Structural Bioinformatics

Elodie Laine

Master BIM-BMC Semestre 3, 2014-2015

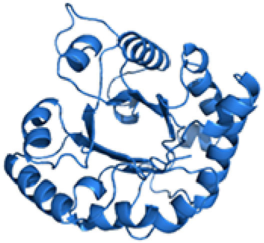
Laboratoire de Biologie Computationnelle et Quantitative (LCQB)
e-documents: http://www.lcqb.upmc.fr/laine/STRUCT
e-mail: elodie.laine@upmc.fr

Lecture 6 – Intrinsically Disordered Proteins

Elodie Laine – 14.10.2014

Structured & disordered protein building blocks

Structured domain

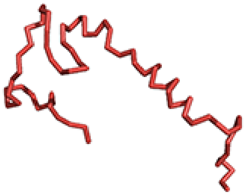


TIM barrel
fully folded
(stable folded monomer)

sequence
↓
structure
↓
function
(e.g. enzyme catalysis)

Disordered region

sequence
↓
disorder
↓
function
(e.g. binding)

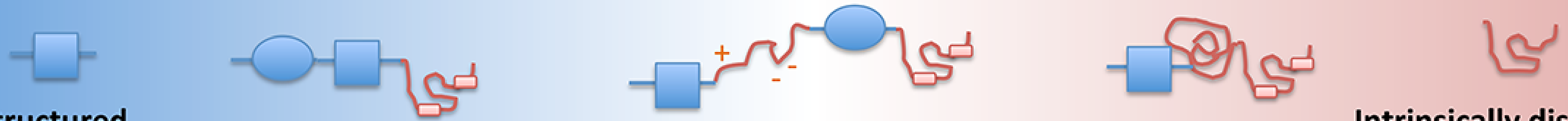


p27
Conformational ensemble
(disordered monomer)

structure-function paradigm
(established)

disorder-function paradigm
(emerging)

Proteome



Structured protein

Proteins with structured domains and disordered regions

Intrinsically disordered protein (IDP)

Disorder becomes apparent

- **Denaturation arises from loss of structure** (Hsien 1931): if particular conditions, such as acid, urea, or high temperature, cause a protein to lose its unique and ordered structure, then it loses its ability to carry out function and is considered to have become denatured

- **Conformational selection** :
  - high flexibility enables antibodies to bind different antigens (Pauling 1940)
  - bovine serum albumin binding sites are heterogeneous (Karuth 1950)

- **Unstructured proteins were observed** in intact cells in early proton NMR experiments (Daniels 1978)

- **The rapid rise of genomic data** (~1990) has led to the extensive use of sequence analysis for the identification of intrinsically unstructured sequences.

A recent discovery

- **Classic biochemical methods** are strongly biased towards the production and characterization of folded, active proteins. They discover a detectable activity and isolate it by purifying the protein.

- **The use of genetic methods** to isolate function, using mutants and knockouts, has been the way that most unfolded proteins have been identified so far.

- This **identification process** involves:

  - Formulating a function (for example, control of transcription)
  - Mapping the function to a particular gene and to a particular area
  - Transcribing the gene, producing the protein and purifying the protein
  - Examining its structure by circular dichroism and NMR spectroscopy
  - If the protein is unfolded, trying to co-express it with a binding partner (also identified through genetic mapping)

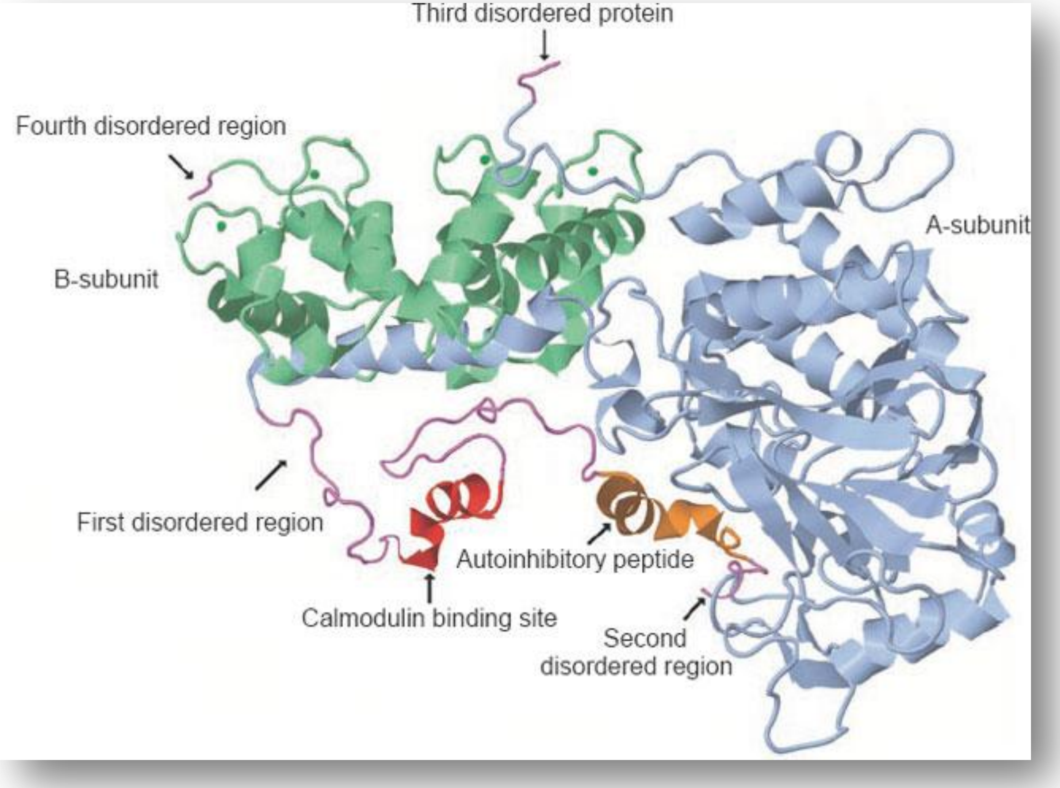
Definition

A protein is **disordered** if it does not adopt a well-defined structure when isolated in solution under near-physiological conditions (Eliezer 2009)

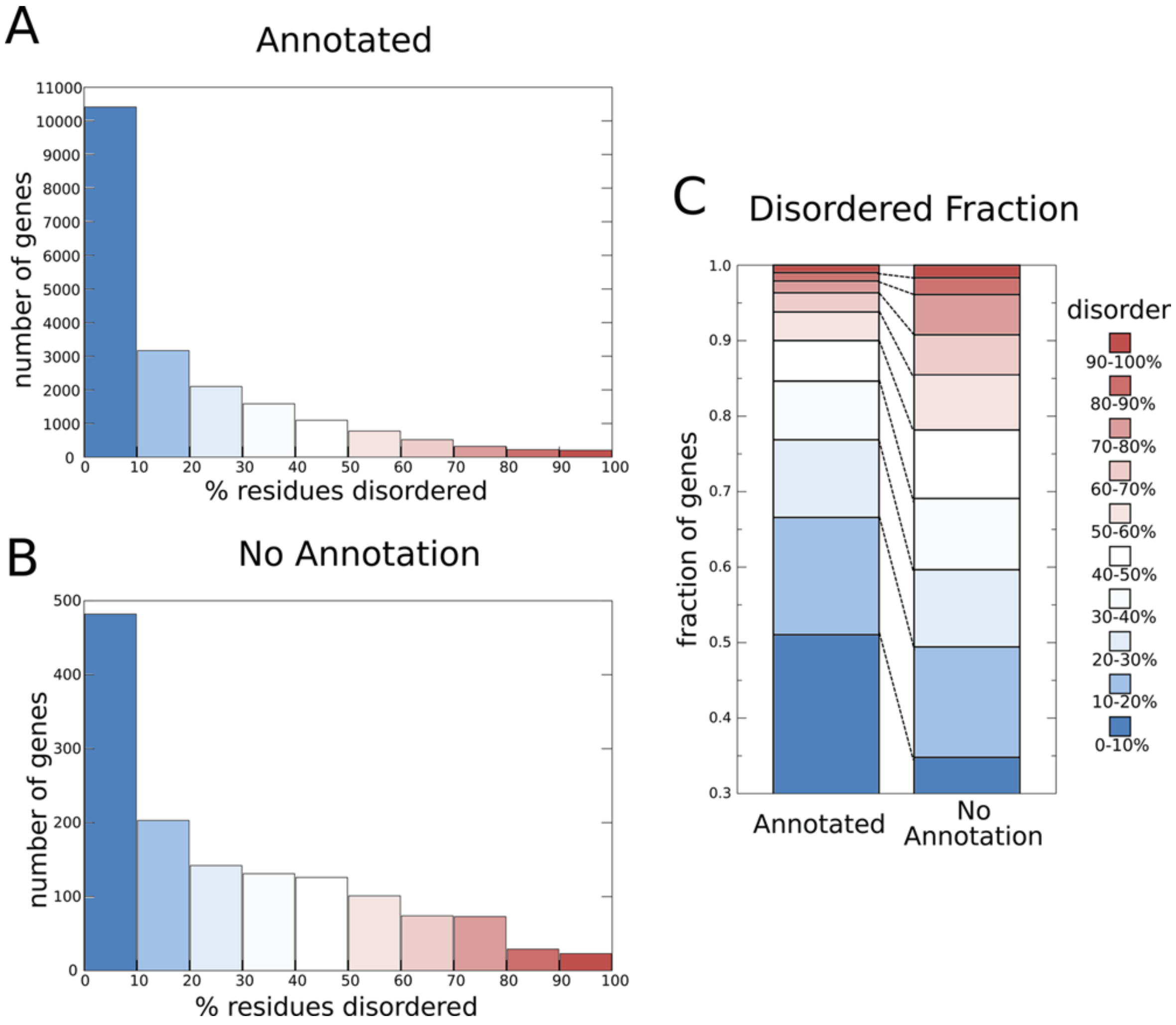
2 types :

- **Denatured state ensembles** (DSEs)
- **Intrinsically disordered proteins** (IDPs)

Structure of **calcineurin** with essential disorder



Disorder in human genes

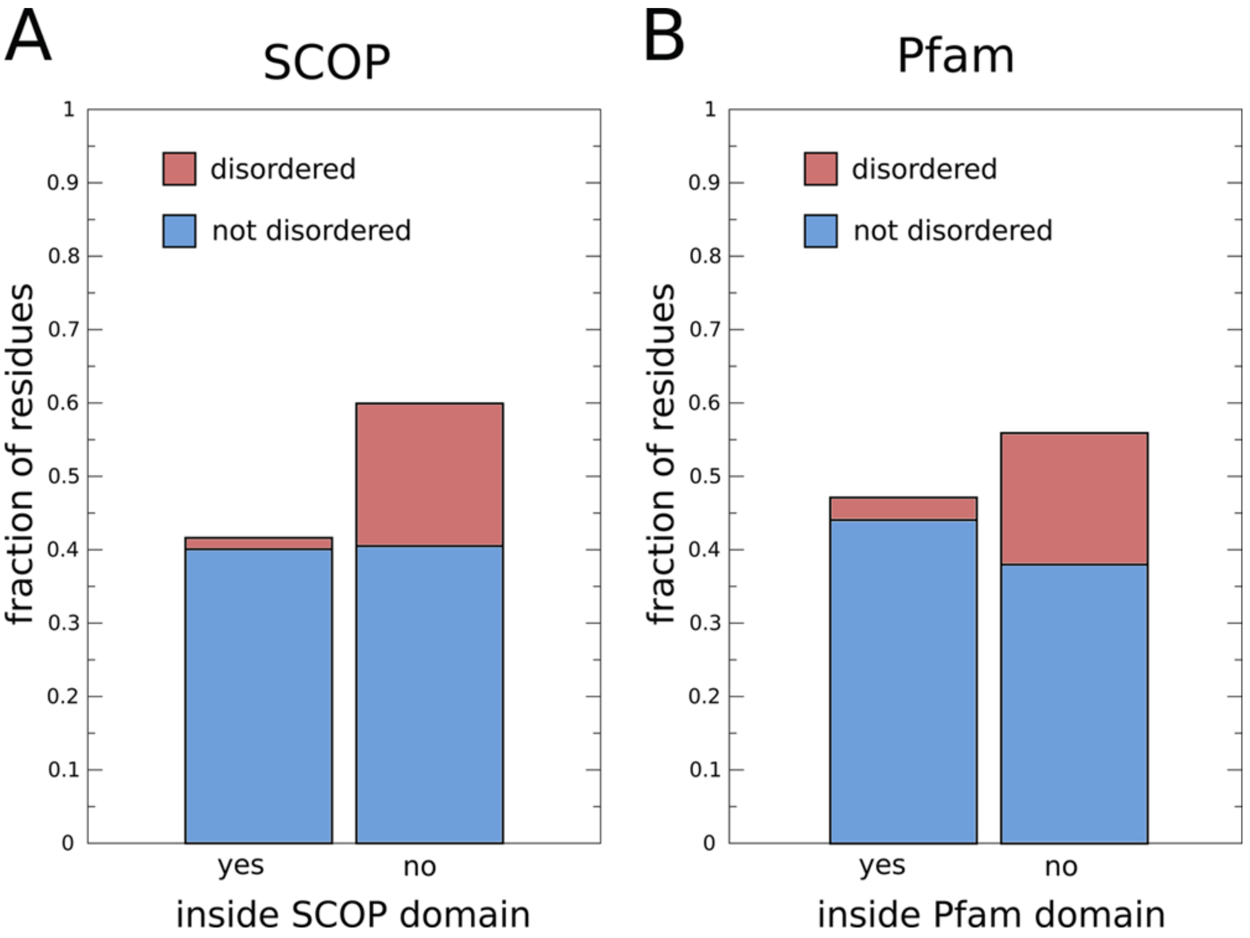


44% of human protein-coding genes contain disordered segments of >30 amino acids in length. In the human genome, 6.4% of all protein-coding

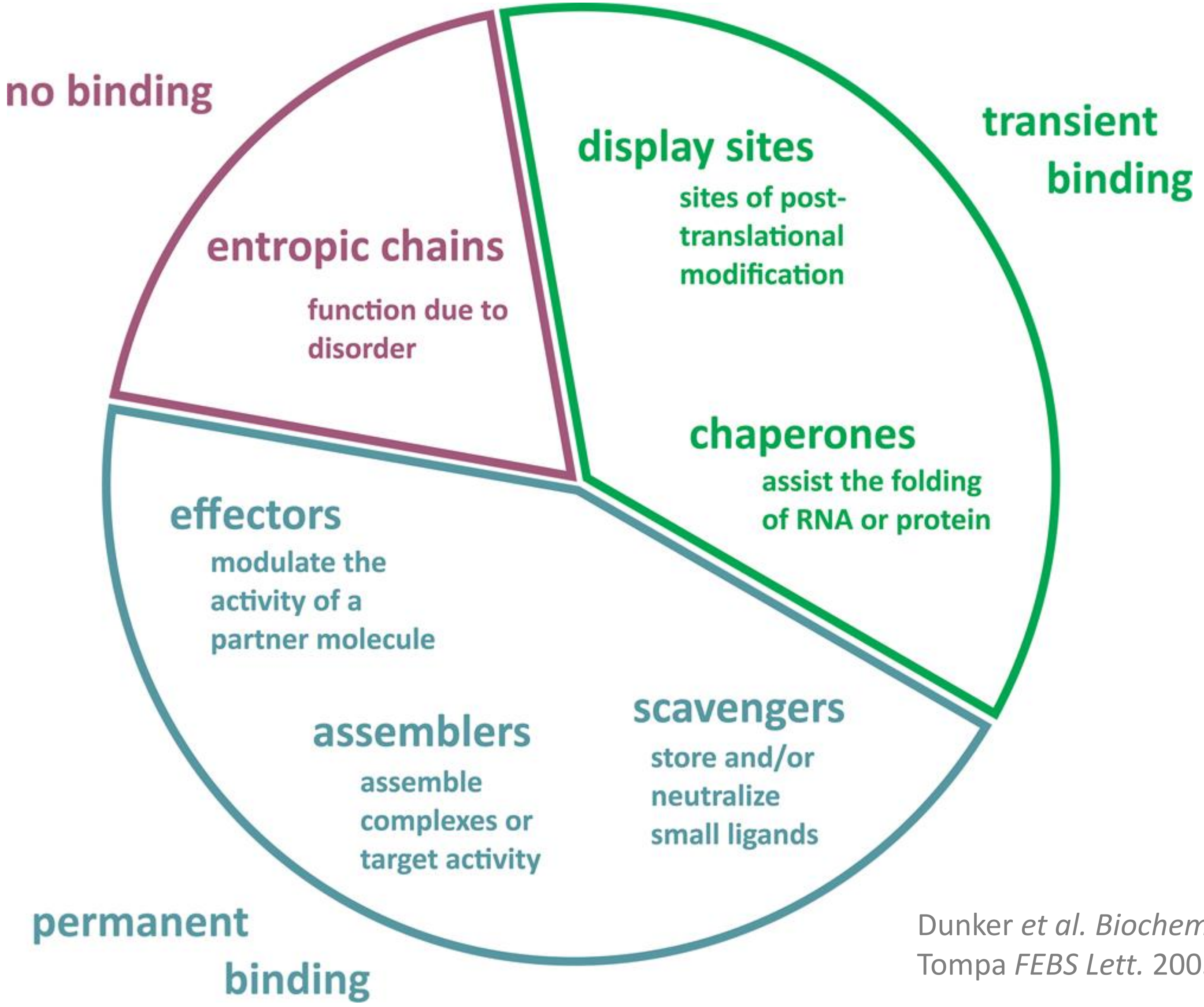
genes do not have any function annotation in their description in Ensembl. Genes with no annotation contain proportionally more IDRs.

Disorder in human genes

Less than one half of all residues in the human proteome fall within structured domains. Not only do most residues of human proteins fall outside domains, a large fraction of these residues are also disordered.



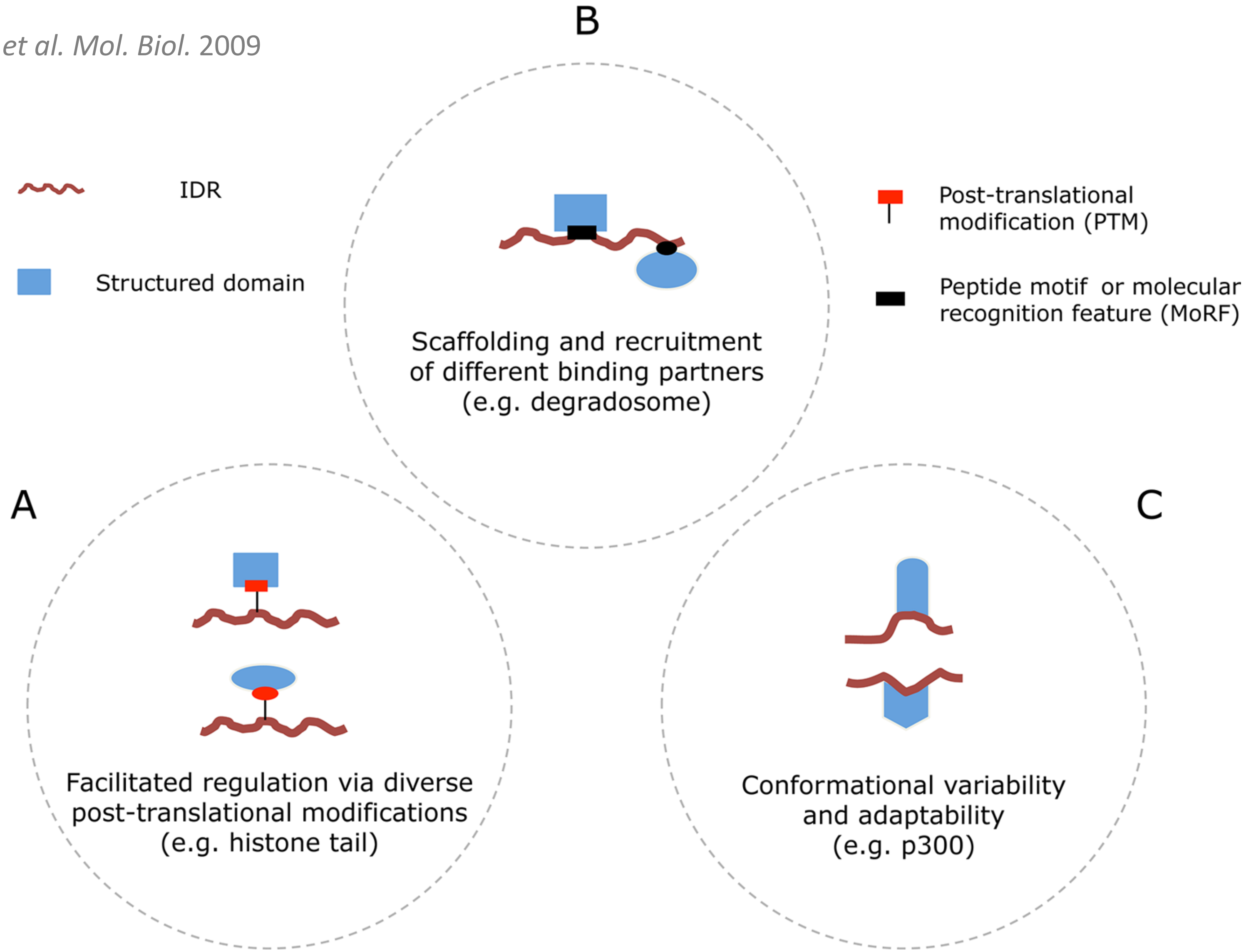
Functional classification of IDRs


no binding
entropic chains
function due to disorder
transient binding
display sites
sites of post-translational modification
chaperones
assist the folding of RNA or protein
effectors
modulate the activity of a partner molecule
assemblers
assemble complexes or target activity
scavengers
store and/or neutralize small ligands
permanent binding

Dunker *et al. Biochemistry* 2002
Tompa *FEBS Lett.* 2005

Functional classification of IDRs

Gsponer *et al. Mol. Biol.* 2009



IDR

Structured domain

Post-translational modification (PTM)

Peptide motif or molecular recognition feature (MoRF)

B

Scaffolding and recruitment of different binding partners (e.g. degradosome)

A

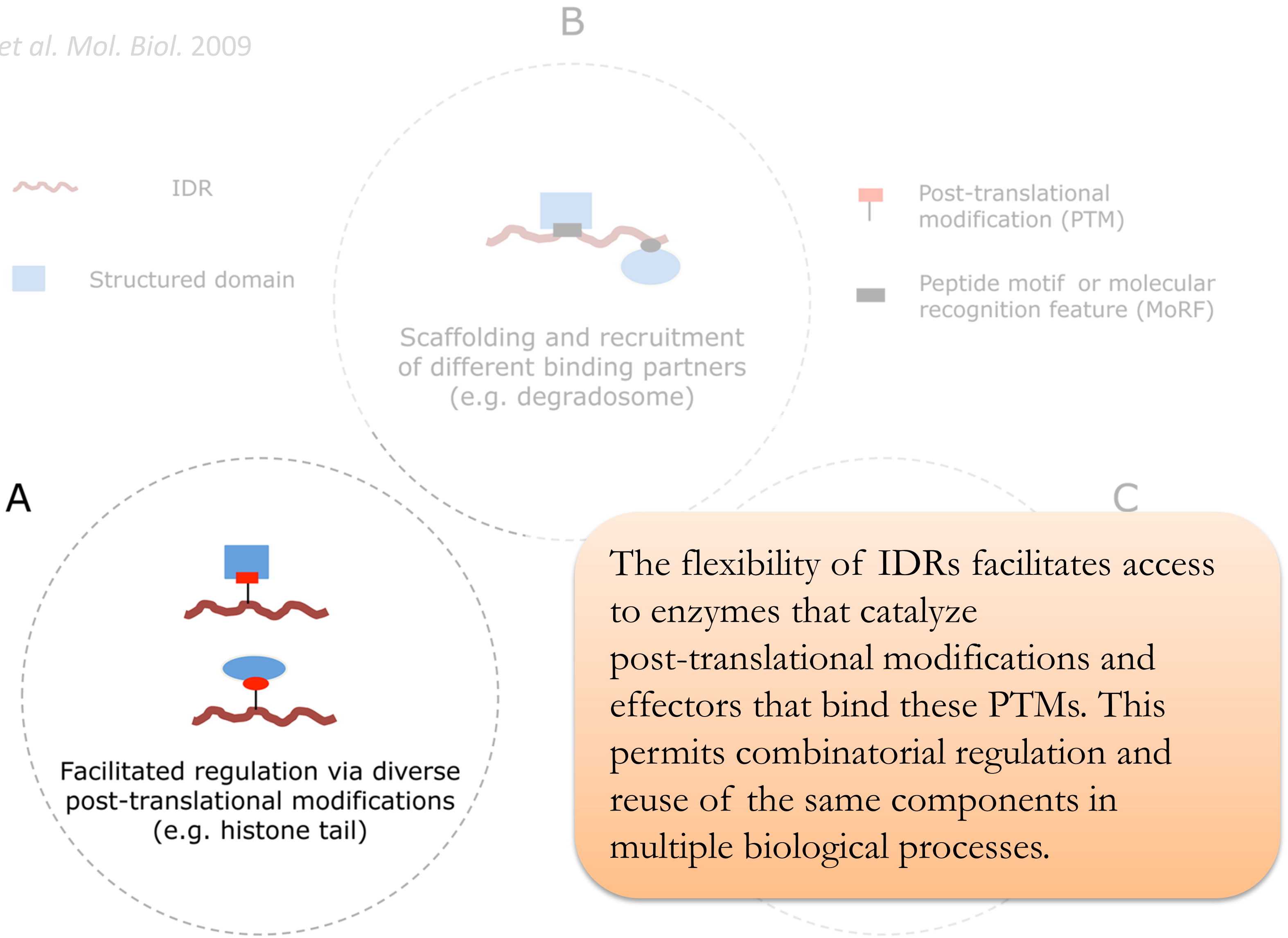
Facilitated regulation via diverse post-translational modifications (e.g. histone tail)

C

Conformational variability and adaptability (e.g. p300)

Functional classification of IDRs

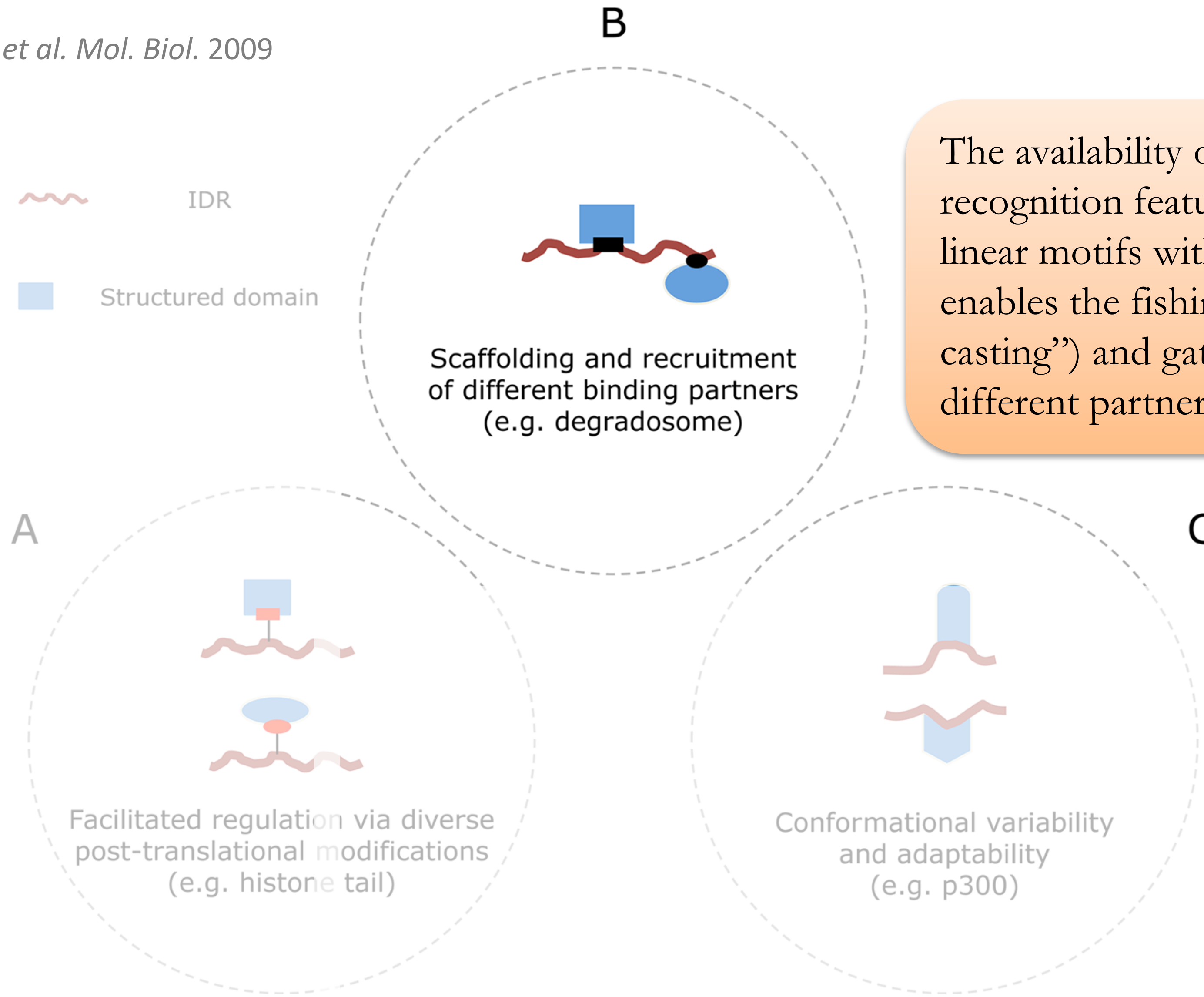
Gsponer *et al. Mol. Biol.* 2009


B
IDR
Structured domain
Post-translational modification (PTM)
Peptide motif or molecular recognition feature (MoRF)
Scaffolding and recruitment of different binding partners (e.g. degradosome)
A
Facilitated regulation via diverse post-translational modifications (e.g. histone tail)
C

The flexibility of IDRs facilitates access to enzymes that catalyze post-translational modifications and effectors that bind these PTMs. This permits combinatorial regulation and reuse of the same components in multiple biological processes.

Functional classification of IDRs

Gsponer *et al. Mol. Biol.* 2009



Scaffolding and recruitment of different binding partners (e.g. degradosome)

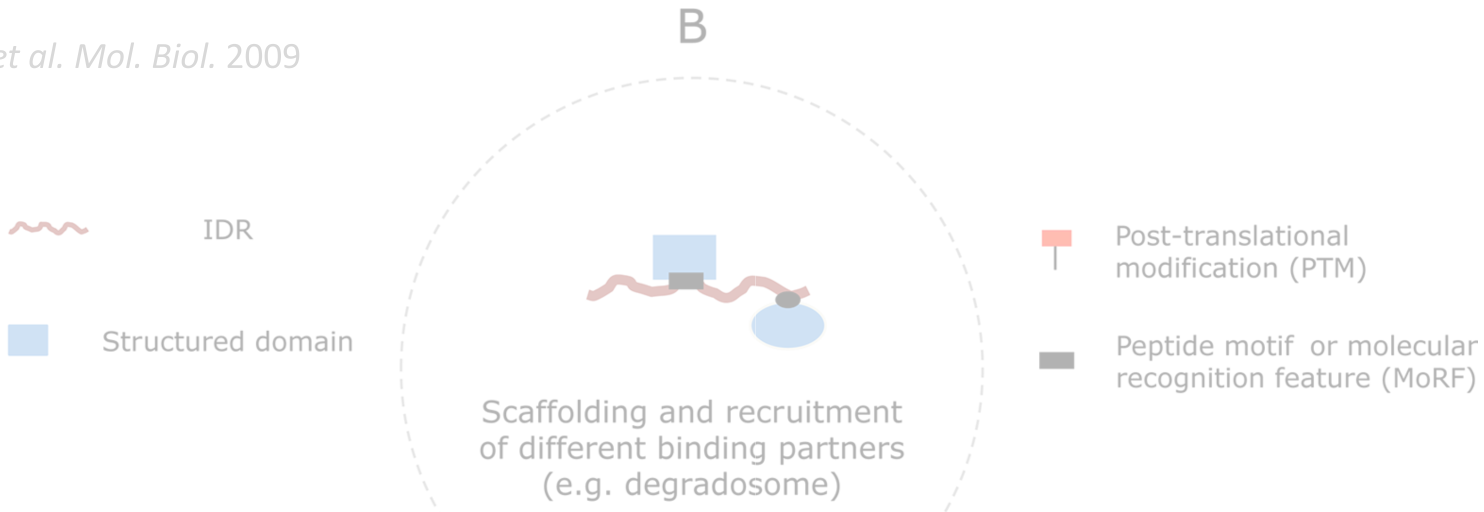
Facilitated regulation via diverse post-translational modifications (e.g. histone tail)

Conformational variability and adaptability (e.g. p300)

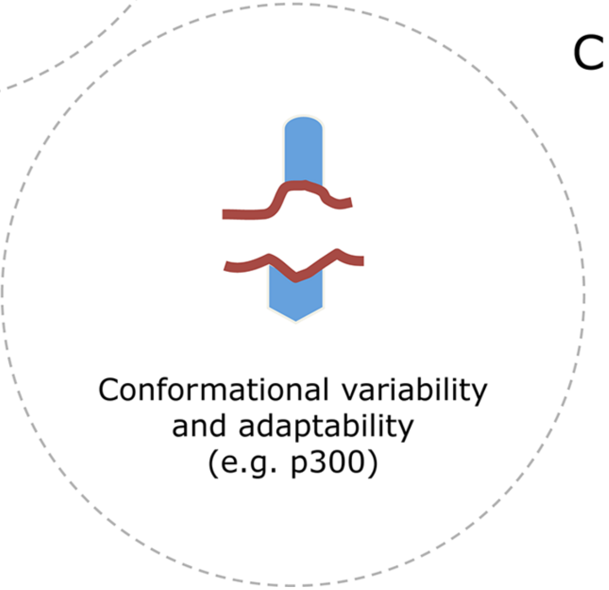
The availability of molecular recognition features and linear motifs within the IDRs enables the fishing for ("fly casting") and gathering of different partners.

Functional classification of IDRs

Gsponer *et al. Mol. Biol.* 2009

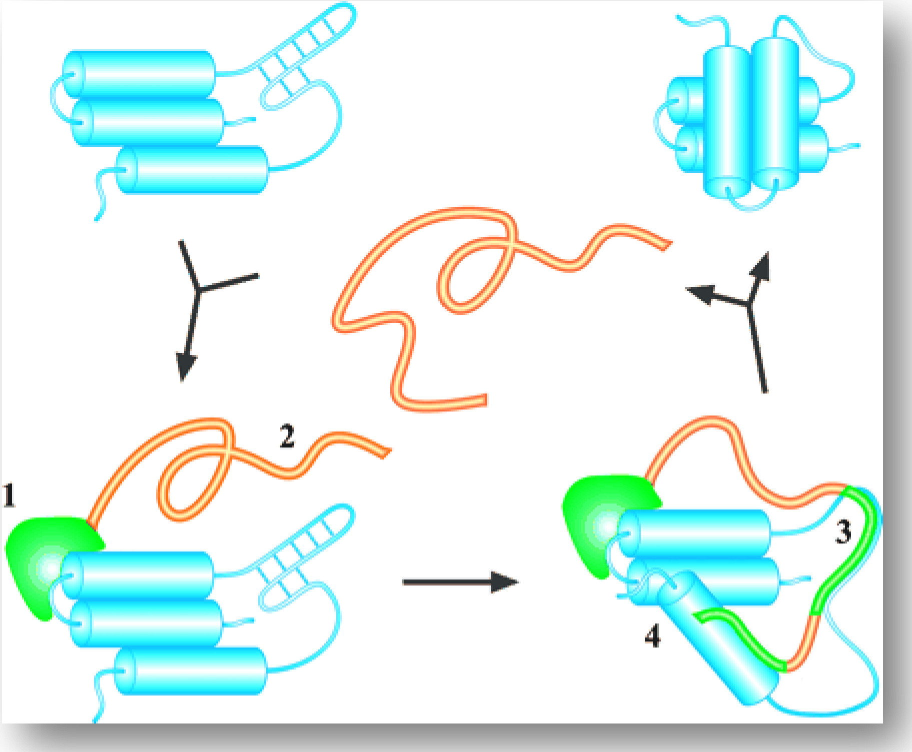

B
IDR
Structured domain
Post-translational modification (PTM)
Peptide motif or molecular recognition feature (MoRF)
Scaffolding and recruitment of different binding partners (e.g. degradosome)

Conformational variability enables a nearly perfect molding to fit the binding interfaces of very diverse interaction partners. Context-dependent folding of an IDR can activate signaling processes in one case or inhibit them in another, resulting in completely different outcomes



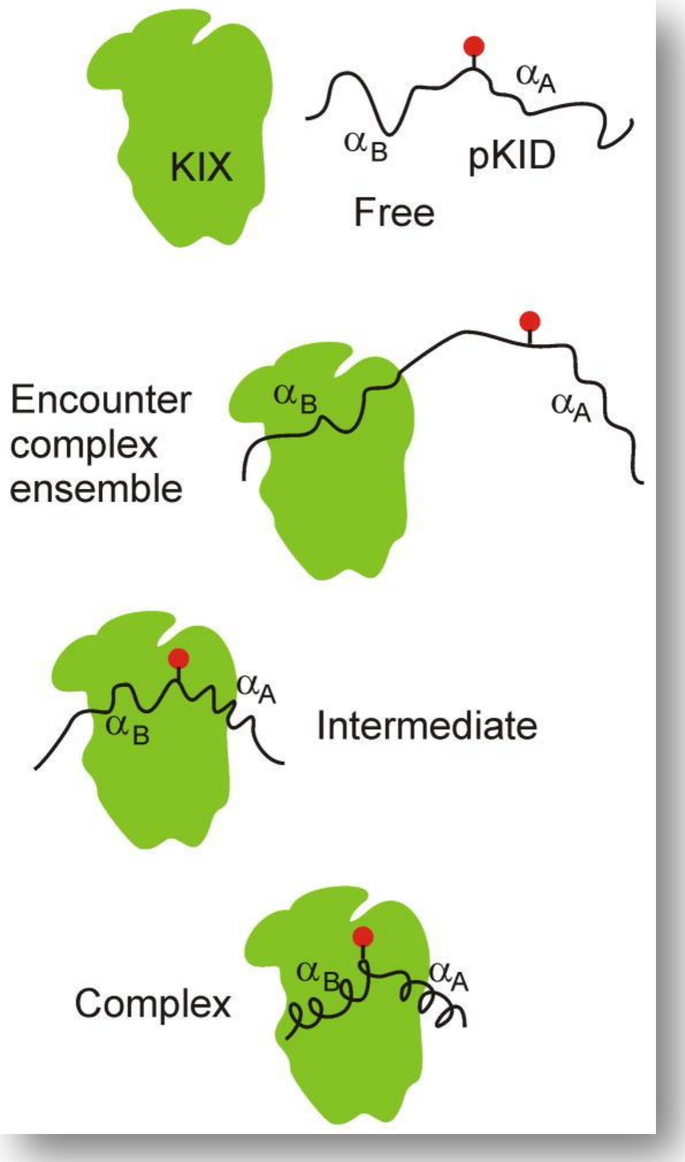
Entropy transfer model

Tompa and Csermely *FASEB J.* 2004

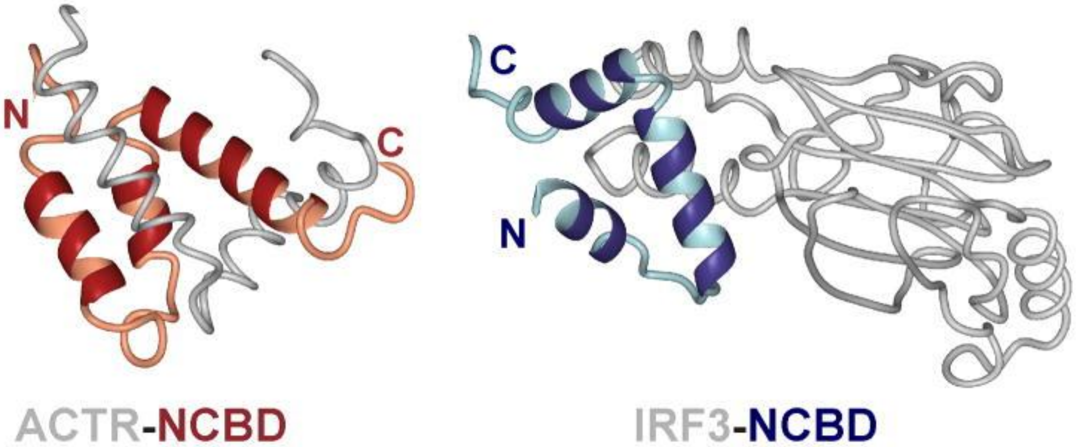


The rigidification of the chaperone upon binding enables the target protein to gain some flexibility and to refold in a correct manner.

Structuration upon binding

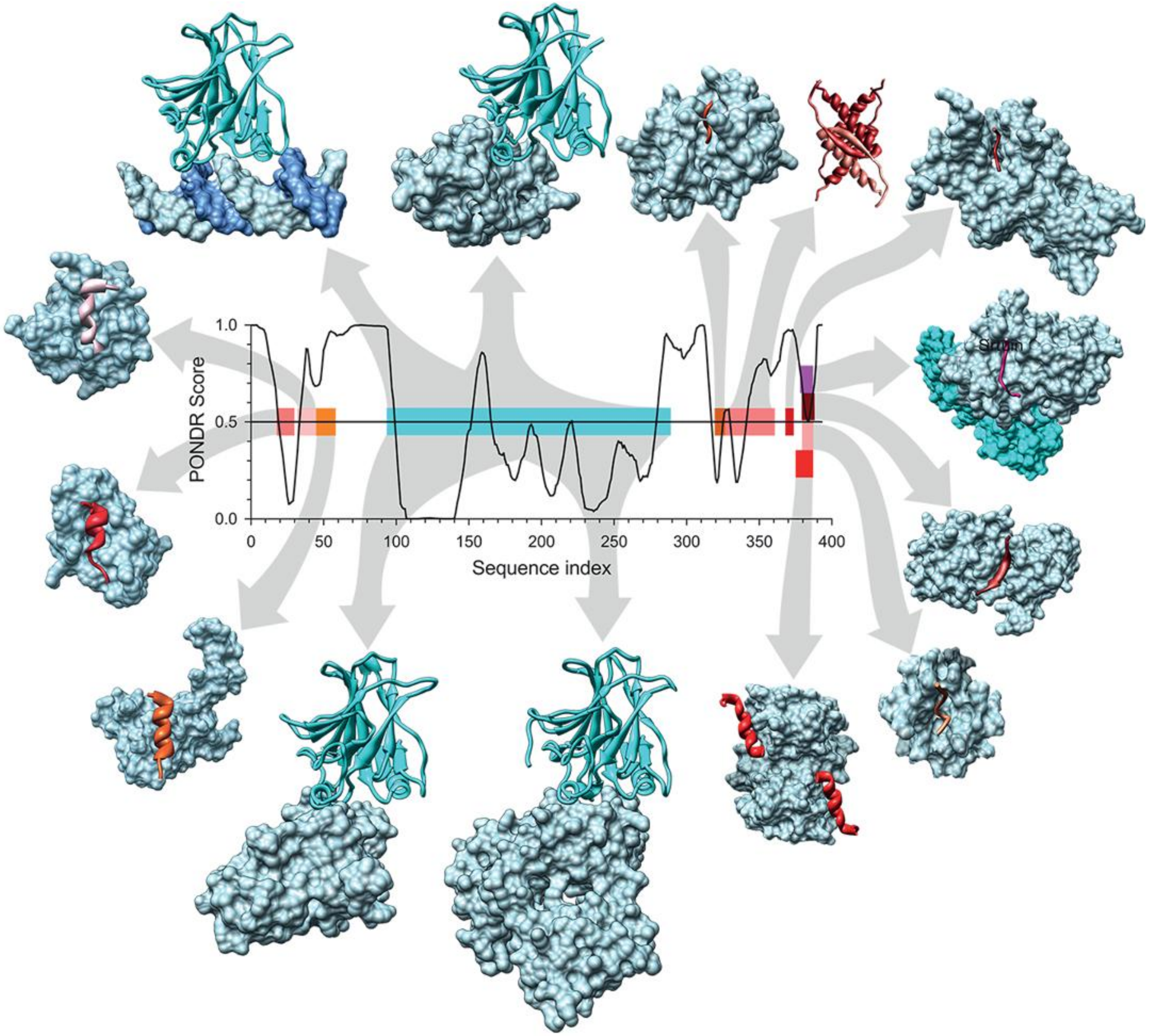


Wright and Dyson *Curr. Opin. Struct. Biol.* 2009

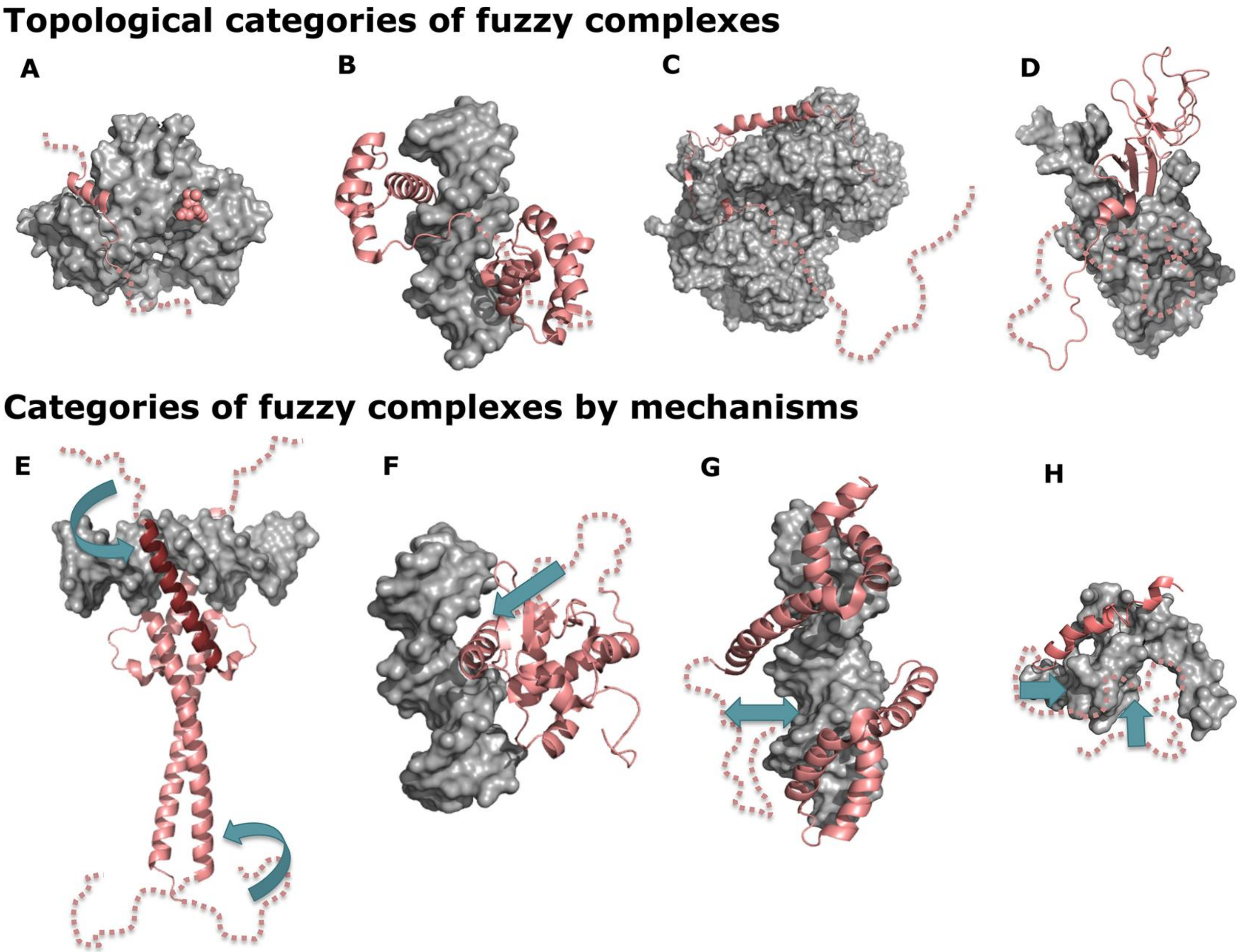


NCBD domain folds into topologically different helical arrangements unpon binding to two different partners

p53 multiple partners



Fuzzy complexes



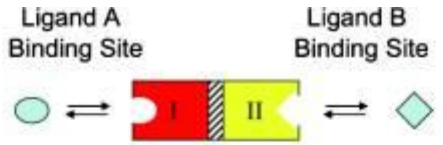
Topological categories of fuzzy complexes

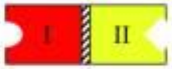
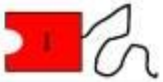
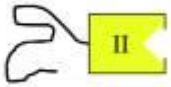
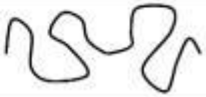
Categories of fuzzy complexes by mechanisms

Model for allosteric coupling

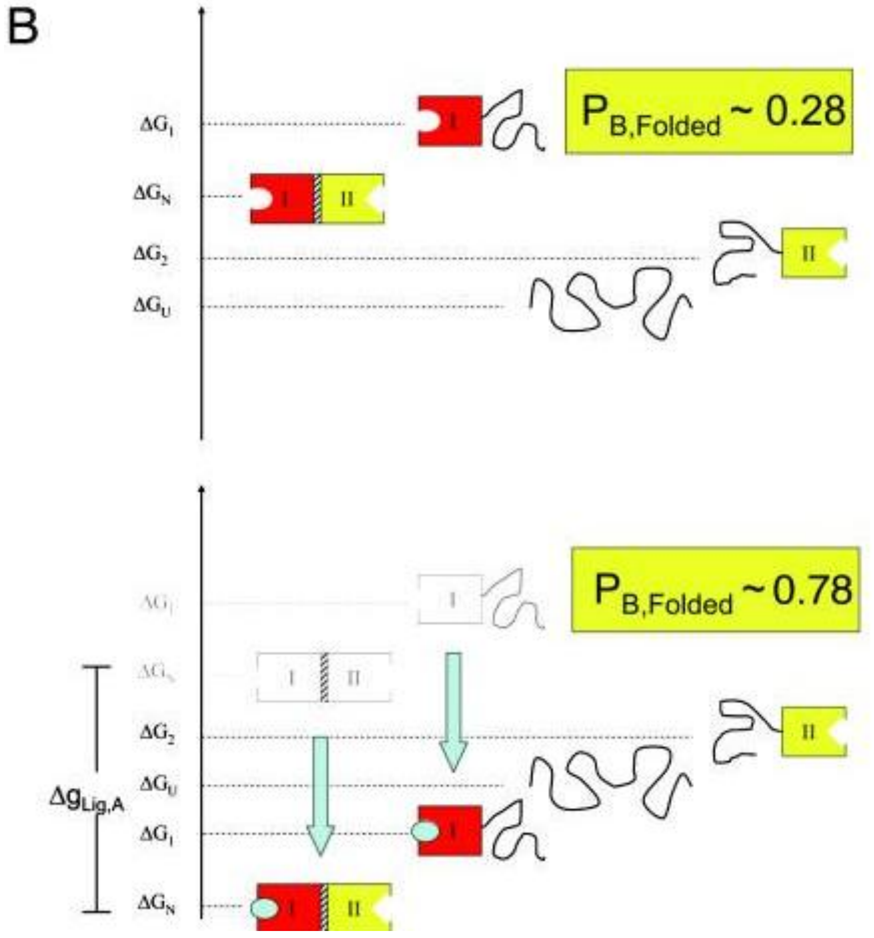
Hilser *et al.* *PNAS* 2007

A



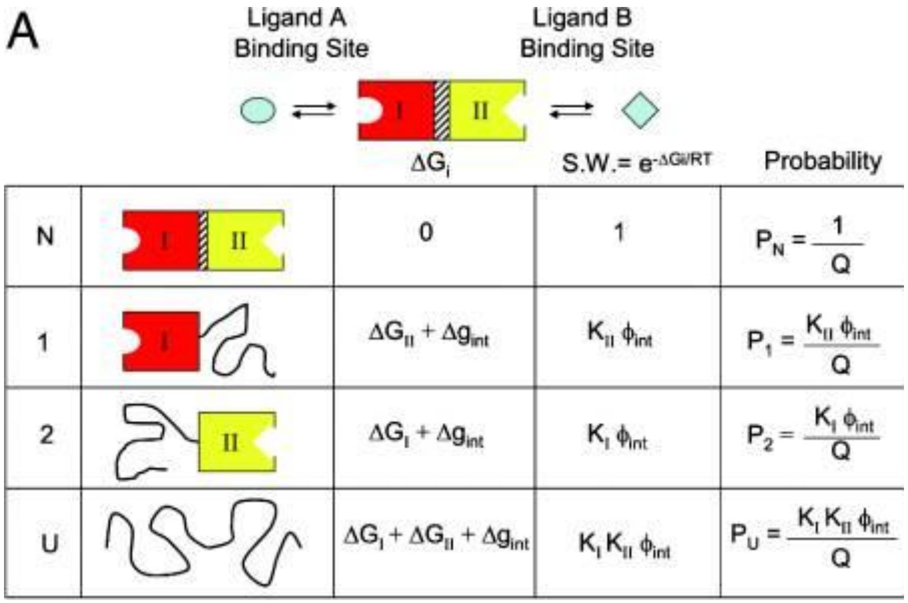
		ΔG_i	S.W.= $e^{-\Delta Gi/RT}$	Probability
N		0	1	$P_N = \frac{1}{Q}$
1		$\Delta G_{II} + \Delta g_{int}$	$K_{II}\ \phi_{int}$	$P_1 = \frac{K_{II}\ \phi_{int}}{Q}$
2		$\Delta G_I + \Delta g_{int}$	$K_I\ \phi_{int}$	$P_2 = \frac{K_I\ \phi_{int}}{Q}$
U		$\Delta G_I + \Delta G_{II} + \Delta g_{int}$	$K_I K_{II}\ \phi_{int}$	$P_U = \frac{K_I K_{II}\ \phi_{int}}{Q}$

The binding of ligand A redistributes the ensemble probabilities



Model for allosteric coupling

Hilser *et al.* *PNAS* 2007



		ΔG_i	S.W.= $e^{-\Delta Gi/RT}$	Probability
N		0	1	$P_N = \frac{1}{Q}$
1		$\Delta G_{II} + \Delta g_{int}$	$K_{II}\,\phi_{int}$	$P_1 = \frac{K_{II}\,\phi_{int}}{Q}$
2		$\Delta G_{I} + \Delta g_{int}$	$K_{I}\,\phi_{int}$	$P_2 = \frac{K_{I}\,\phi_{int}}{Q}$
U		$\Delta G_I + \Delta G_{II} + \Delta g_{int}$	$K_I K_{II}\,\phi_{int}$	$P_U = \frac{K_I K_{II}\,\phi_{int}}{Q}$

$$Q = 1 + K_{II}\phi_{\text{int}} + K_I\phi_{\text{int}} + K_I K_{II}\phi_{\text{int}}$$

$$Q = Z_{Lig,A}(1 + K_{II}\phi_{\text{int}}) + K_I\phi_{\text{int}} + K_I K_{II}\phi_{\text{int}}$$

$$\Delta g_{Lig,A} = -RT \bullet \ln Z_{Lig,A} = -RT \bullet \ln(1 + K_{a,A}[A])$$

without ligand:

$$P_{B,Folded} = P_N + P_2 = \frac{1 + K_I\phi_{\text{int}}}{1 + K_{II}\phi_{\text{int}} + K_I\phi_{\text{int}} + K_I K_{II}\phi_{\text{int}}}$$

with ligand:

$$P_{B,Folded} = \frac{Z_{Lig,A} + K_I\phi_{\text{int}}}{Z_{Lig,A}(1 + K_{II}\phi_{\text{int}}) + K_I\phi_{\text{int}} + K_I K_{II}\phi_{\text{int}}}$$

Disorder prediction methods

Three general prediction strategies currently exists:

- **based on sequence properties**

  IUPred estimates residue interaction energies

  FoldIndex considers weakly hydrophobic regions of high net charge

- **machine learning**

  DISOPRED2 uses linear SVMs trained on PSI-BLAST sequence profiles around UR

  PONDR XL1 employs a feed-forward NN trained on sequence attributes of UR

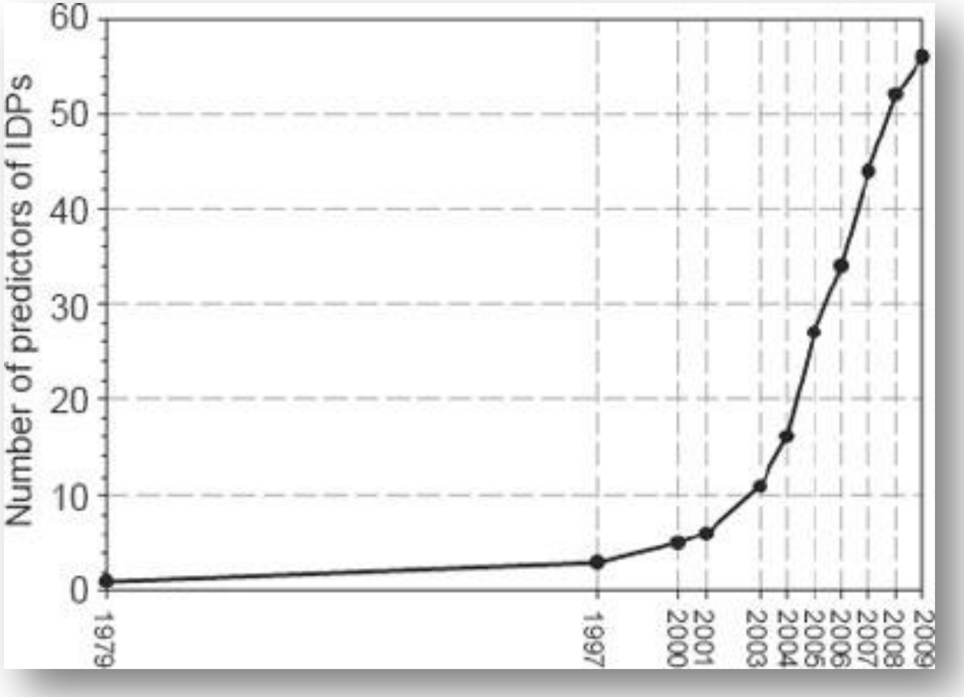
- **meta-predictors**

  metaPrDOS and MFDp apply SVM to individual prediction methods

  MobiDB and C^{2}P^2 databases provide a consensus overview

Disorder prediction methods

The total number of IDP predictors



He *et al.* (2009) *Cell Research*

Since the first predictors were published, more than 50 predictors of disorder have been developed. Many of these predictors can be accessed via public servers and evaluate intrinsic disorder on a per-residue basis.

Disorder prediction methods

- Williams (1979) *Biol Rev Camb Philos Soc.*

$$r = \frac{\#(\text{charged aas})}{\#(\text{hydrophobic aas})}$$

Classifier for non-folding vs structured proteins

- very good performance on a few examples
- disappointing results upon generalization

- Romero (1997) *Proc IEEE Int Conf Neural Networks*

First formal predictor based on aa composition

Predictors of Natural Disordered Regions (PONDR)

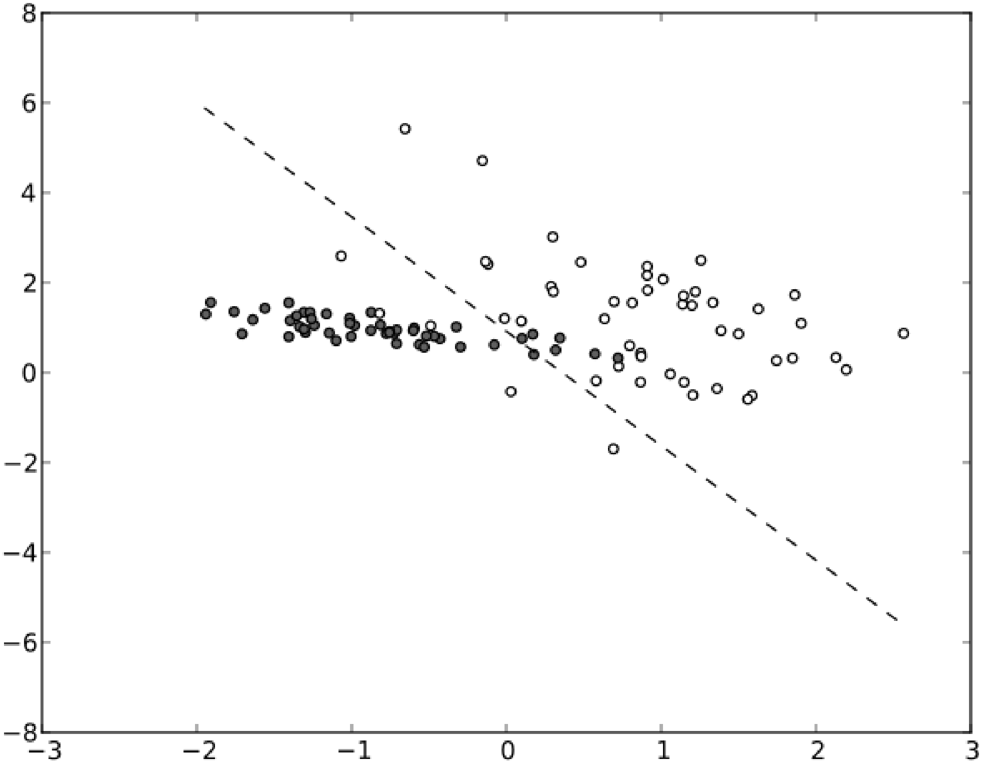
Inputs:

- amino acid composition
- attributes such as sequence complexity
- attributes such as hydropathy, net charge,…

Combined in a non-linear manner: ANNs, SVMs, logistic regression…

Support Vector Machines

SVMs are supervised learning models.



Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds **a model that assigns new examples into one category or the other**.

It constructs **hyperplane(s) in a high- to infinite-dimensional space**.

Intuitively, a good separation is achieved by the **hyperplane that has the largest distance to the nearest training data point of any class** (so-called functional margin).

PONDR methods

PONDR VL-XT (Romero 2001). The disordered structure characteristics might depend on the location of the disordered region in the sequence. Three training sets:

- Variously characterized Long (>30 aas) disordered regions
- X-ray characterized Terminal regions (N- & C-term)

VL3-E (Peng 2005). Combination of the ANN-based predictors:

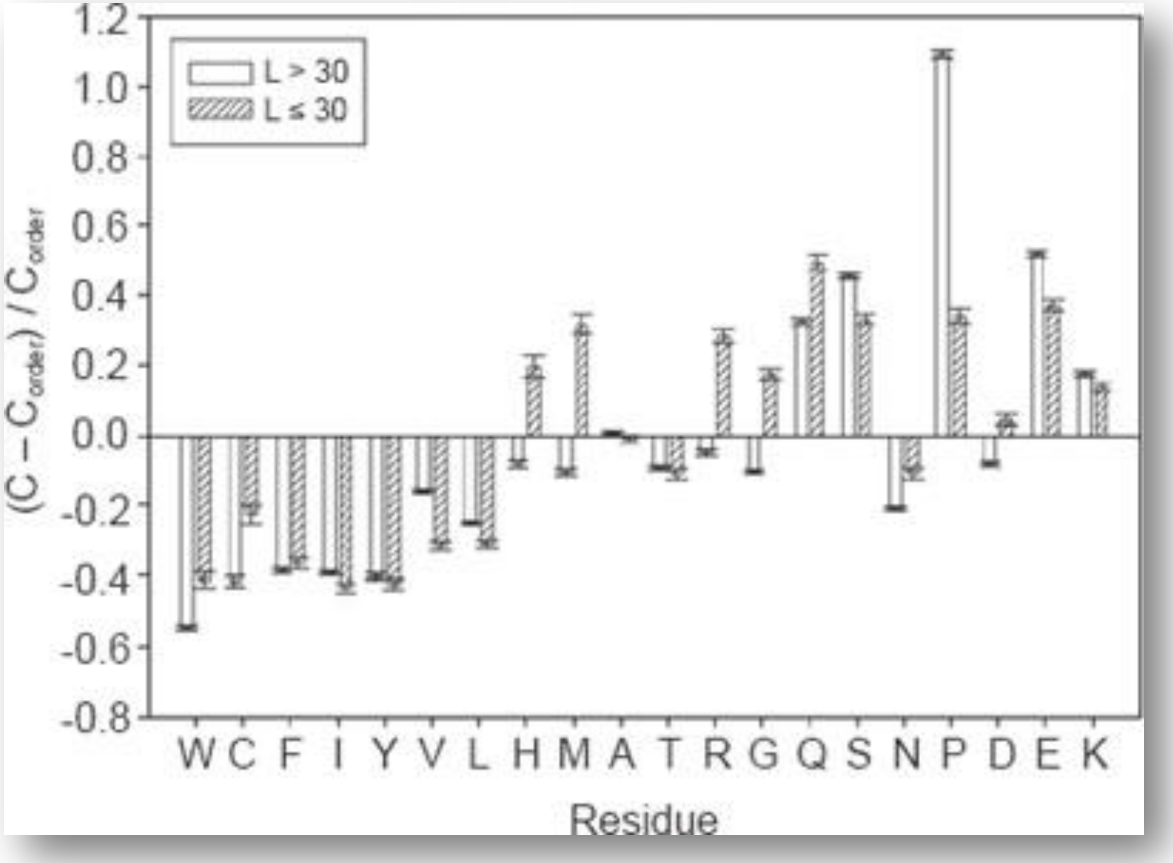
- VL3-H: searches for homologous sequences to enrich training set
- VL3-P: considers PSI-BLAST profiles as input attributes

PONDR VSL. Variously characterized Short and Long disordered regions

- VSL1 (Obradovic 2005): ANN + logistic regression models
- VSL2 (Peng 2006): SVMs + logic regression model

Short and long disordered regions

Relative aa compositions for short and long disordered regions



Short disordered regions are more depleted in I, V, and L, while **long disordered regions** are more enriched in K, E, and P but are less enriched in Q. In addition, long disordered regions are depleted in G and N, while short disordered regions are enriched in G and D.

Disorder prediction methods

- Uversky (2000) *Proteins*

folding of a protein is governed by a balance between attractive forces (*e.g.*, hydrophobic interactions) and repulsive forces (Coulomb or electrostatic repulsion).

$$r = \frac{\text{mean net charge}}{\text{mean hydropathy}}$$

binary predictor in the form of a CH-plot

- Prilusky (2005) *Bioinformatics*

FoldIndex: per-residue disorder predictor that computes the CH ratio along the protein and predict if a local region in given sequence is in a disordered structure.

- Linding (2003) *Nucleic Acids Res.*

GlobPlot:

- sum function using amino acid scale based on $p(c_i)/p(ss_i)$
- digital low-pass filter based on the Savitzky-Golay algorithm
- numerical estimation of the first order derivative
- plot using the DISLIN 8.0 package
- selection of putative globular and disordered segments by a peak finder algorithm

Disorder prediction methods

- Linding (2003) *Structure*

  DisEMBL: three separate ANN predictors, to predict three kinds of disordered structures in proteins

  - loops/coils (as defined by DSSP)
  - hot loops (loops with high B-factors)
  - Remark 465 (missing from the PDB X-ray structures)

- Jones (2003) *Proteins*

  DISOPRED: feed-forward neural network with a large number of hidden units

  - over-fitting and slow training

- Ward (2004) *Bioinformatics*

  DISOPRED2: support vector machines directly trained on the whole sequence.
  Training set: 715 high resolution (<2Å) X-ray structures with less than 25% seq id.
  (176 550 ordered and 4 590 disordered residues)
  Various combinations of descriptors:

  - binary-encoded aa sequence
  - secondary structure predictions
  - PSI-BLAST profiles

Disorder prediction methods

- Weathers (2004) *FEBS Lett*

  Linear combination of the composition vectors using either the full or a reduced amino acid alphabet.
  Training set: 1 190 ordered proteins + 718 disordered segments
  Representation: each protein is represented by a vector set, on vector of *n* elements for each aa (*n* = 20, 15, 10, 8, 4)
  Optimization: SVMs to find optimal weights by taking linear combinations of compositions vectors (e.g. dot kernel function K(sj,x) = sj · x)

- MacCallum

  DRIPRED: based on Kohonen's self-organizing map (SOM)

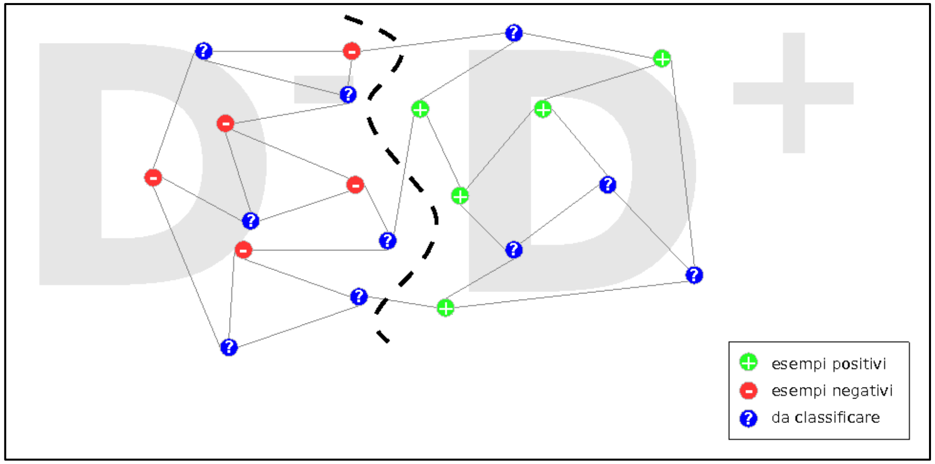
  1. selected data are made non-redundant using a single-pass hashing approach
  2. PSI-BLAST to obtain PSSM for each sequence of length L
  3. profile windows of sequences are mapped into an SOM
  4. predictions are made based on hit frequencies to certain areas in the map

  Sequences that mapped to parts of UniProt space unpopulated by known structures are assumed to correspond to disordered regions.

Disorder prediction methods

- Shimizu (2007) *BMC Bioinformatics*

Uses information from structure-unknown proteins to avoid training data sparseness.
Training set: huge amount of structure-unknown & structure-known sequences
Classification technique: Spectral Graph Transducer (SGT)



The SGT is used to construct a **k-nearest-neighbor graph**, which takes into account the information on the unlabeled data. SGT assigns a label to U by dividing the graph D into two subgraphs, D^+ and D^- . The data with structure-unknown information can **expand the training set** and **improve the accuracy** of the disorder prediction.

Disorder prediction methods

❖ Ishida (2007) *Nucleic Acids Res*

PrDOS uses the alignment of homologs with templates that have been determined to improve predictions.

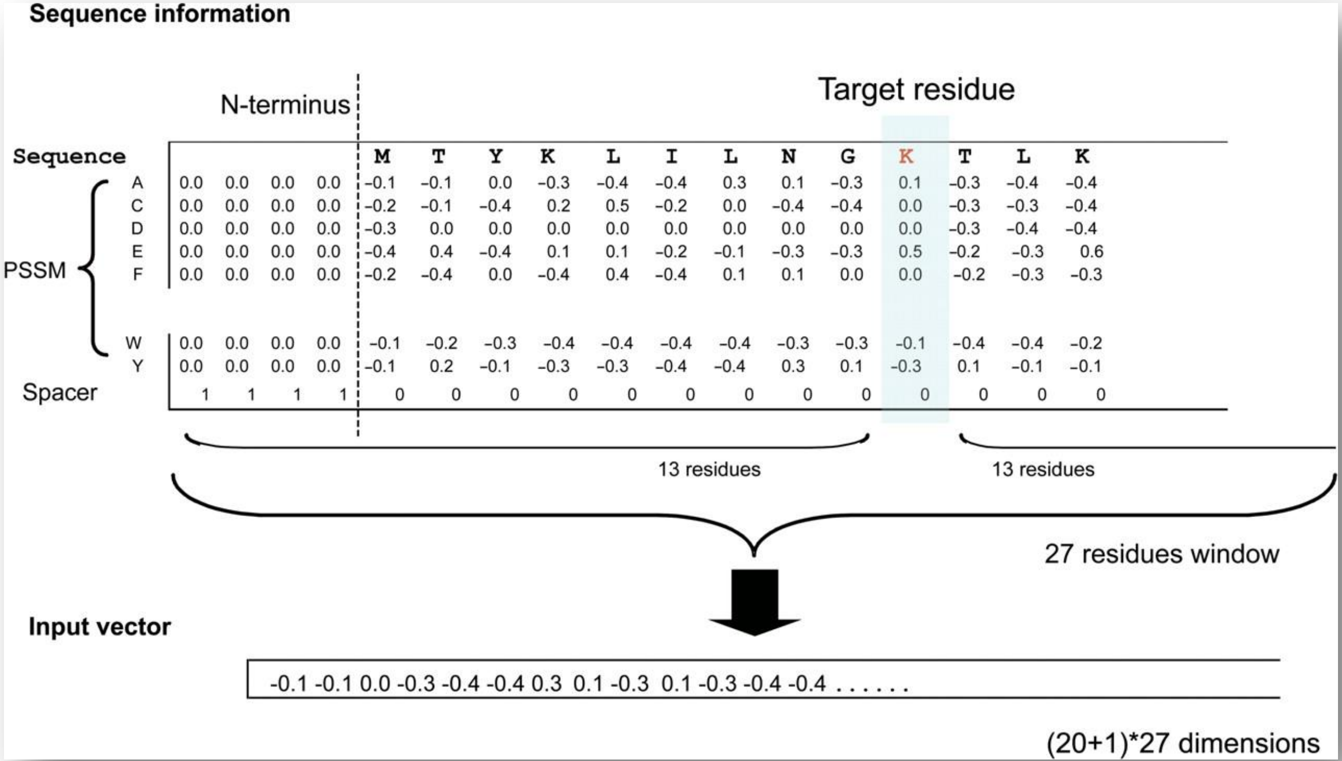
1. Generate a PSSM using two rounds of PSI-BLAST
2. Predict disorder using a SVM based on local amino acid sequence information

Disorder prediction methods

- Ishida (2007) *Nucleic Acids Res*

PrDOS uses the alignment of homologs with templates that have been determined to improve predictions.

1. Generate a PSSM using two rounds of PSI-BLAST
2. Predict disorder using a SVM based on local amino acid sequence information



Sequence						M	T	Y	K	L	I	L	N	G	K	T	L	K
PSSM	A	0.0	0.0	0.0	0.0	–0.1	–0.1	0.0	–0.3	–0.4	–0.4	0.3	0.1	–0.3	0.1	–0.3	–0.4	–0.4
	C	0.0	0.0	0.0	0.0	–0.2	–0.1	–0.4	0.2	0.5	–0.2	0.0	–0.4	–0.4	0.0	–0.3	–0.3	–0.4
	D	0.0	0.0	0.0	0.0	–0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	–0.3	–0.4	–0.4
	E	0.0	0.0	0.0	0.0	–0.4	0.4	–0.4	0.1	0.1	–0.2	–0.1	–0.3	–0.3	0.5	–0.2	–0.3	0.6
	F	0.0	0.0	0.0	0.0	–0.2	–0.4	0.0	–0.4	0.4	–0.4	0.1	0.1	0.0	0.0	–0.2	–0.3	–0.3
	W	0.0	0.0	0.0	0.0	–0.1	–0.2	–0.3	–0.4	–0.4	–0.4	–0.4	–0.3	–0.3	-0.1	–0.4	–0.4	–0.2
	Y	0.0	0.0	0.0	0.0	–0.1	0.2	–0.1	–0.3	–0.3	–0.4	–0.4	0.3	0.1	-0.3	0.1	–0.1	–0.1
Spacer		1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0

-0.1 -0.1 0.0 -0.3 -0.4 -0.4 0.3 0.1 -0.3 0.1 -0.3 -0.4 -0.4 . . . . . .

Disorder prediction methods

- Ishida (2007) *Nucleic Acids Res*

PrDOS uses the alignment of homologs with templates that have been determined to improve predictions.

1. Generate a PSSM using two rounds of PSI-BLAST
2. Predict disorder using a SVM based on local amino acid sequence information
3. Predict disorder based on the alignments with homologues with known structures

$$P_i = \frac{\sum_{j=1}^{n} \alpha_j I_j}{n}$$

n: number of alignments
Ij: sequence identity of the jth hit
αj: set to 1 if the aligned residue in the jth hit is disordered

4. Combine the results of the two independent predictions by computing the weighted average (w1=1.0, w2=0.11)

Disorder prediction methods

- Bulashevska (2008) *J Theor Biol*

Bayesian classifier: Each protein sequence belonging to a certain class can be considered as a realization of an independent random process that emits symbols from an alphabet of 20 amino acids.

The probability of a sequence *s* to come from a certain class *c* is given by a multinomial probability function governed by its vector of parameters $\theta_c=(\theta_{c1},\ldots,\theta_{c20})\in[0,1]$:

$$p(s|\theta_c) = \frac{n!}{\prod_{i=1}^{20} n_i!} \prod_{i=1}^{20} \theta_{c_i}^{n}$$

n: length of the sequence
ni: occurrences of amino acid i
θci: *c* th class-conditional probability of amino acid *i* to occur in a sequence

According to Bayes' rule, the class for an unlabeled sequence *s* can be inferred using the posterior probability

$$p(c|s) = \frac{p(c)\,p(s|c)}{p(s)} = \frac{p(c)\,p(s|c)}{\sum_c p(c)\,p(s|c)}$$

Consensus methods

Metapredictors combine the outputs of several individual predictors. They can be applied either at the residue level or at the whole sequence level.

- Oldfield(2005) *Biochemistry*

  based on two distinct binary classifiers, the CH-plot and the CDF analysis

- Xue (2009) *FEBS Lett*

  combination of several CDF predictors developed from several disorder predictors, including PONDR®s VLXT, VSL2, and VL3, TopIDP, IUPred, and FoldIndex. A neural network is used to combine these individual CDF-based predictions. (improved accuracy by 5%-10%)

- Ishita and Kinoshita (2008) *Bioinformatics*

  metaPdDOS: metapredictor for per-residue estimates of order and disorder. SVM to integrate residue-level predictions from PrDOS, DISOPRED2, DisEMBL, DISPROT, DISpro, IUPred, and POODLE-S (AUC =0.877)

Conclusion

- **A large number of proteins** possess regions that are disordered. This disorder can be an intrinsic property of the regions, depending on their amino acid composition.
- **Intrinsically disordered regions** often play regulatory roles through the binding of (multiple) partners which are recruited through motifs or post-translational modification
- **Disorder prediction algorithms** are based on the analysis of the input sequences and generally employs machine learning methods.