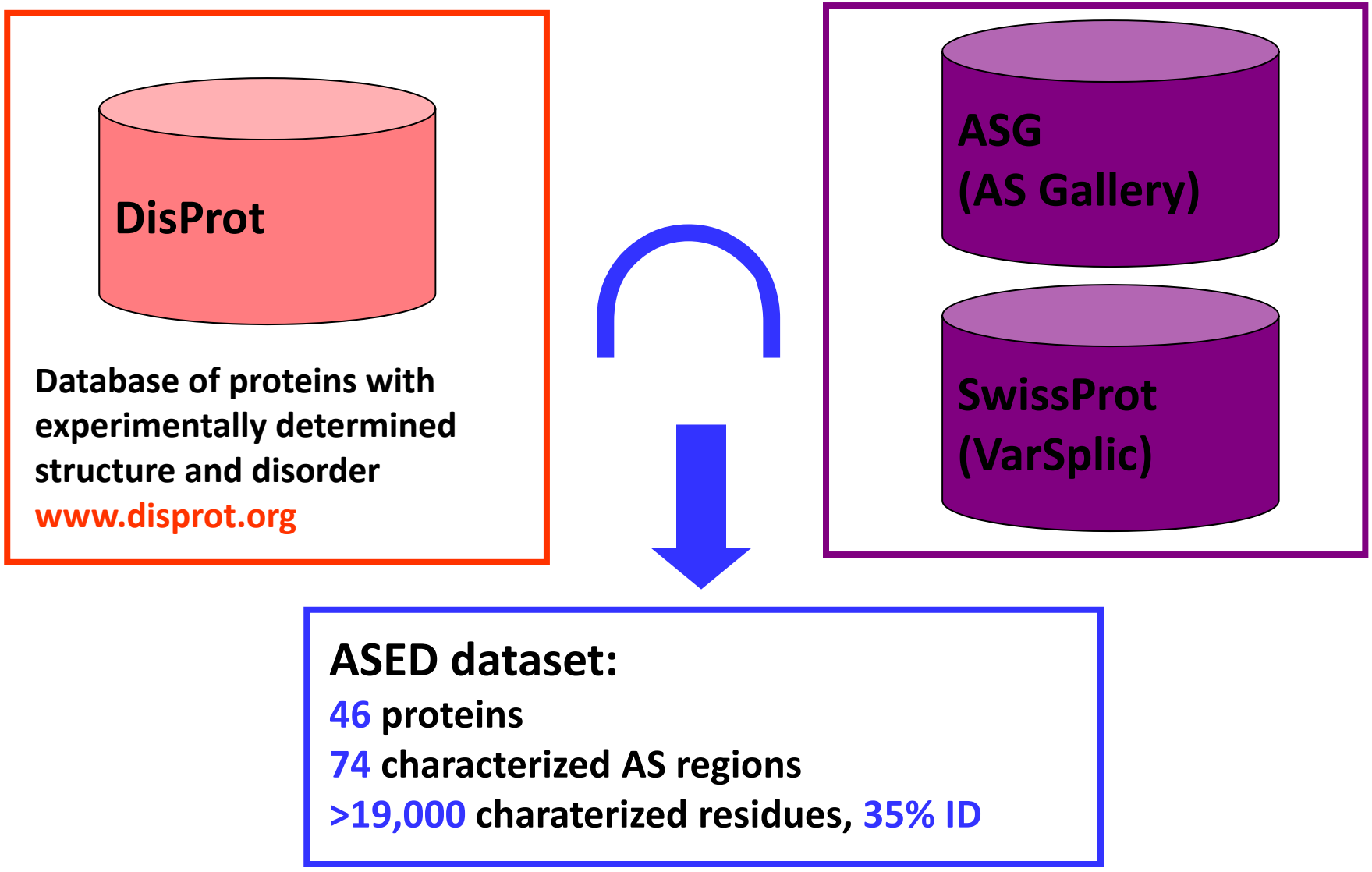
Pyrophosphorylase



RAC1
Tumor necrosis factor

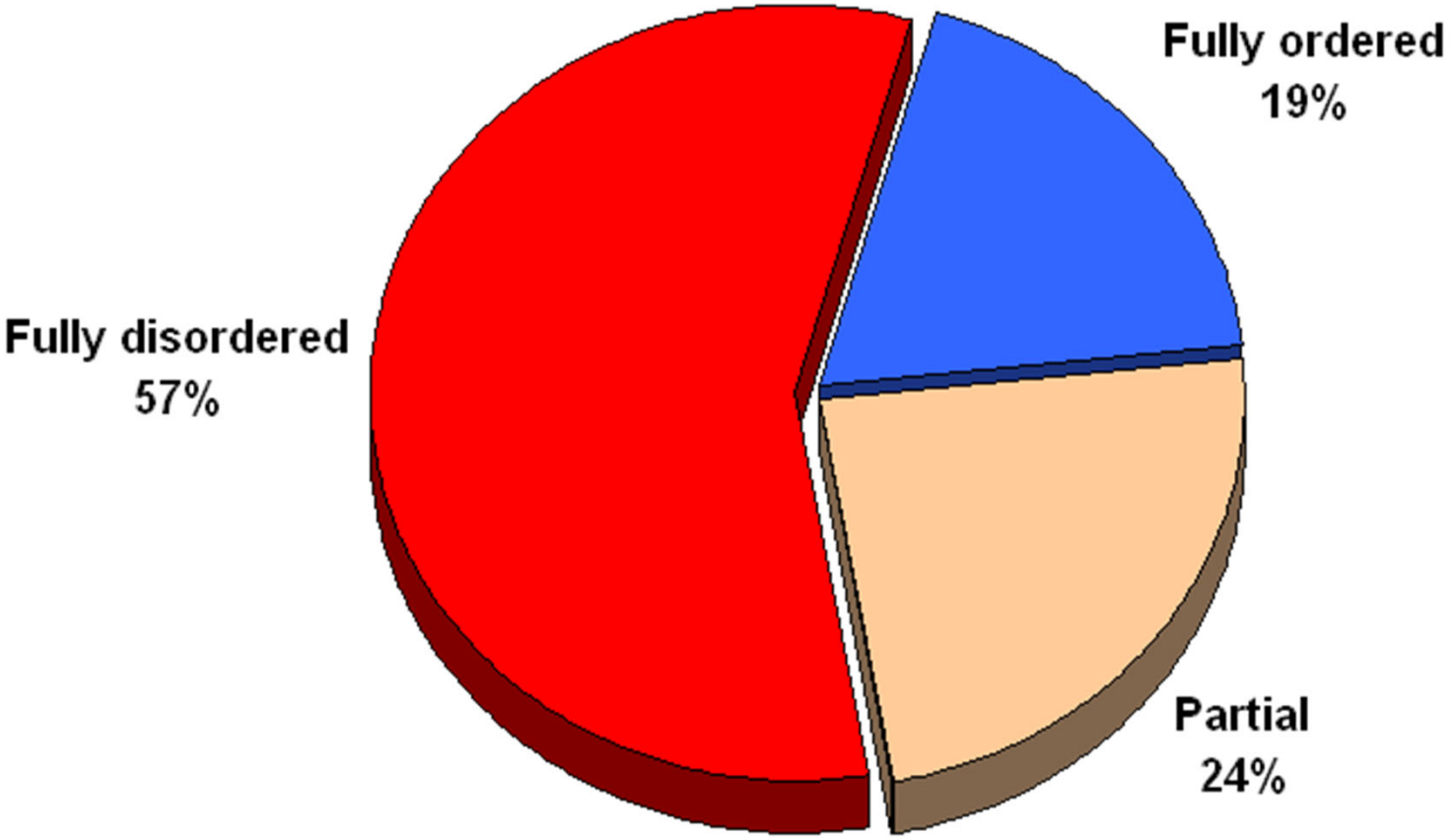
Sulphotransferase

Studying the Relationship ID↔AS

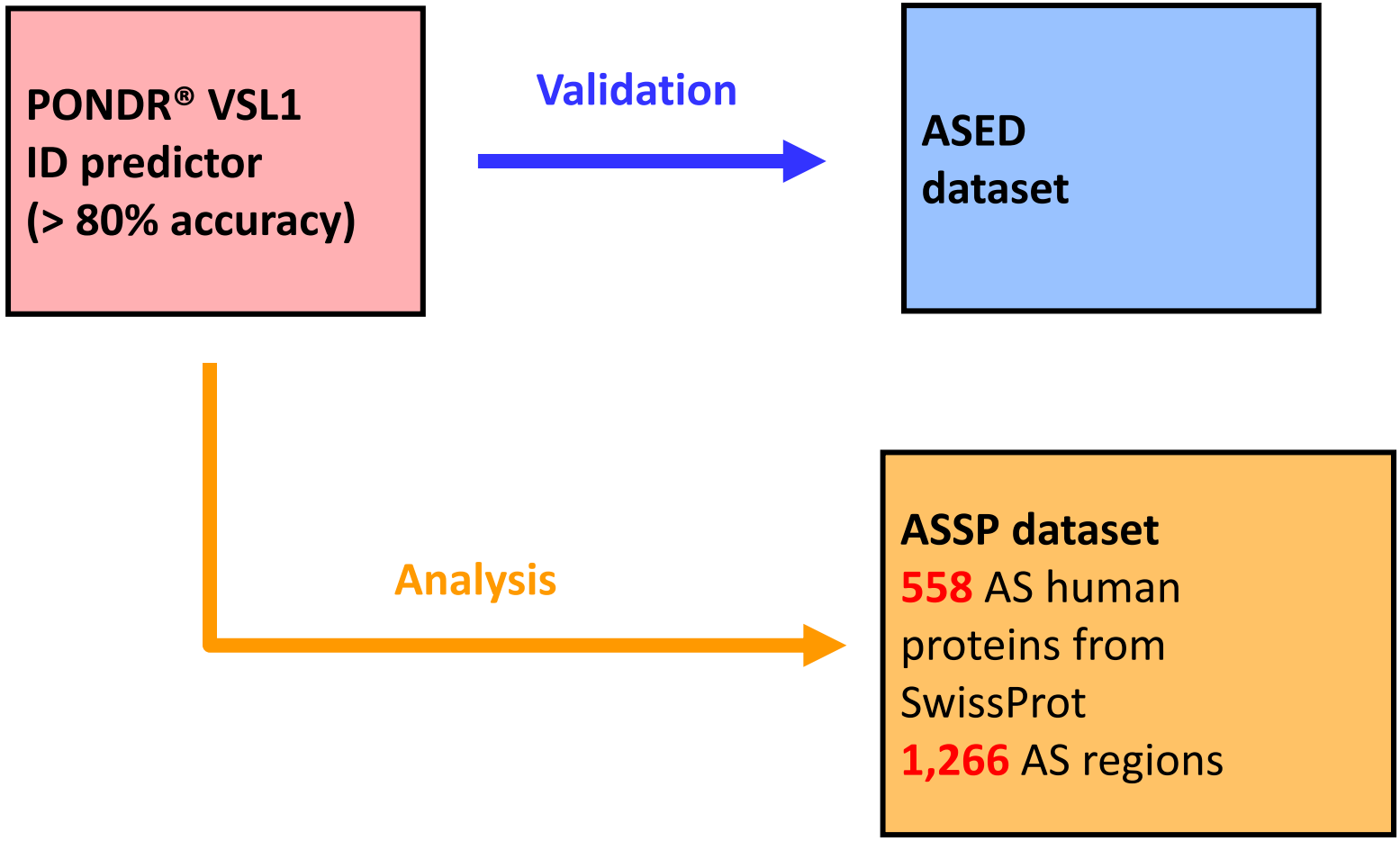


Results on ASED

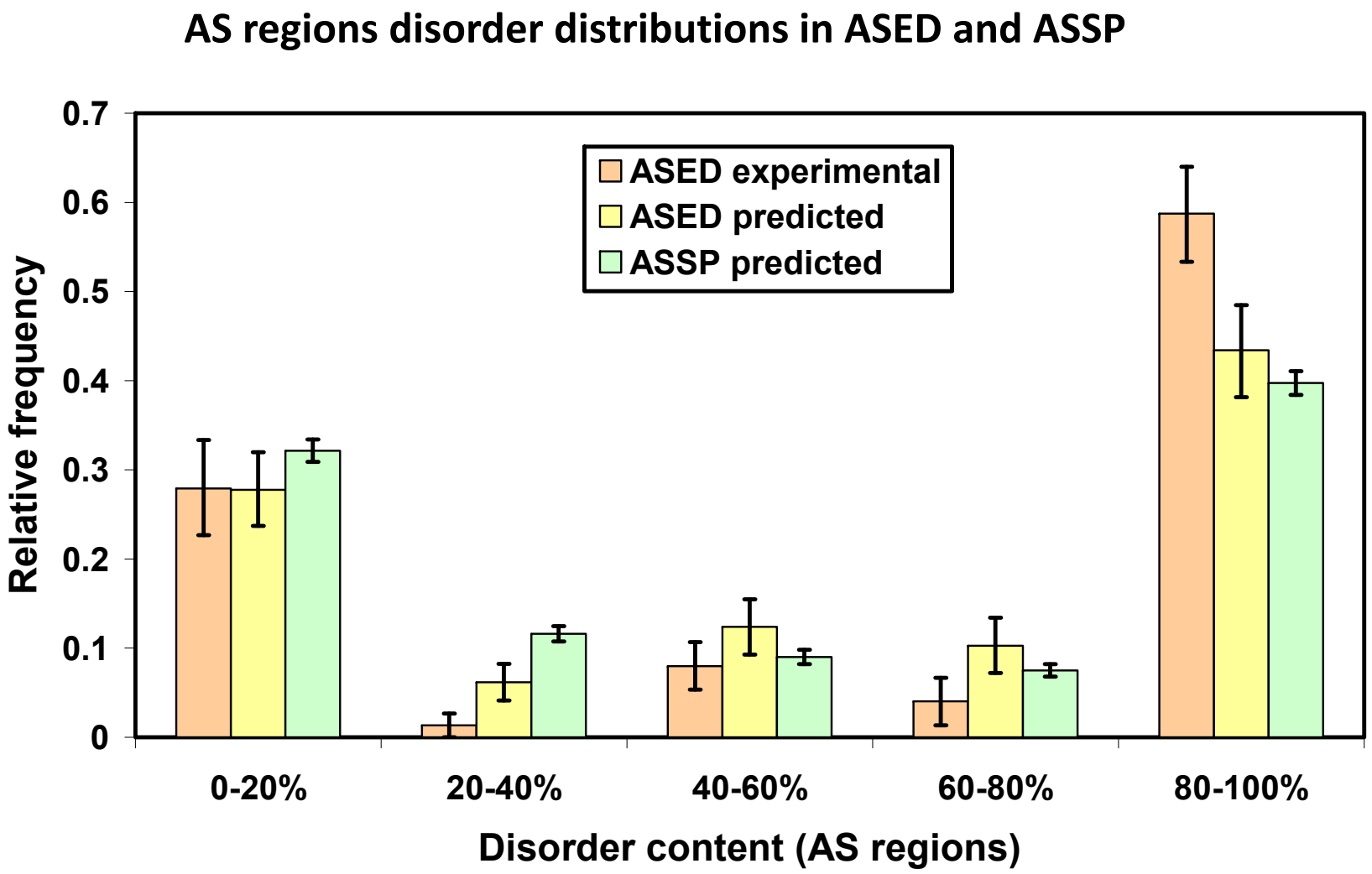
Distribution of structurally characterized AS regions



Enlarging the Dataset



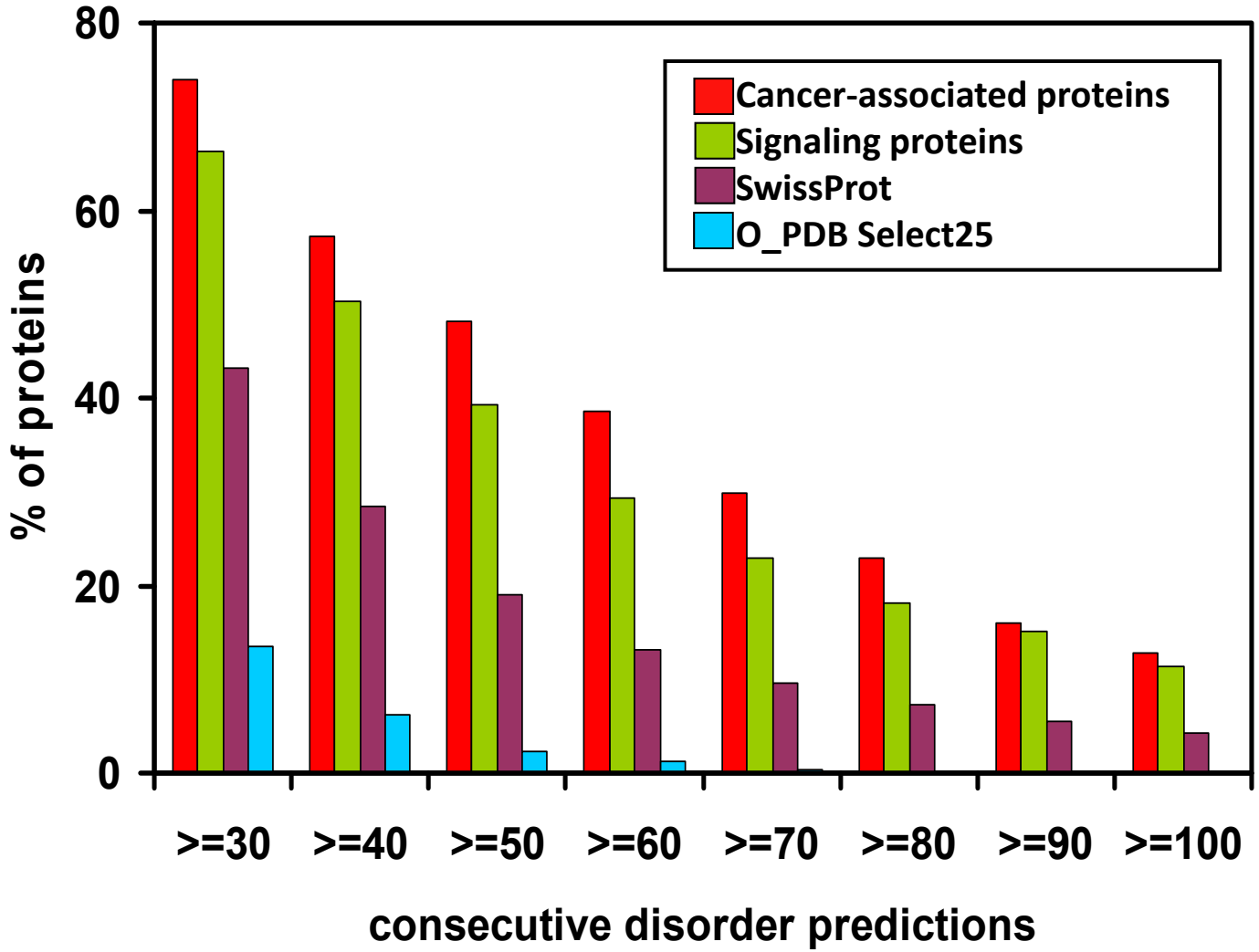
Global Results



Alternative Splicing and Disorder

- Ordered Proteins: active site residues non-local in sequence, become associated by protein folding
- Disordered Proteins and regions: functional residues localized in sequence
- Functional regions for signaling and regulation are located one after another
- Alternative splicing edits functional sets and thereby leads to regulatory and signaling diversity

Disorder and Cell Signaling



Iakoucheva I et al., J Mol Biol 323:573-584 (2002)

Disorder and Drug Discovery

- The p53-MDM2 interaction is blocked by several drugs; one is in clinical trials and shows promise as an anti-cancer drug.
- The drug molecules bind to the ordered partner, preventing the disordered partner from binding.
- Such interactions are typically weak per unit of surface area, and the interaction surfaces can be small, thus such interactions are ideal drug targets.
- Molecular Kinetics has strategy to find all druggable MoRF-based interactions; bioinformatics indicates that more than one hundred are in cancer-associated proteins.
- Is this approach a new drug discovery pathway?