Intrinsically Disordered Protein, Alternative Splicing & Post-Translational Modification (IDP-AS-PTM): A Toolkit for Developmental Biology

A. Keith Dunker
Indiana University School of Medicine,
(kedunker@iupui.edu)

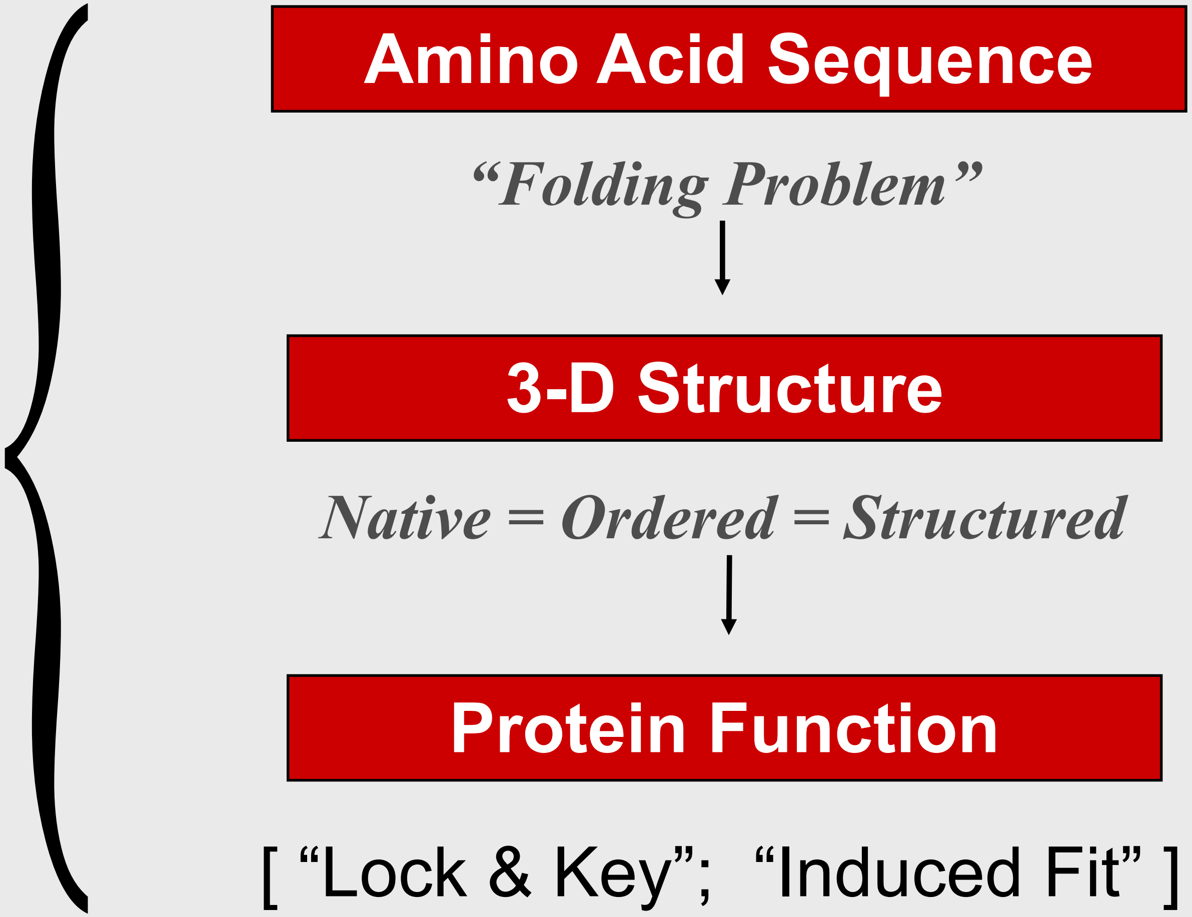
Wednesday, February 27, 2019
Institute for Quantitative Biomedicine
Center for Biotechnology & Medicine
Rugters University
New Brunswick, NJ

Center For Computational Biology and Bioinformatics

School of Medicine
INDIANA UNIVERSITY

Textbook Protein Structure/Function

Currently
Dominant
Protein
Structure/
Function
Paradigm



Amino Acid Sequence

"Folding Problem"

3-D Structure

Native = Ordered = Structured

Protein Function

["Lock & Key"; "Induced Fit"]

Definition: Intrinsically Disordered Proteins (IDPs) and IDP Regions

Whole proteins and regions of proteins are intrinsically disordered if:

- **they *lack stable 3D structure* under physiological conditions, and if:**
- **they are flexible molecules that form *dynamic ensembles* with *inter-converting configurations* and *without particular equilibrium values* for their *coordinates*.**

Intrinsically Disordered Proteins (IDPs)

- Karl Landsteiner (1939) & Linus Pauling (1940) suggested that unfolded proteins exist and that they fold into different structures as they bind separately to multiple, differently shaped partners.

- IDPs first characterized in the 1950s by OR & ORD.

- Thousands now characterized by X-ray, NMR, etc., & especially by computational biology & bioinformatics.

- IDP discovery represents a true paradigm shift.

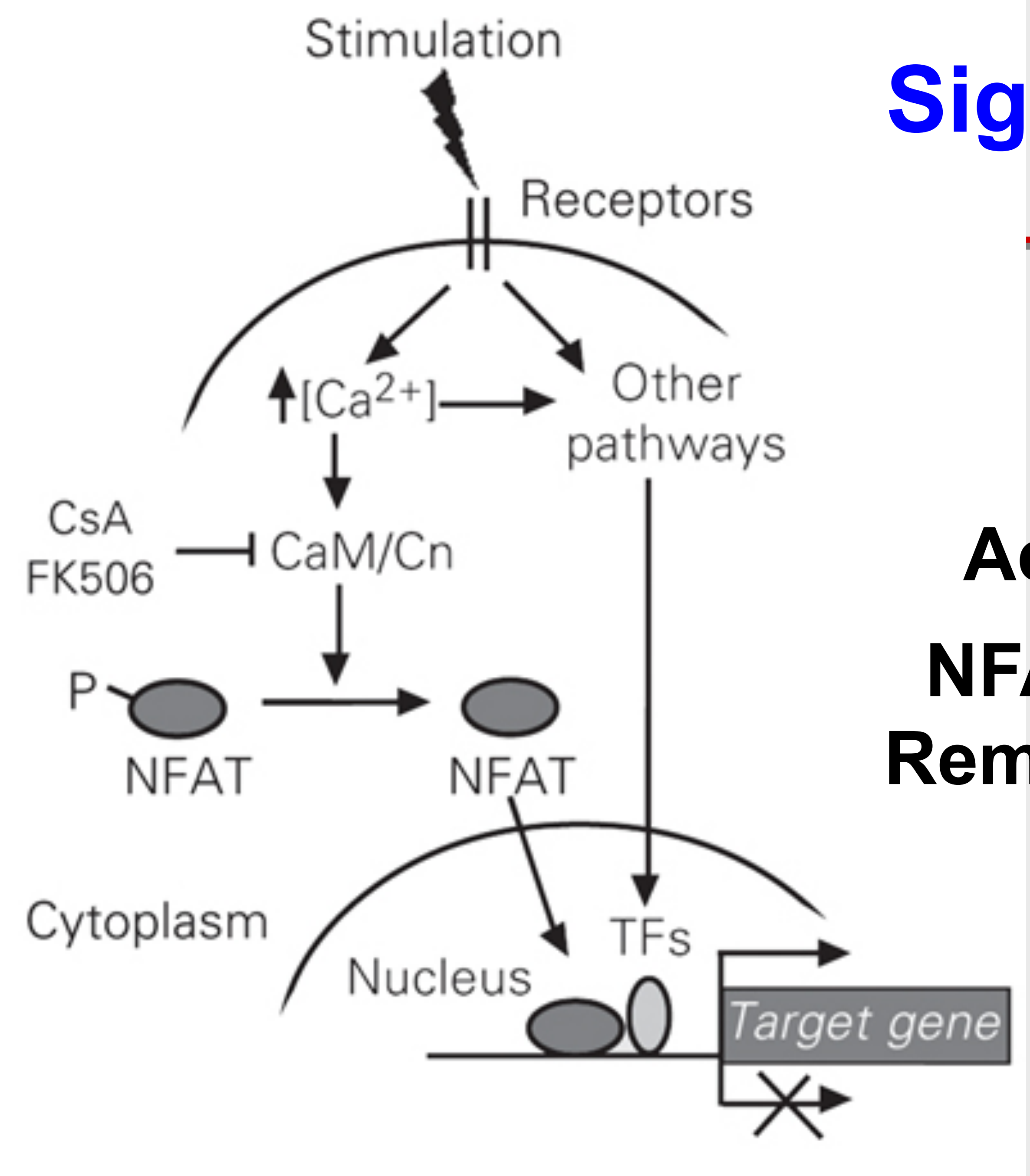
Reviewed in Dunker AK & Oldfield CJ *Adv Exp Med Biol* 870: 1-34 (2015)

Trigger for Dunker's IDP Research

Seminar describing an important IDP
12 Noon to 1 PM, 15 November, 1995
Washington State University

Given By Chuck Kissinger
BS / MS Washington State University
PhD University of Washington
Johns Hopkins / MIT Post Doc
Aguoron Pharmaceuticals

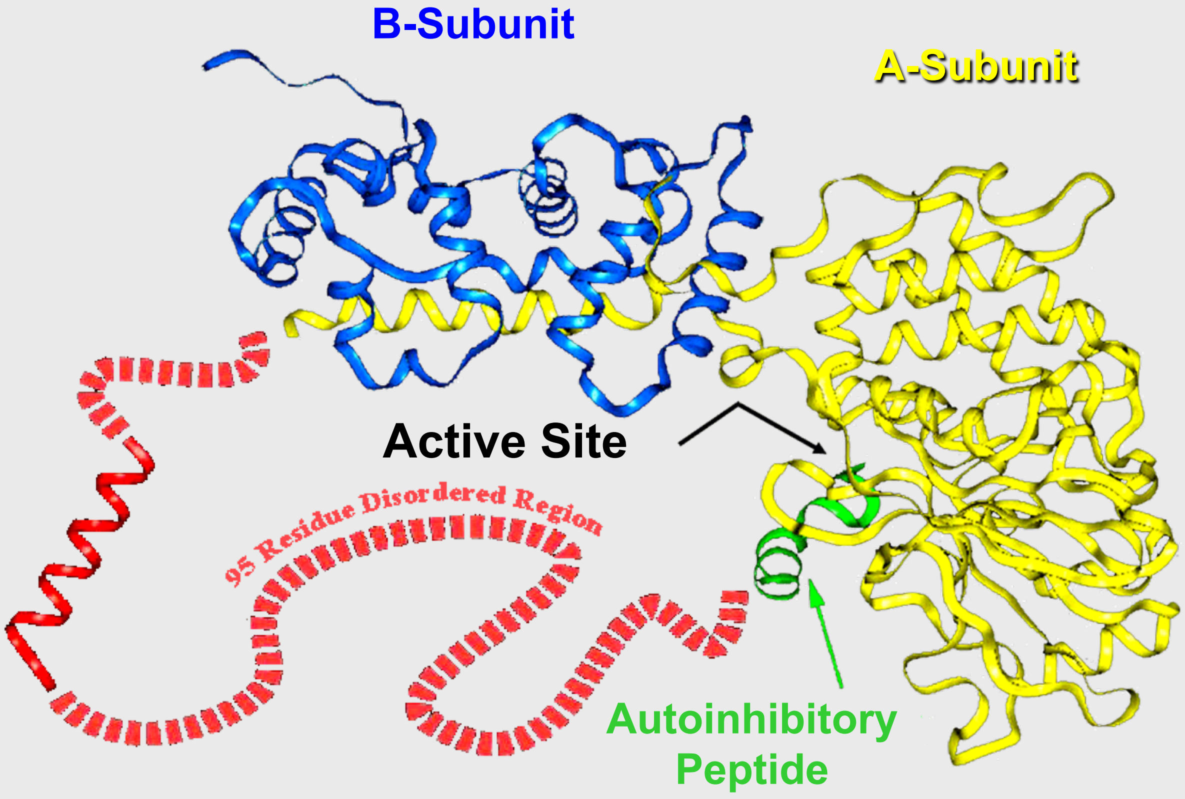
Signaling Pathway



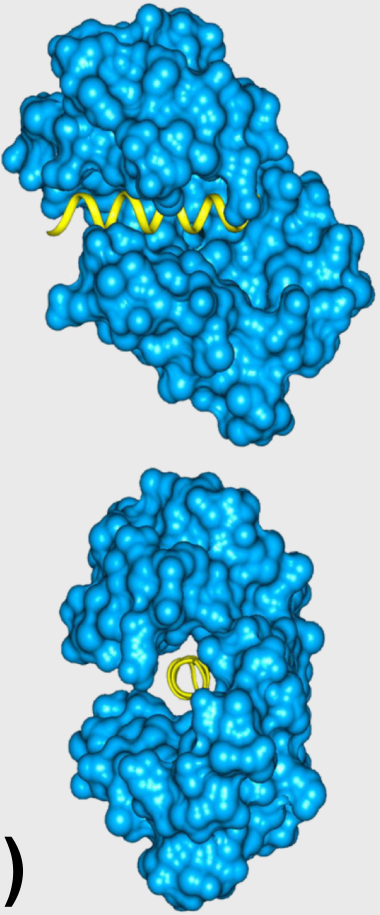
Calmodulin (CaM)
Calcineurin (Cn)
Nuclear Factor of Activated T- Cells (NFAT)

NFAT-poly-P in an IDP tail.
Remove Ps, activates NLS
→ NFAT → nucleus
→ turns on genes
→ T-cells activated
→ reject transplant

Calcineurin and Calmodulin



Meador W et al., *Science* 257: 1251-1255 (1992)



Kissinger C et al., *Nature* 378:641-644 (1995)

Intrinsically Disordered Proteins (IDPs)
After Seminar Questions:

- Why don't IDPs and IDP regions fold into 3D structure?

- How common are IDPs and IDP regions?

- What are the functions of IDPs and IDP regions?

Why don't IDPs fold into 3D structure?

- ***Amino acid composition*** determines whether a protein will fold or remain unfolded.

- For compositions that favor structure, the sequence patterns of hydrophobic / hydrophilic groups determine which ***3D structure*** is formed.

Shakhnovich, E.I. and Gutin, A.M. Engineering of stable and fast-folding sequences of model proteins. *Proc. Natl. Acad. Sci. USA* 90: 7195 – 7199 (1993).

Why don't IDPs fold into 3D structure?

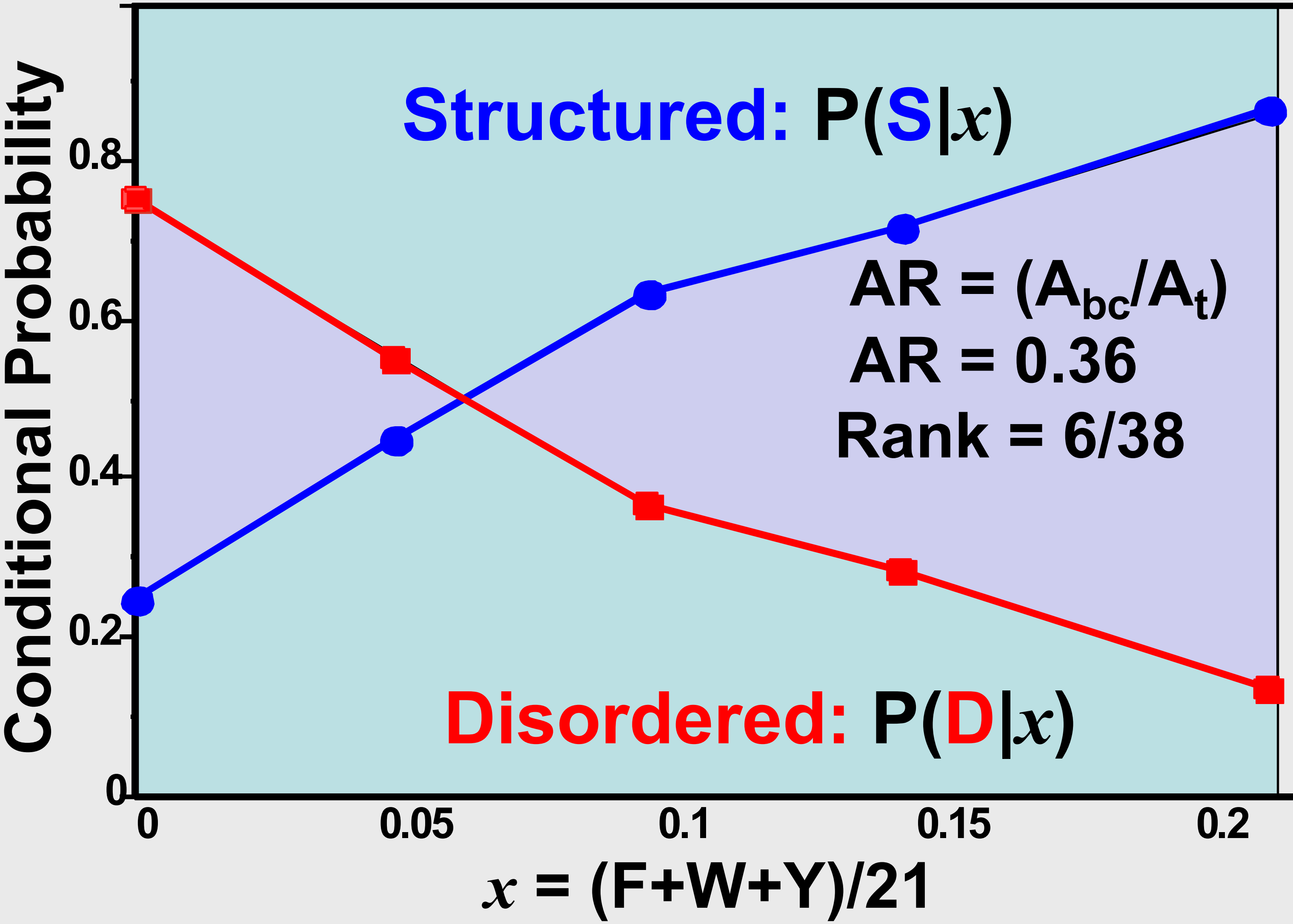
Xie et al., *Genome Informatics* 9: 193-200 (1998)

Qian Xie

Ethan Garner

Zoran Obradovic

Pedro Romero



Why don't IDPs fold into 3D structure? Amino acid sequence favors nonfolding!

- IDPs have too few aromatics – aromatics are important for the stability of hydrophobic cores;
- IDP ratio of hydrophilic amino acids to hydrophobic amino acids is too high for folding;
- IDPs have too low of a sequence complexity
- IDPs have too large of a ***net*** charge – charge repulsion inhibits folding;
- IDPs have too many prolines – prolines cannot form backbone H–bond, so helices and sheets are destabilized by prolines.

Intrinsically Disordered Proteins (IDPs)

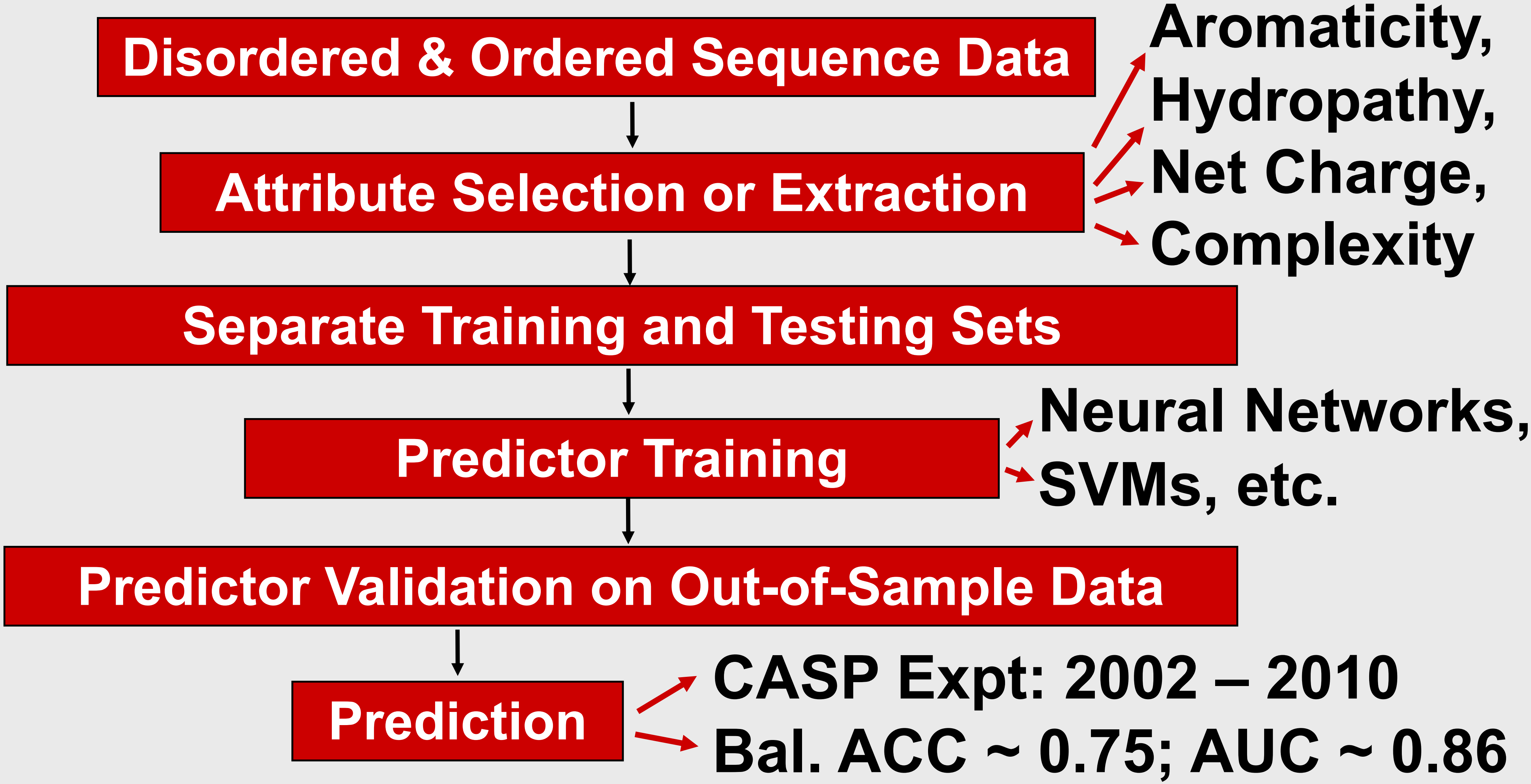
How common are IDPs and IDP regions?

Step 1: Develop predictor of IDPs and IDP regions.

Step 2: Apply to multiple proteomes.

Dunker *et al., Genome Informatics* 11: 161-171 (2000)
(repeated by many others, and by us)

Step1: Predictor Intrinsic Disorder



Disordered & Ordered Sequence Data

Attribute Selection or Extraction

Aromaticity,
Hydropathy,
Net Charge,
Complexity

Separate Training and Testing Sets

Predictor Training

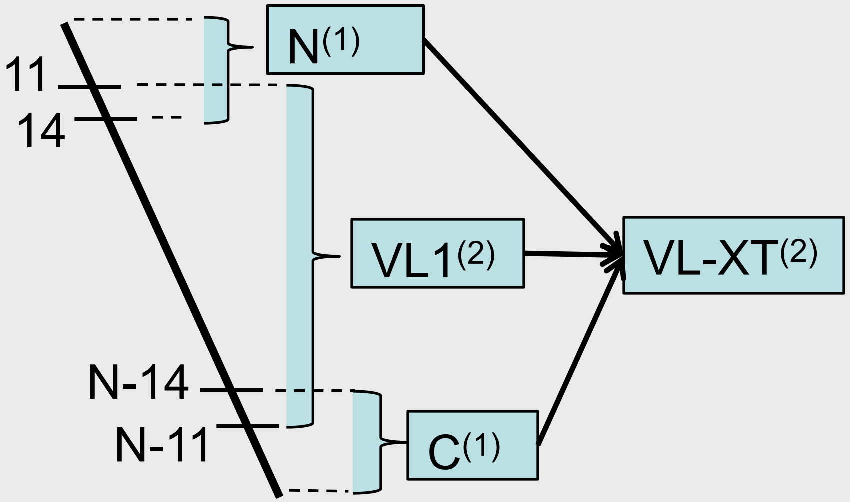
Neural Networks,
SVMs, etc.

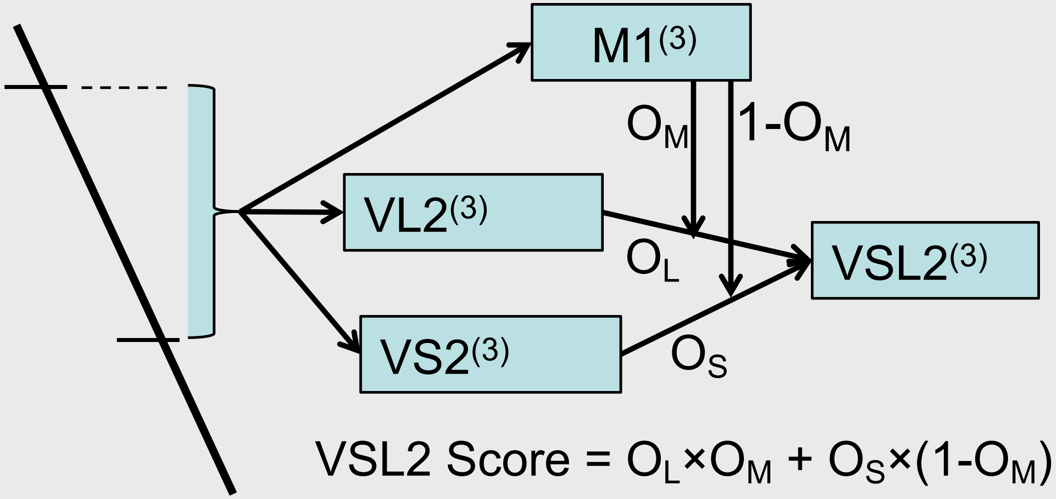
Predictor Validation on Out-of-Sample Data

Prediction

CASP Expt: 2002 – 2010
Bal. ACC ~ 0.75; AUC ~ 0.86

Predictors of Natural Disordered Regions PONDR®VL-XT and PONDR®VSL2





VSL2 Score = $O_L \times O_M + O_S \times (1-O_M)$

N, VL1, and C are neural networks

N-term: 8 inputs
VL1: 10 inputs
C-term: 8 inputs

M1, VSL2-L, and VSL2-S are support vector machines

M1: 54 inputs
VL2: 20 inputs
VS2: 20 inputs

[1] **Li X et al., *Genome Informat*. 9:201-213 (1999)**

[2] **Romero P et al., *Proteins* 42:38-48 (2001)**

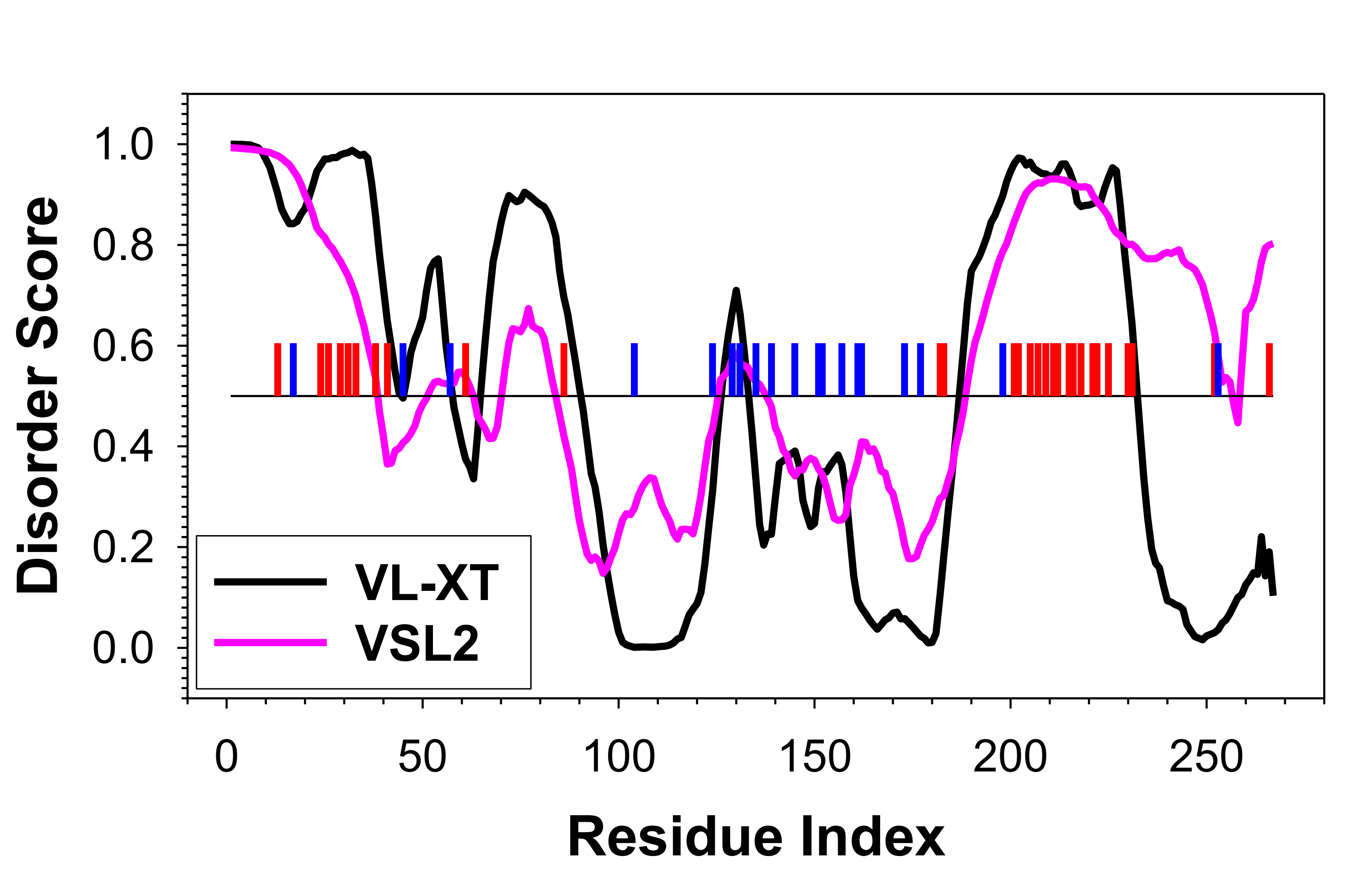
[3] **Peng K et al., *BMC Bioinfo.* 7:208 (2006)**

PONDR®VL-XT and PONDR®VSL2

(+) Disordered

XPA

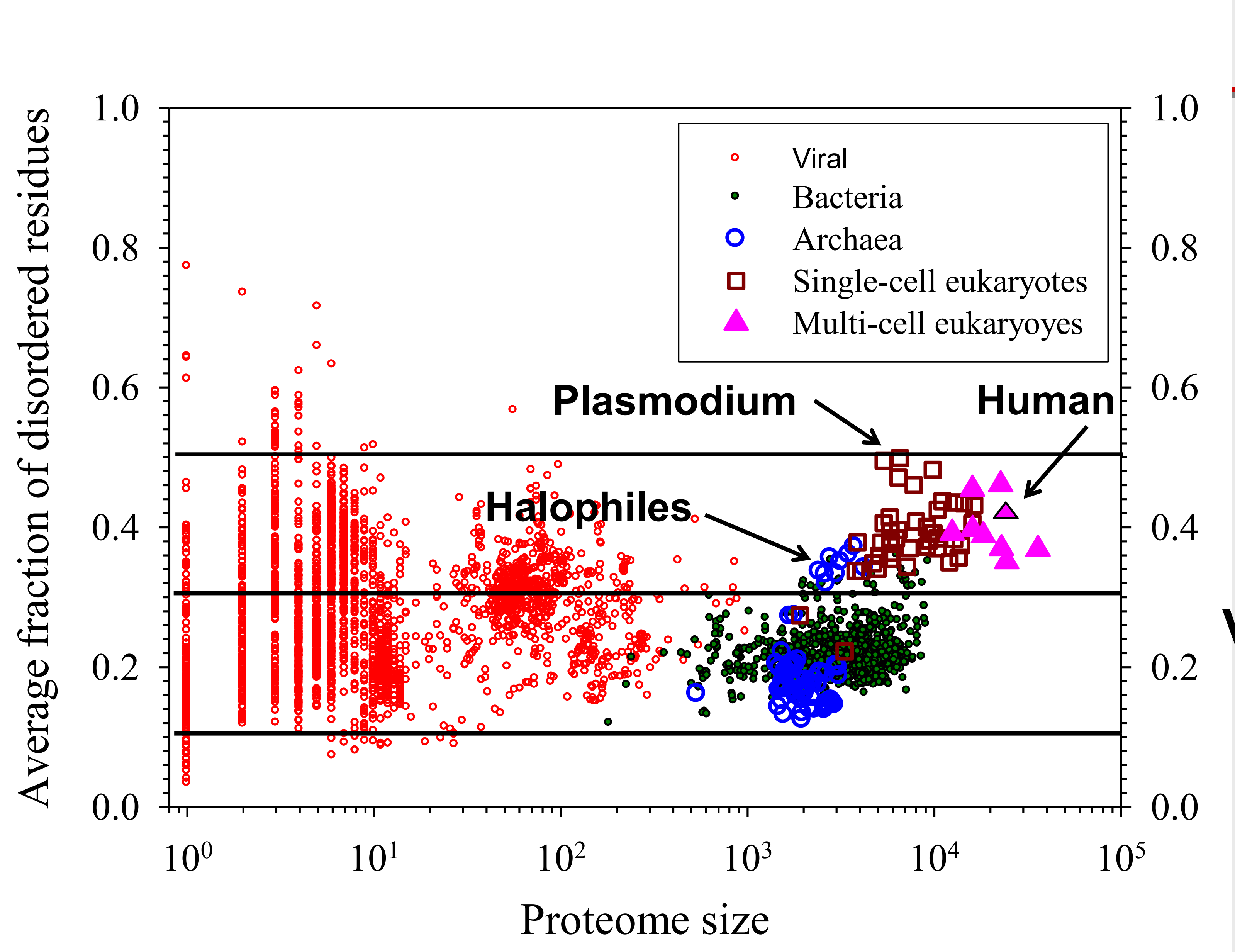
(–) Structured



Iakoucheva L et al., *Protein Sci* 3: 561-571 (2001)

Dunker AK et al., *FEBS J* 272: 5129-5148 (2005)

Step 2: How common are IDPs?



Bin Xue

Vladimir Uversky

Xue et al., *J Biomol Struct Dyn* 30: 137-149 (2012)

How common are IDPs?
More recent, improved approach

Combine structure / disorder prediction and structure prediction *by sequence similarity* to all currently known protein 3 D structures.

For the human proteome:

Fukuchi, S., *et al.,* Binary classification of protein molecules into intrinsically disordered and ordered segments. BMC Struct Biol. 11:29 (2011); For Human: 35% residues are in IDPs or IDP regions. (Weakness → used Pfam for structured proteins)

For 1,765 proteomes (8 different order / disorder predictors):

Oates, M.E. *et al.,* D^2P^2: database of disordered protein predictions. *Nucleic Acids Res*. 41(Database issue):D508-516 (2013). For Human: 35% - 50% residues in IDPs or IDP regions.

(Strength → used SUPERFAMILY for structured proteins)

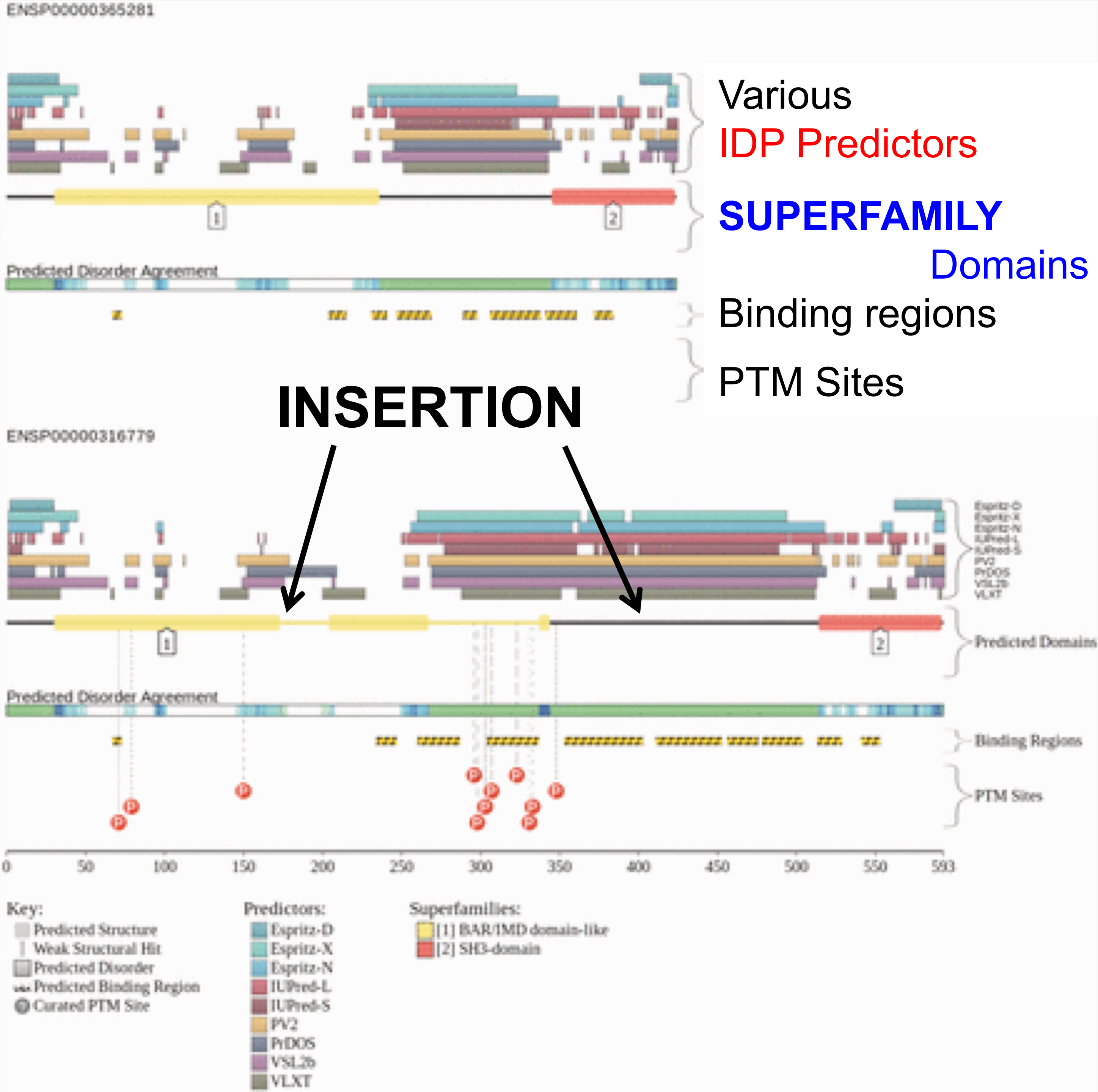
Human BIN1 from D^{2}P^2

Two transcripts from one gene;

Matt Oates

Julian Gough

Insertion from alternative splicing.



ENSP00000365281

Predicted Disorder Agreement

Various IDP Predictors

SUPERFAMILY Domains

Binding regions

PTM Sites

INSERTION

ENSP00000316779

Predicted Disorder Agreement

Key:
Predicted Structure
Weak Structural Hit
Predicted Disorder
Predicted Binding Region
Curated PTM Site

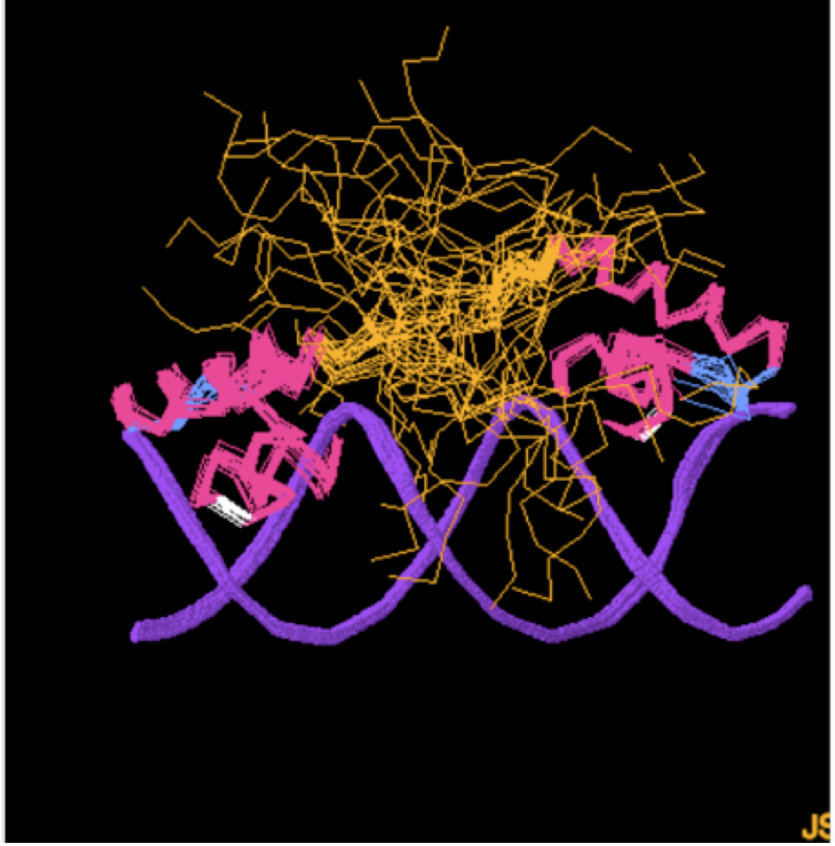
Predictors:
Espritz-D
Espritz-X
Espritz-N
IUPred-L
IUPred-S
PV2
PrDOS
VSL2b
VLXT

Superfamilies:
[1] BAR/IMD domain-like
[2] SH3-domain

Oates M *et al., NAR* 41: D508-516 (2013)

IDP Functions: Lac Repressor

Kalodimos *et al., Science* 305:386-389 (2004)

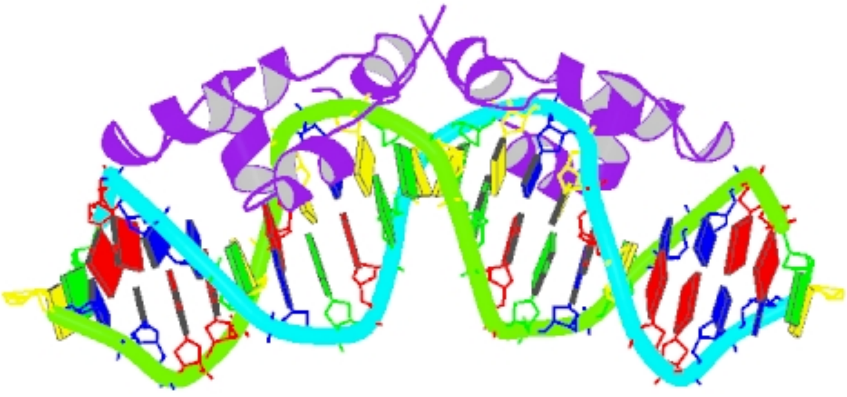


- Upon binding random DNA, a 12 residue linker remains disordered & binds DNA phosphates transiently, helping the Lac Repressor slide along the DNA.

- Upon encountering its binding sequence, the IDP region → structure and is involved in recognizing the cognate DNA binding sequence, in increasing the affinity, &

  in helping bend the DNA.

The picture can't be displayed.



Images: Proteopedia, Life in 3D, the free, collaborative, 3D Encyclopedia:

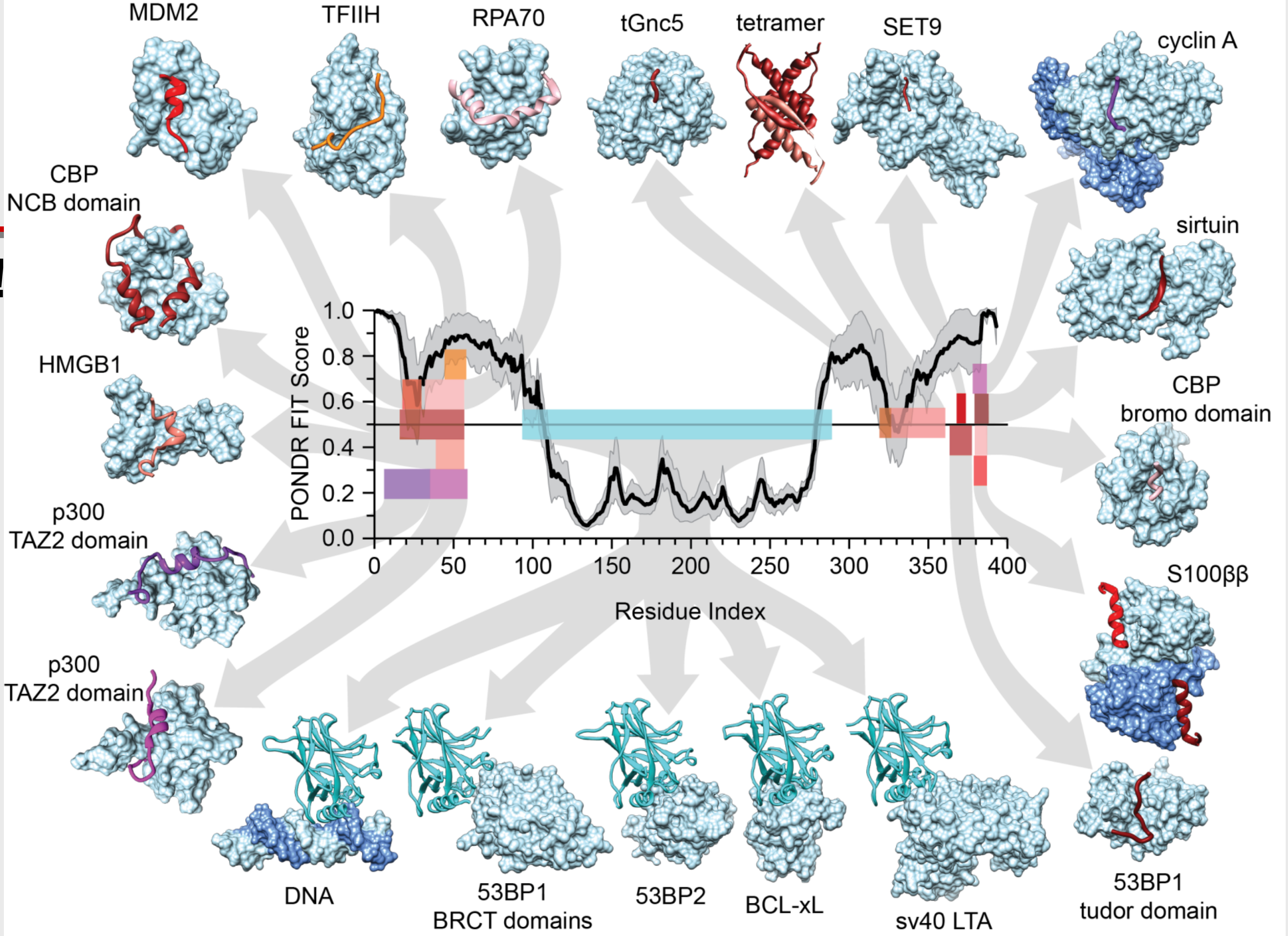
– provided by: Joel Sussman

p53 binding

Note IDP tails!

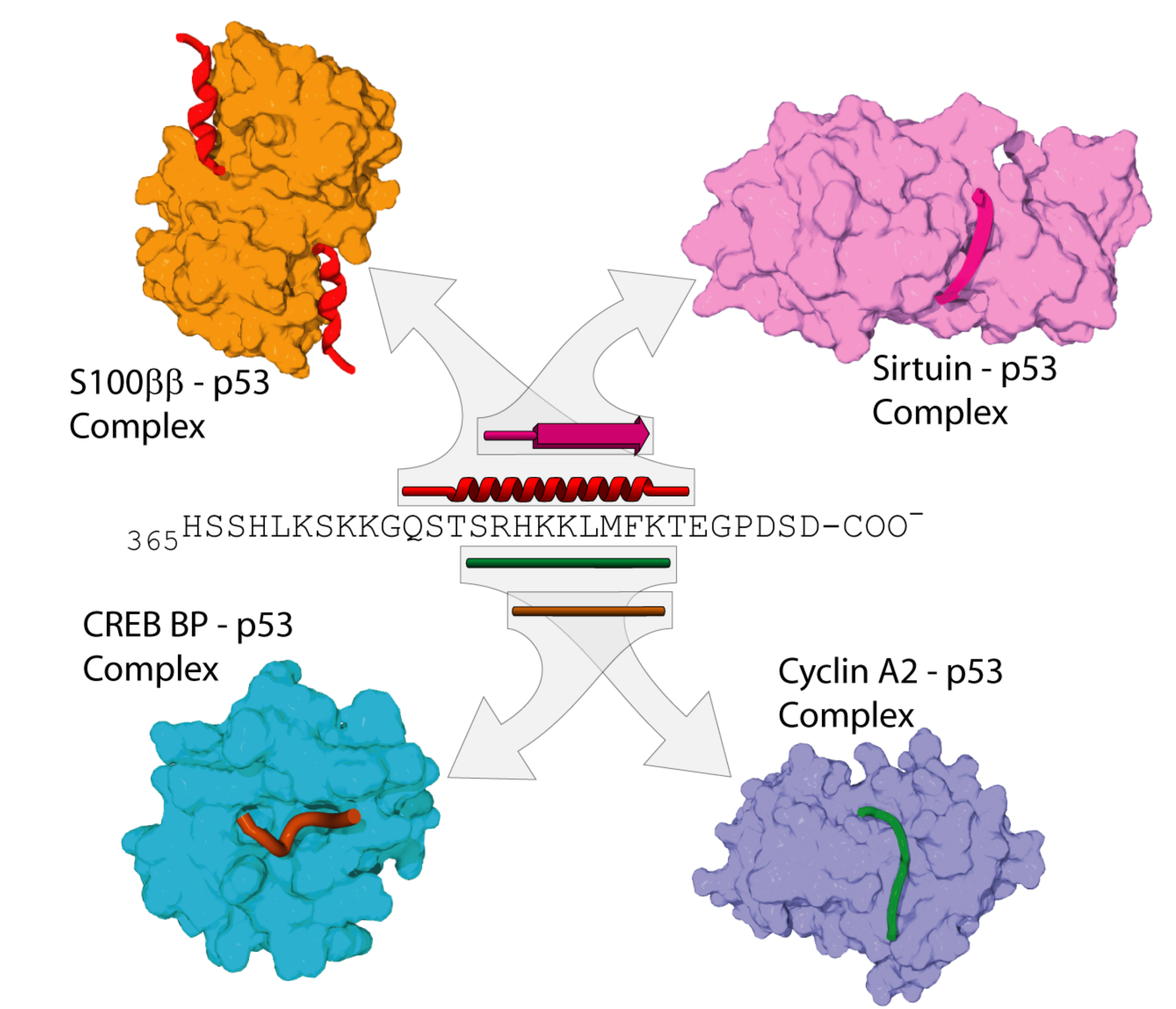
Molecular Recognition Features (MoRFs)

Chris Oldfield



Modified from: Oldfield & Dunker, *Ann Rev Biochem* 83: 553 – 584 (2014)

p53 C-terminal Domain: Secondary Structure and Overlap

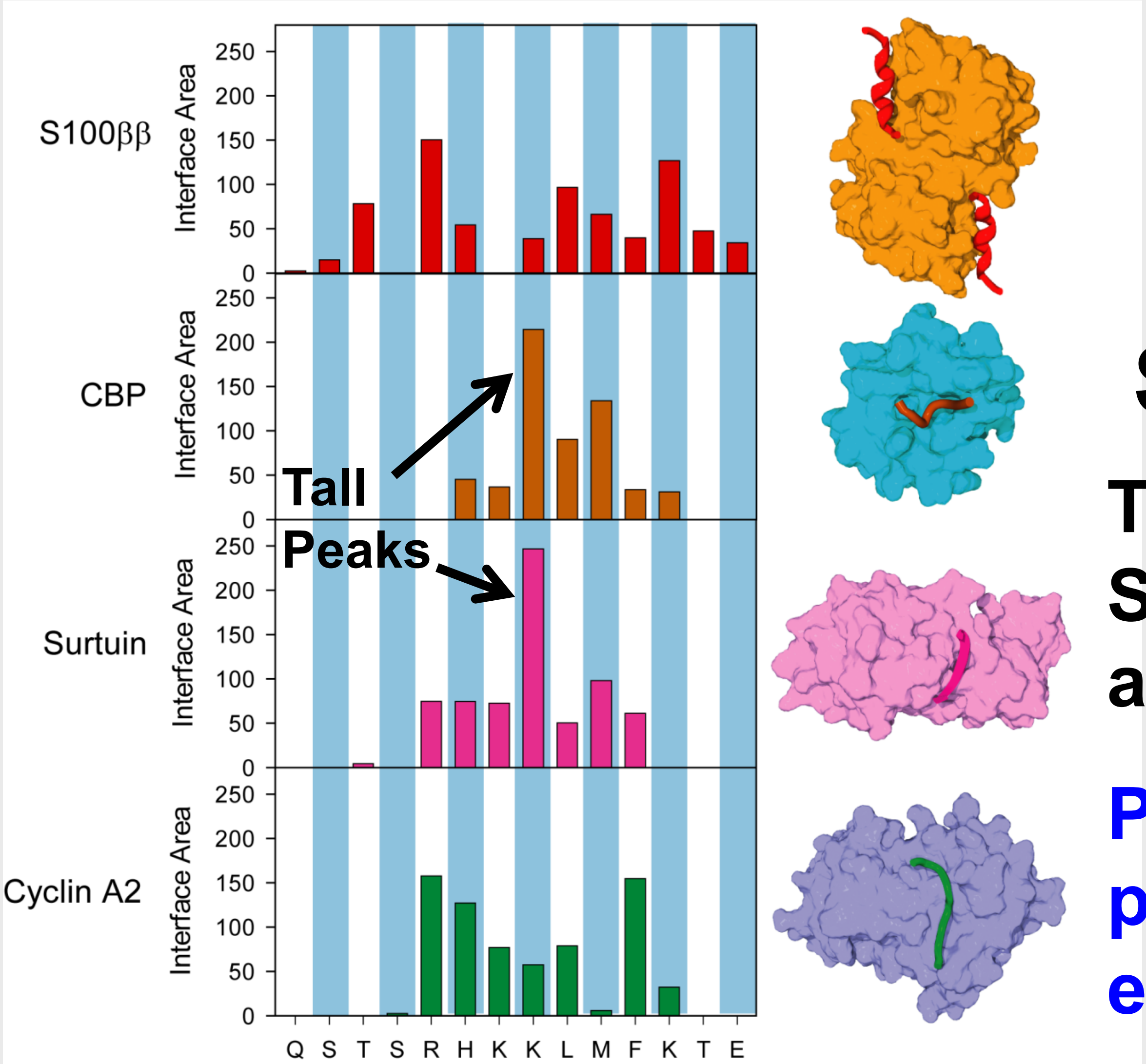


Oldfield CJ, et al., *BMC Genomics* 9 (Suppl 1) S1 (2008)

Confirms Landsteiner-Pauling 1939 - 1940 hypothesis of changes in structure due to folding upon binding to different partners!!

Pauling L, J Am Chem Soc 62: 2643-2657 (1940)

p53 C-terminal IDP region: Residue-specific Interface Area



Tall Peaks

Oldfield CJ, et al., *BMC Genomics* 9 (Suppl 1) S1 (2008)

STSRHKKLMFKE

Tall peaks in CBP and Sirtuin; due to buried acetyl group.

PTMs often contribute to partner switching. Many examples observed.

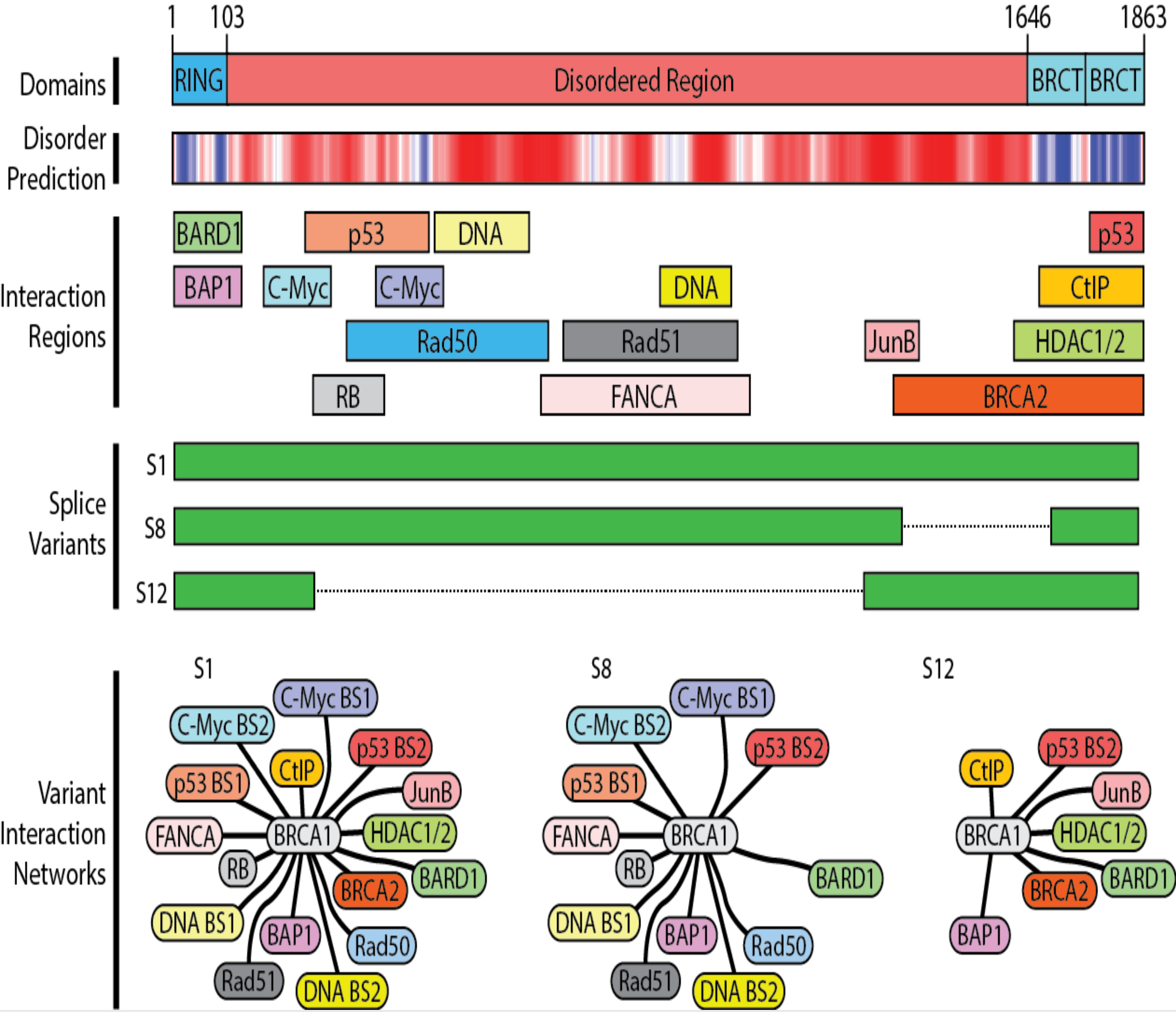
BRCA1

1863 residues;

103 ordered at N-term;

217 ordered at C-term;

1543 disordered in between.

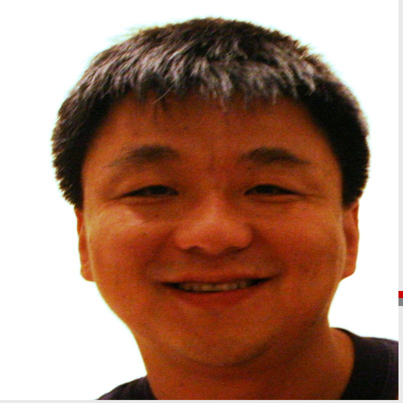


The IDP–AS–PTM Toolkit Hypothesis

IDP, AS, & PTM ***all shown to enable signalling complexity***:

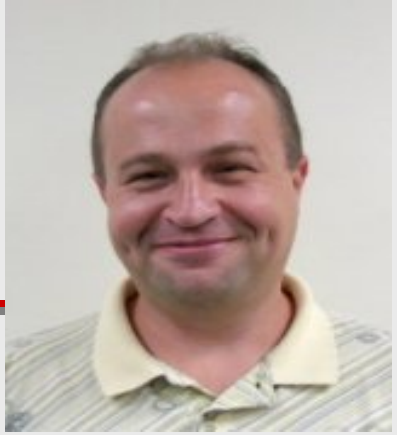
- IDPs change shape & thereby bind to multiple partners.
- PTMs within IDP regions bring about partner switching.
- AS of mRNA coding for IDP regions rewires protein-protein & protein-DNA interaction networks – often tissue-specific!

Hypothesis: IDP, AS, PTMs are colocalized & thus collaborate to further increase signaling complexity.



Hongbao Xie

IDPs & Function Global Analysis



Zoran Obradovic

- Collect SwissProt function-specific sequences;
- Collect 1,000 matching, random-function sequences; Matching = same size, same # chains.
- Predict disorder for each function-specific & 1,000 random-function sets → all RFS sets ~ Gaussian;
- Rank structure- and disorder-associated functions by Z-scores (Z-score = [x – <x>]/σ); Set <x> = 0.

– values = more structure, + values = more disorder

IDPs & Function

Functional Key Word Categories	Number
High-prediction of disorder (> +1)	238
Intermediate (Z-score, –1 to +1)	170
Low-prediction of disorder (< –1)	302
TOTAL	710

Xie H et al. *J. Proteome Res.*. 6: 1882- 1898;
6:1899-1916; & 6:1917-1932 (2007)

Top 10 Biological Processes Most Strongly Associated with Low-prediction of Disorder (e.g. with Structure)

KEYWORDS	Proteins (number)	Families (number)	Length (Ave)	Z – Score
GMP Biosynthesis	225	3	473	–17.6
Amino-acid Biosynthesis	7098	212	361	–17.1
Transport	19888	2199	378	–14.9
Electron Transport	4633	346	272	–13.7
Lipid A Biosynthesis	533	13	291	–13.2
Aromatic Catabolism	320	105	300	–12.4
Glycolysis	2255	50	390	–12.1
Purine Biosynthesis	1208	28	445	–11.9
Pyrimidine Biosynthesis	1310	27	383	–11.7
Carbohydrate Metabolism	1797	180	404	–11.7

Xie H, et al., *J. Proteome Res* 6: 1882-1932 (2007)

Top 10 Biological Processes Most Strongly Associated with High-Prediction of Disorder

KEYWORDS	Proteins (number)	Families (number)	Length (Ave)	Z – Score
Differentiation	1406	422	439	18.8
Transcription	11223	1653	442	14.6
Transcription Regulation	9758	1554	413	14.3
Spermatogenesis	332	189	280	13.9
DNA Condensation	317	130	300	13.3
Cell Cycle	4278	612	494	12.2
mRNA Processing	1575	249	516	10.9
mRNA Splicing	716	180	459	10.1
Mitosis	718	215	620	9.4
Apoptosis	810	211	465	9.4

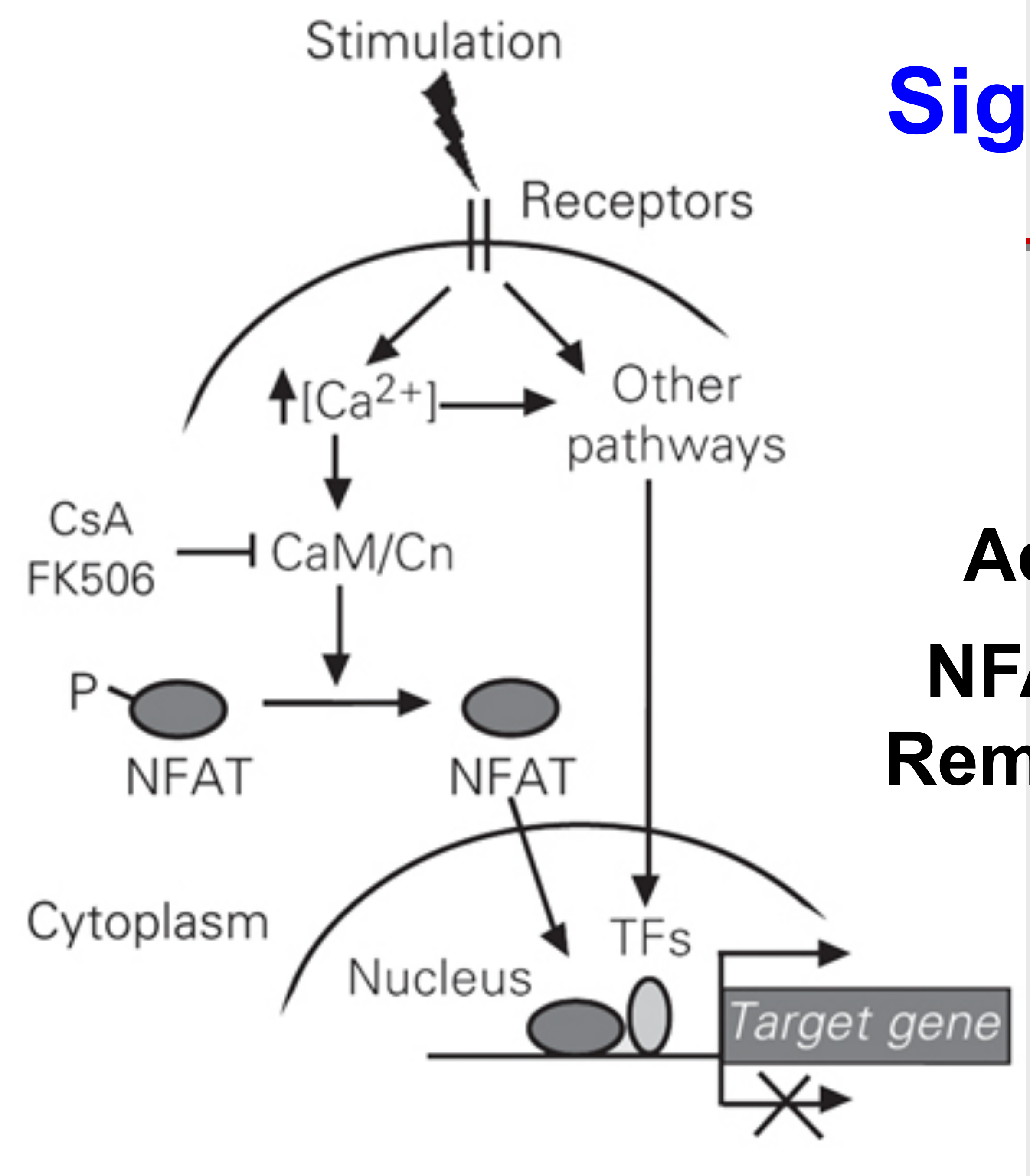
Xie H, et al., *J. Proteome Res* 6: 1882-1932 (2007)

Functions of Structured Proteins vs. IDPs

- Sequence → Structure → Function ($Z < -1$)
  - Catalysis,
  - Membrane transport,
  - Binding with DNA, RNA, Proteins, IDPs & molecules
- Sequence → IDP Ensemble → Function ($Z > +1$)
  - Signaling, Dunker AK, et al., *Biochemistry* 41: 6573-6582 (2002)
  - Regulation, Dunker AK, et al., *Adv. Prot. Chem*. 62: 25-49 (2002)
  - Recognition, Xie H, et al., *Proteome Res.* 6: 1882-1898 (2007)
  - Control. Vucetic, S. et al., *Proteome Res* 6: 1899-1916 (2007)

Xie H, et al., *Proteome Res* 6: 1917-1932 (2007)

Signaling Pathway



Calmodulin (CaM)
Calcineurin (Cn)
Nuclear Factor of Activated T- Cells (NFAT)

NFAT-poly-P in an IDP tail.
Remove Ps, activates NLS
→ NFAT → nucleus
→ turns on genes
→ T-cells activated
→ reject transplant

Nuclear Factor of Activated T-cells (NFAT) Transcription Factor (TF) Family

NFAT: Phosphorylation → Inactivation

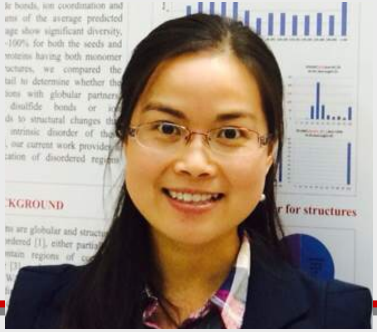
Ca^{2+}/CaM → CaN Activation

Plays key roles in the following biological processes:

- T-cell Activation
- Myocardial development
- Cancer metastasis
- And many more
- Angiogenesis
- Skeletal muscle development

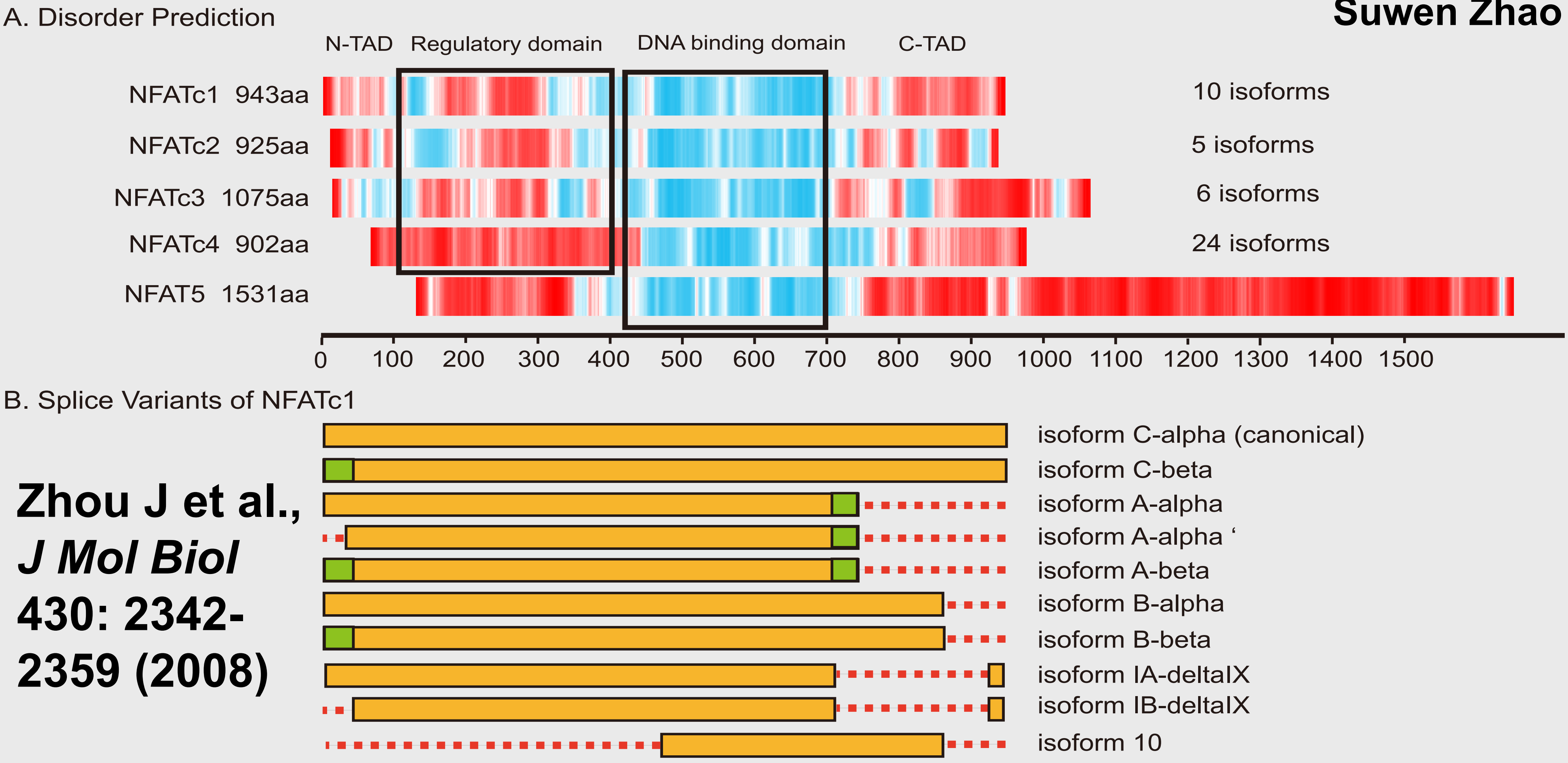
Pan MG et al., *Curr Mol Med* 13:543-554 (2013).

NFAT Family of TFs`



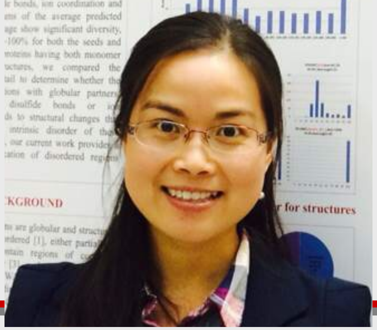
Jianhong Zhou

Suwen Zhao



Zhou J et al.,
J Mol Biol
430: 2342-2359 (2008)

NFAT Family of TFs`

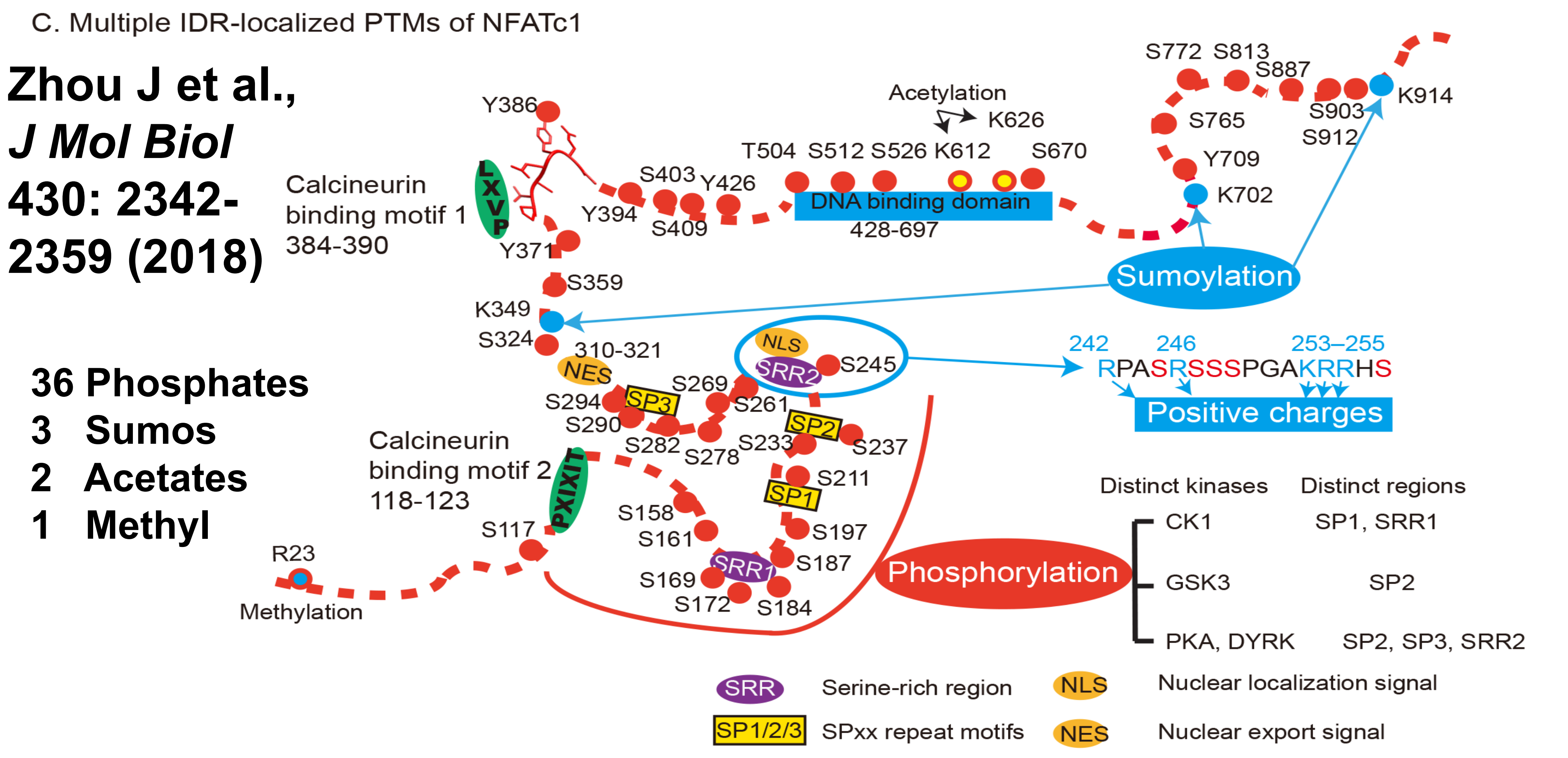


Jianhong Zhou

Suwen Zhao

Zhou J et al., *J Mol Biol* 430: 2342-2359 (2018)

36 Phosphates
3 Sumos
2 Acetates
1 Methyl



C. Multiple IDR-localized PTMs of NFATc1

Distinct kinases	Distinct regions
CK1	SP1, SRR1
GSK3	SP2
PKA, DYRK	SP2, SP3, SRR2

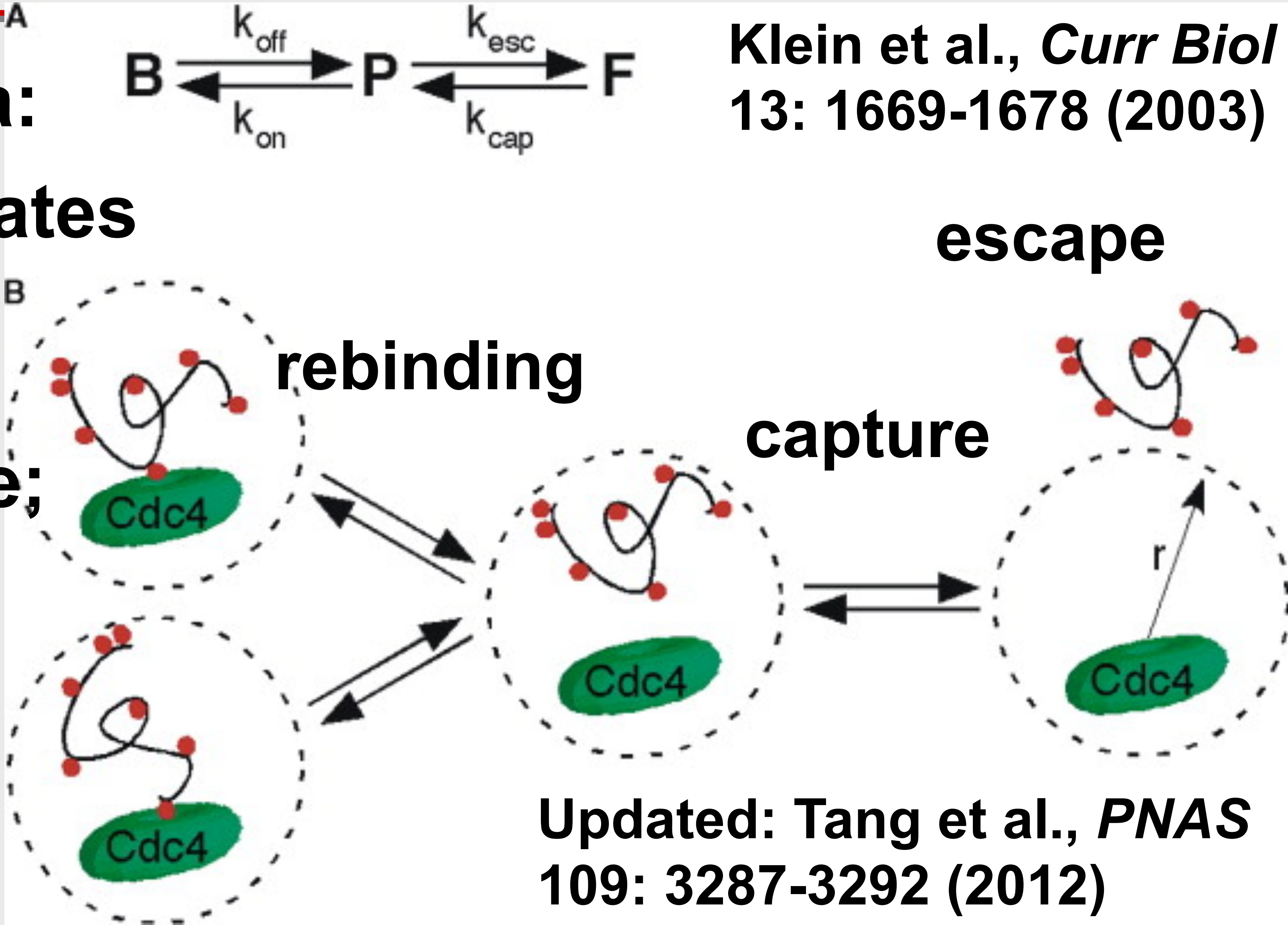
SRR	Serine-rich region	NLS	Nuclear localization signal
SP1/2/3	SPxx repeat motifs	NES	Nuclear export signal

Cdc4-Sic1 & NFAT-NLS: ON-OFF Switches From Multiple Phosphates in IDP Regions

Overall Idea:

If # phosphates under threshold, then escape;

If over threshold, then rebinding.



rebinding

capture

escape

Klein et al., *Curr Biol* 13: 1669-1678 (2003)

Updated: Tang et al., *PNAS* 109: 3287-3292 (2012)

Many Proteins have PTM Clusters

Proteins	Suggested Concept	Reference
Histones	Histone Code	Strahl BD & Allis CD Nature 403:41-45 (2000)
p53, tubulin, Cdc25c, RNAP II	Molecular Barcode	Yang JX Oncogene 24: 1653-1662 (2005)
Transcription Factors	PTM Code	Benavoun BA & Veitia, Trends Cell Biol 19:189-197 (2009)
Various	Combinatorial PTMs	Lothrop AP et al., FEBS Lett 587:1247-1257 (2013)
P300 / CBP	Coactivator Code	Gamble MJ & Freedman LP, TIBS 27:165-167 (2002)
RNAP II CTD	Hyper-/Hypo-phosphorylation	Xu YX et al., Genes & Dev 17: 2765-2776 (2003)
Forkhead Box	FoxO Code	Calnan DR & Brunet A, Oncogene 27: 2276-2288 (2008)
p53	Cooperative Integrators	Meek DW & Anderson CW CSH Perspect Biol 1: a000950 (2009)

Cited in Pejaver et al., Protein Sci 23: 1077-1093 (2014)

PTM Clusters → PTM Codes

- For **PTM clusters** in Histones, p53, tubulin, Cdc25c, FoxO, RNAP II CTD, etc., the concept is that **different PTM patterns** lead to **different signaling consequences**.
- Thus, a **Histone** or **PTM code** likely exits.
- Predictions & experiments show that these **PTM clusters** are located in **IDP regions**.
- Bioinformatics extensions suggest that **PTM clusters** in **IDP regions** are very common.
- Thus, **PTM codes** are almost certainly very widely used for modulating **cell signaling**.

Pedja Radivojac

Vikas Pejaver

Pejaver et al., Protein Sci 23: 1077-1093 (2014)

PTM Codes: Located in IDP Regions Modulated by AS, Thus IDP-AS-PTM

Proteins	Code Name	HITS	+AS
Histones	*Histone Code*	44,260	184
CREB BP	*Coactivator Code*	10,305	89
RNA Polymerase II	*Hyper/Hypo Phos.*	30,008	615
p53	*Molec. Barcode*	87,167	532
Tubulin		29,998	92
Forkhead Box	*FOXO Code*	2,357	6
Forkhead Box 1	*PTM Code*	3,372	8
Forkhead Box 4		336	4

References for PTM Codes:
New Idea: PTM Codes Modulated by AS

Histone Code – Strahl BD & Allis CD. *Nature* 403:41-45 (2000)

Coactivator Code – Gamble MJ & Freedman LP: TIBS 27:165-167 (2002)

Hyper/Hypo Phos – Xu YX et al., *Genes Dev* 17:2765-2776 (2003)

Molecular Barcode – Yang XJ *Oncogene* 24:1653-1662 (2005)

FOXO Code – Calnan DR & Brunet A: *Oncogene* 27:2276-2288 (2008)

PTM Code – Benayoun BA & Veitia RA: *Trends Cell Biol* 19:189-197 (2009).

The IDP–AS–PTM Toolkit Hypothesis

IDP, AS, & PTM *shown to collaborate to yield complex signaling* for the following proteins:

- NFAT family – transcription factors
- GPCR family – membrane signaling proteins
- Sarc Kinase family – signaling enzymes

Many proteins associated with cancer, cellular differentiation, conversion to stem cells, and so on all contain IDPs, AS, and PTMs, suggesting that this toolkit perhaps used by all these proteins. *Have not yet shown their co-localization and collaboration – for the future.*

Zhou J et al., J Mol Biol 430: 2342-2359 *(2018)*

Key Functions for the Evolution of Complex Multicellular Organisms

Complex multicellular organisms require the following:

- Cell adhesion;
- Communication between cells;
- Developmental programs;

Nicklas & Newman
Evol Devel Biol
15: 41-52 (2013)

- Regulation of the developmental programs;
- Cell-specific biochemistry.

IDPs, AS, & PTMs common (universal?) among proteins that are involved in all of these functions!!

Dunker *AK et al.*, *Semin Cell Devel Biol 37: 44-55* (2015)

IDPs and Gene Regulation

Bin Xue

Shinya Yamanaka (2012 Nobel Prize)
Overexpress 4 transcription factors (TFs):
All 4 of these TFs very rich in predicted IDP AAs:
Sox2 (100%), Oct4 (67%,), Klf4 (97%), c-Myc (80%)
Adult fibroblast cells → induced Pluripotent Stem Cells (iPSCs)

The key TFs identified by >10 years of trial and error from a large number of additional TFs. Many TFs help with transdifferentiation by improving efficiency. Most of these TFs are rich in predicted disorder.

Xue B *et al., Mol BioSys* 8:134-150 (2012)

IDPs and Gene Regulation

Julian Gough

Morgrify: An Algorithm (http://morgrify.net)

Input: gene expression data for different cell types & known regulatory networks; data for 173 cell types, 134 tissues

Output: Atlas of transcription factor sets: (any cell type A) → (any cell type B)

Results: Predicts TF sets for 5 known transdifferentiations
Predicts TF sets for 2 new transdifferentiations
Experiments worked on first try in both cases!!

Rackham OJL *et al.*, *Nature Genetics* 48: 331-335 (2016)

(seminar link: https://www.dropbox.com/s/5rf7s4cfkzrlwu9/CSHL-Asia_2018.pptx?dl=0)

IDPs and Gene Regulation

Rackham OJL *et al., Nature Genetics* 48: 331-335 (2016)
Kamaraj US *et al.,* Cell Cycle 15: 3343-3354 (2016)

Julian Gough

ESC – Embryonic Stem Cell
MSC – Mesanchymal Stem Cells

Previously known Transformations

1. Fibroblasts → Myoblasts (1998)
2. B-cells → Macrophages (2004)
3. Fibroblasts → iPSCs (2007)
4. Fibroblasts → Hepatocytes (2011)
5. Fibroblasts → Heart (2013)

Predicted & Confirmed

1. Fibroblasts → Keratinocytes
2. Keratinocytes → epithelial cells

Since 2016

1. ESC → Endothelial Cell
2. iPSC → Endothelial Cell
3. Fibroblast → Endothelial
4. Fibroblast → Astrocyte
5. ESC → Astrocyte
6. iPSC → Astrocyte
7. MSC → Astrocyte
8. ESC → Keratinocyte
9. iPSC → Keratinocyte

+ 2 more, All of first try!

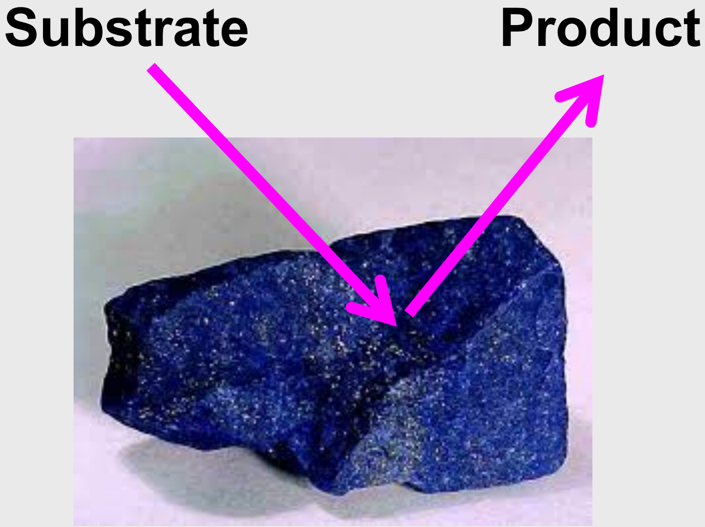
Summary

- Sequence → Structure → Function
  - Catalysis,
  - Membrane transport,
  - Binding with DNA, RNA, Proteins, IDPs & molecules
- Sequence → IDP Ensemble → Function
  - Signaling, Dunker AK, et al., *Biochemistry* 41: 6573-6582 (2002)
  - Regulation, Dunker AK, et al., *Adv. Prot. Chem*. 62: 25-49 (2002)
  - Recognition, Xie H, et al., *Proteome Res.* 6: 1882-1898 (2007)
  - Control. Vucetic, S. et al., *Proteome Res* 6: 1899-1916 (2007)

    Xie H, et al., *Proteome Res* 6: 1917-1932 (2007)

A STRUCTURE-BASED Toolkit

→



A rock-like **structured protein**

Lock and key,
induced fit;

many proteins,

many functions

The IDP-AS-PTM Developmental Toolkit

→

An IDP or IDP Region,

+ PTMs

+ AS

One IDP, many shapes, many functions, provides a toolkit for complex signaling & cellular differentiation

Intrinsically Disordered Proteins

THANK YOU!!!

(kedunker@iupui.edu)

Funding: NIH, NSF, INGEN,
IUPUI Signature Centers Initiative