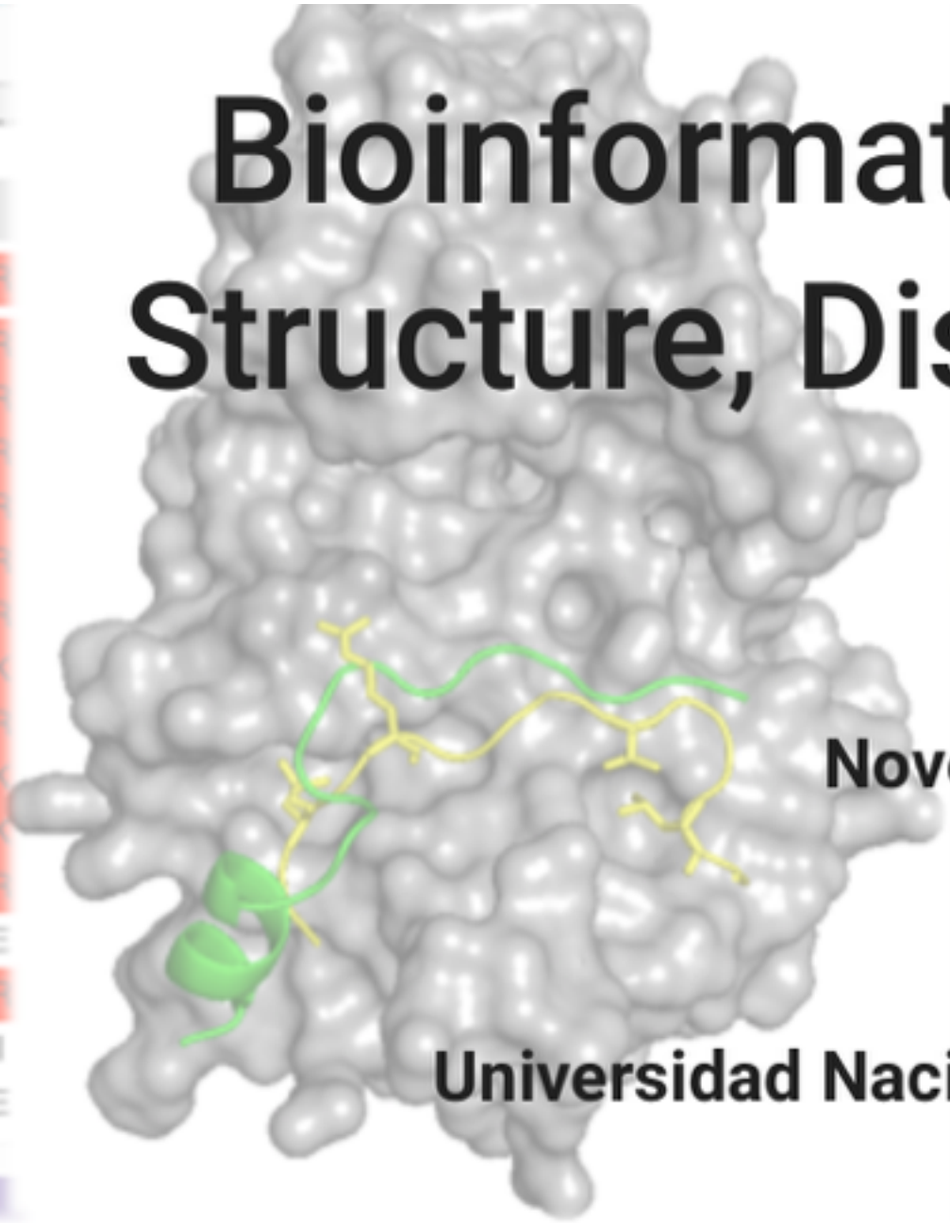
Bioinformatics Tools for Protein Structure, Disorder and Interaction Analysis



November 21st-23rd, 2017

9:00 - 18:00 hrs

Universidad Nacional de San Martín and IIB-INTECH

DAAD

Deutscher Akademischer Austausch Dienst
German Academic Exchange Service

EMBL

MinCyT-DAAD
IIB UNSAM

CONICET

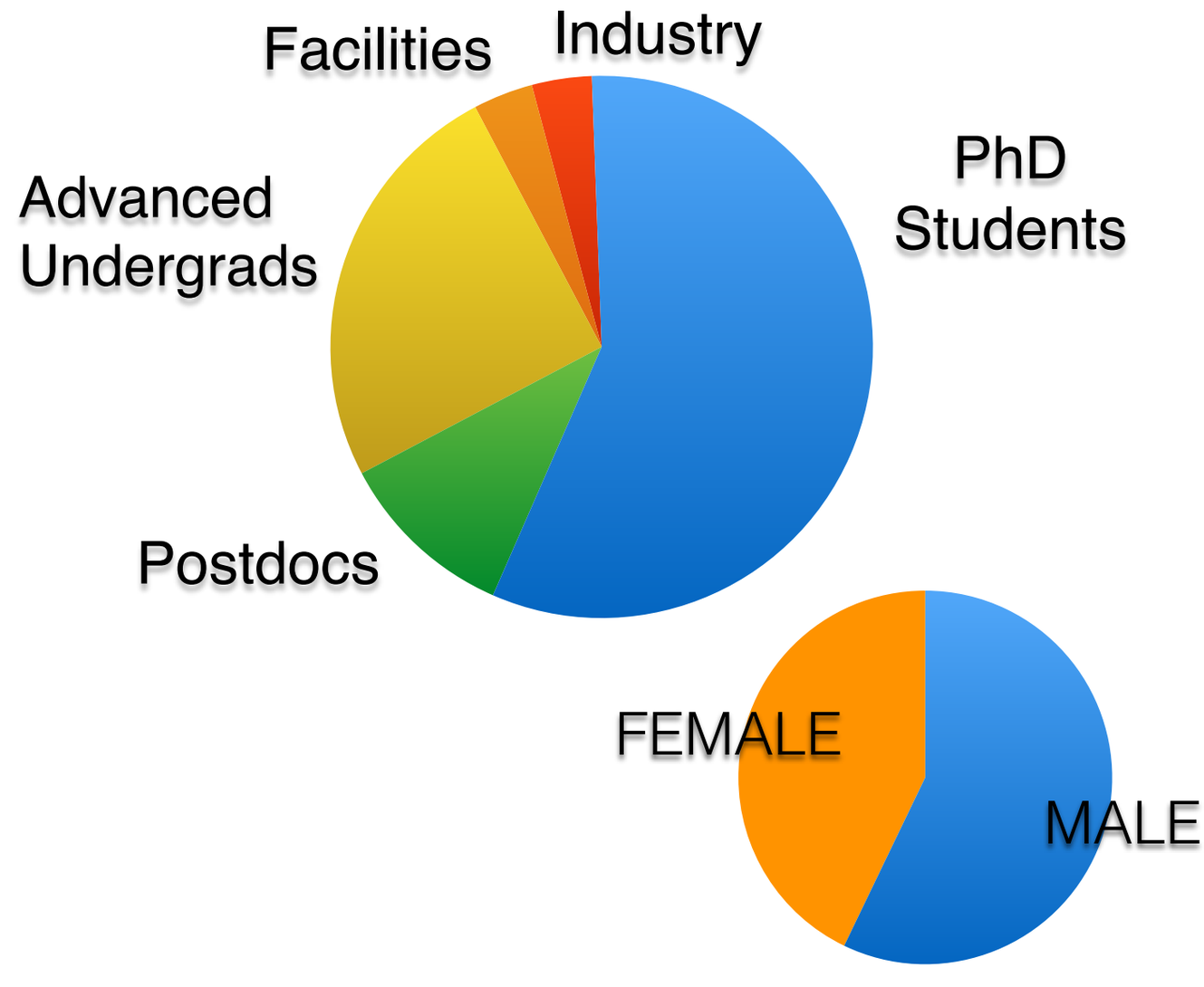
UNIVERSIDAD NACIONAL DE SAN MARTÍN

I I B - I N T E C H

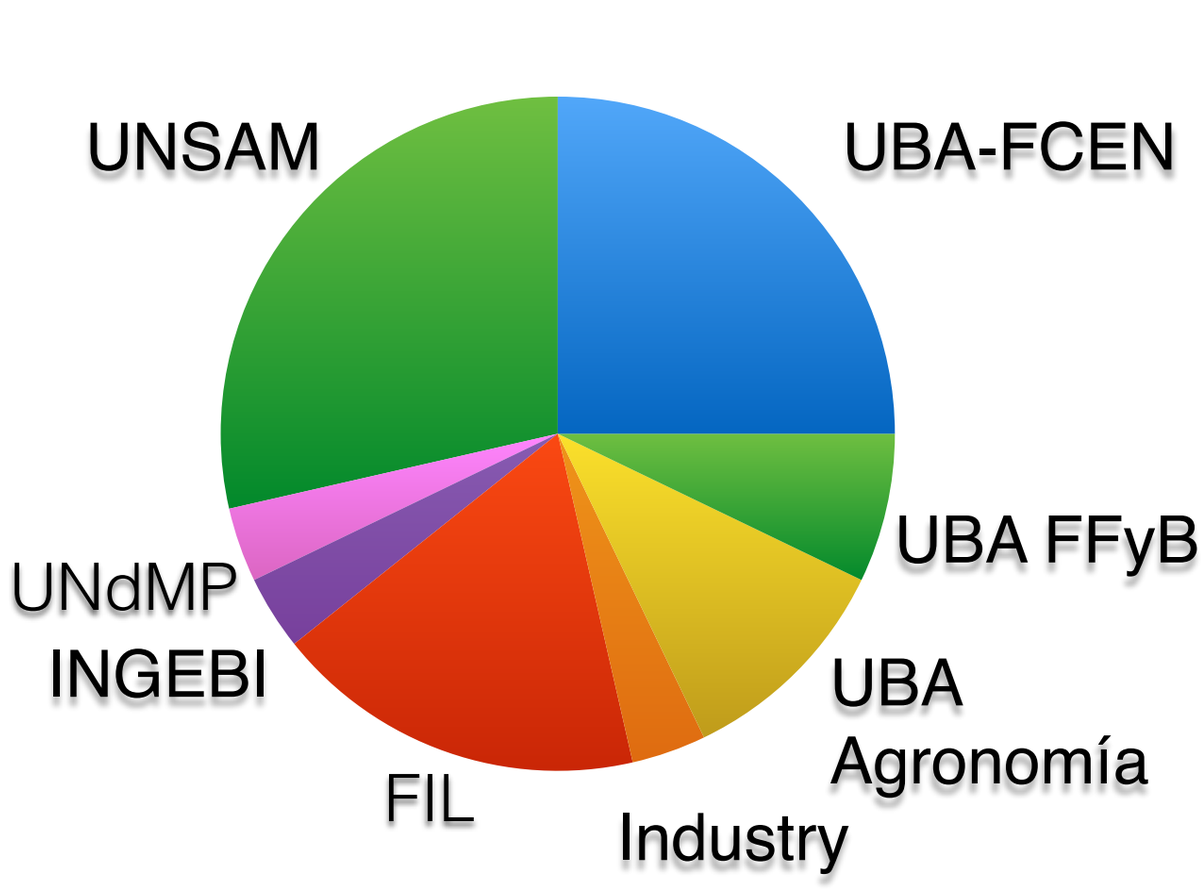
Ministerio de Ciencia, Tecnología e Innovación Productiva
Presidencia de la Nación

Bioinformatics Tools for Protein Structure, Disorder and Interaction Analysis

Who is taking the course?



Where are you coming from?



Bioinformatics Tools for Protein Structure, Disorder and Interaction Analysis

What organisms do you work on?

Cancer biology	*mammalian cells*
DNA Damage	
Yeast genetics/signaling	*S. cerevisiae, S. Pombe*
Plant physiology	*A. thaliana, others*
Bacterial Pathogens	*P. aeruginosa, B. abortus*
Viral Pathogens	*HRSV, Adenovirus*
Eukaryotic Pathogens	*T. cruzi, Leishmania*

What is your motivation for taking the course?

Dynamic protein assemblies

Intrinsically Disordered regions

Protein-protein interaction networks

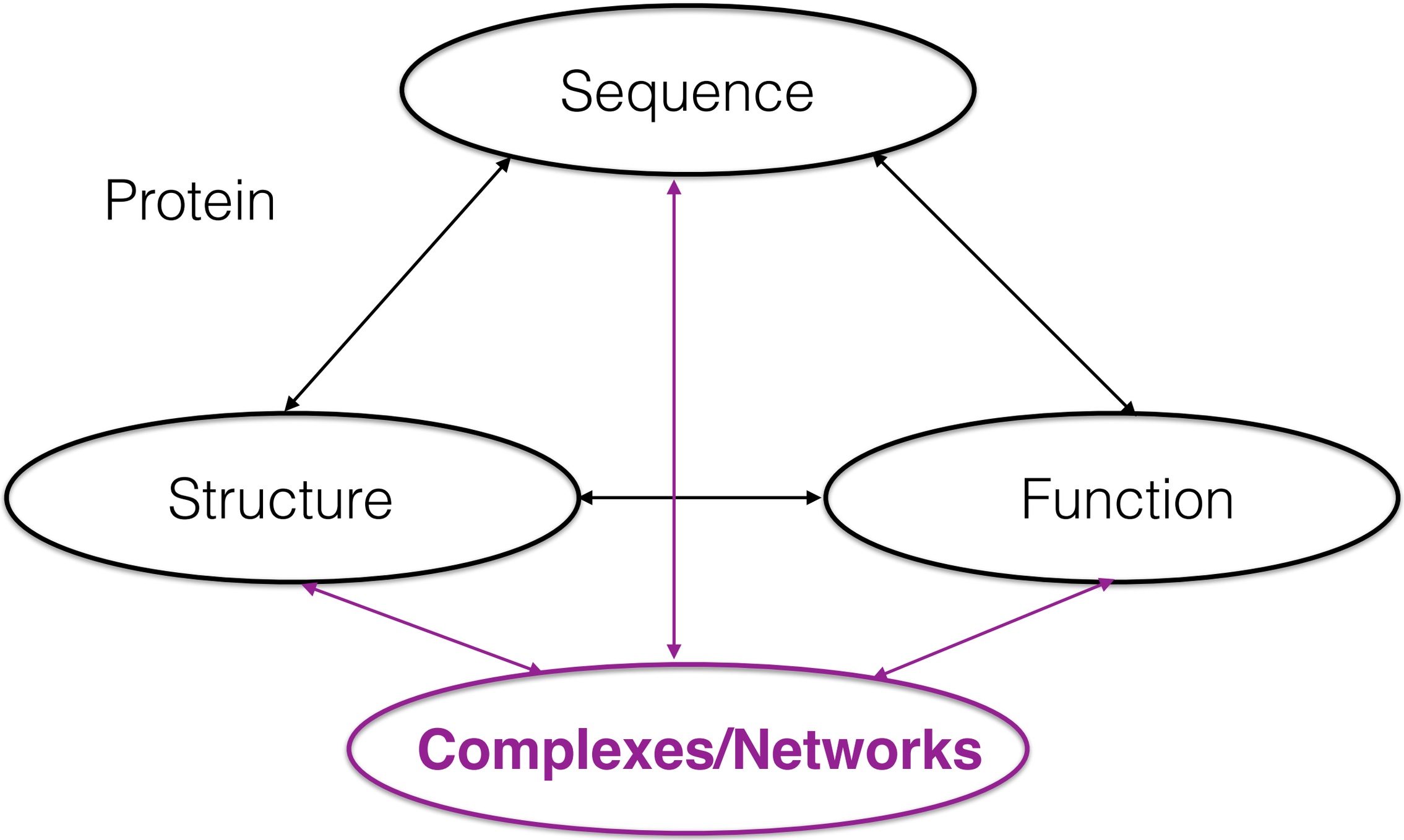
Genomics/Proteomics

Structural biology

How should this course help you?

Challenge:

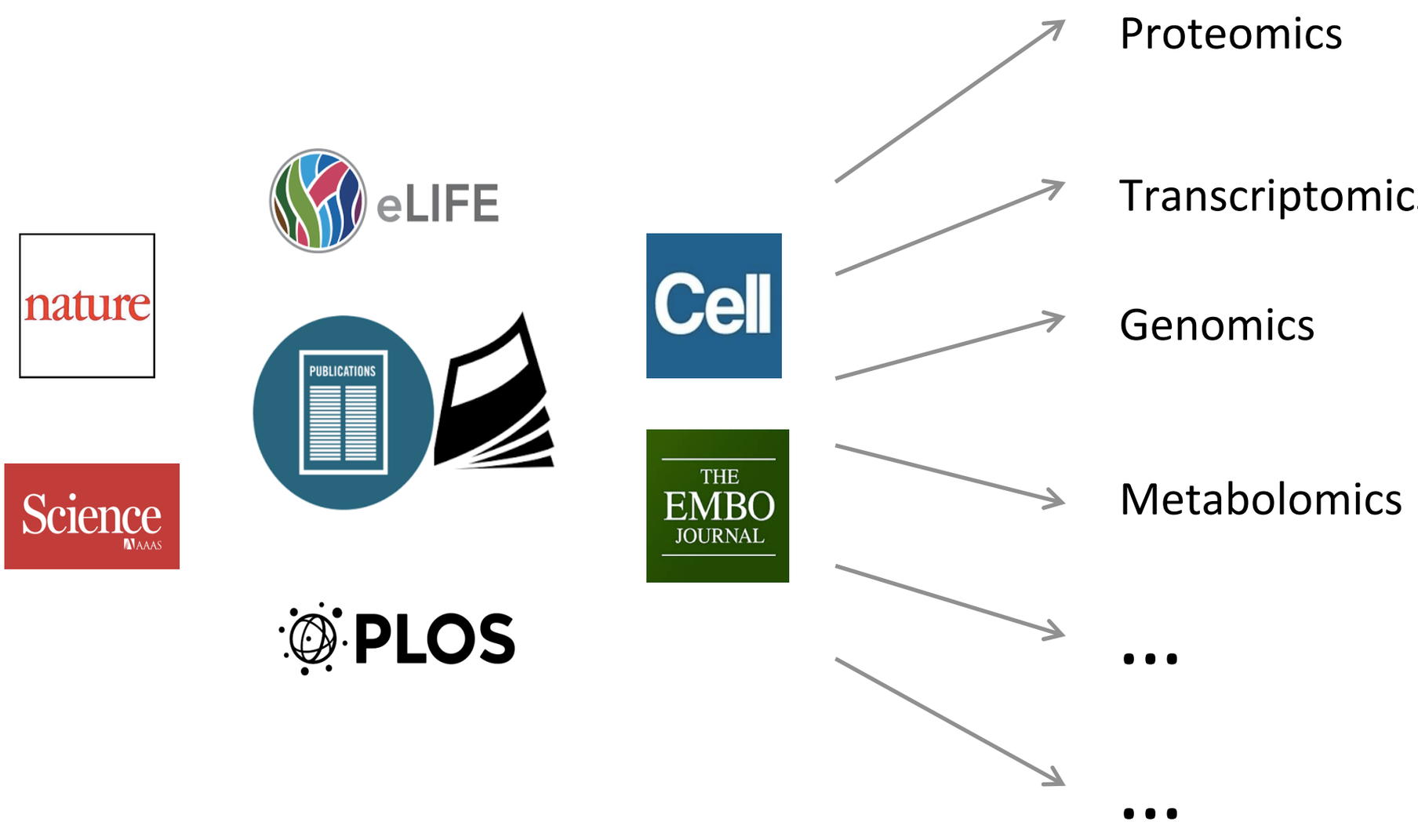
Understanding the complexity of protein sequence-structure-function-interaction relationships



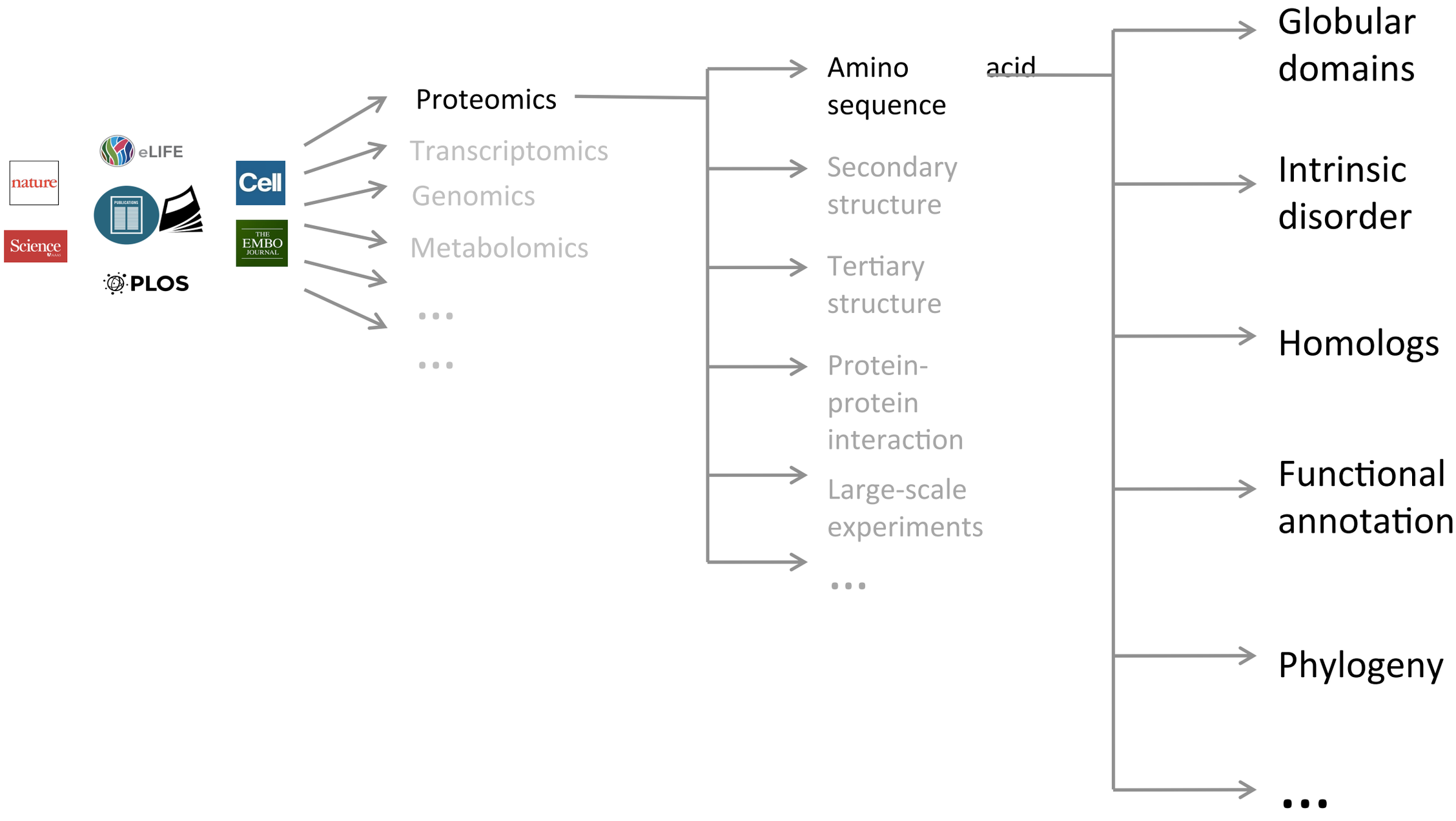
Protein sequence-structure-function-complexes-networks are all interrelated !!!!!

BIOINFORMATICS CAN HELP UNDERSTAND PROTEIN FUNCTION
OK, but HOW?

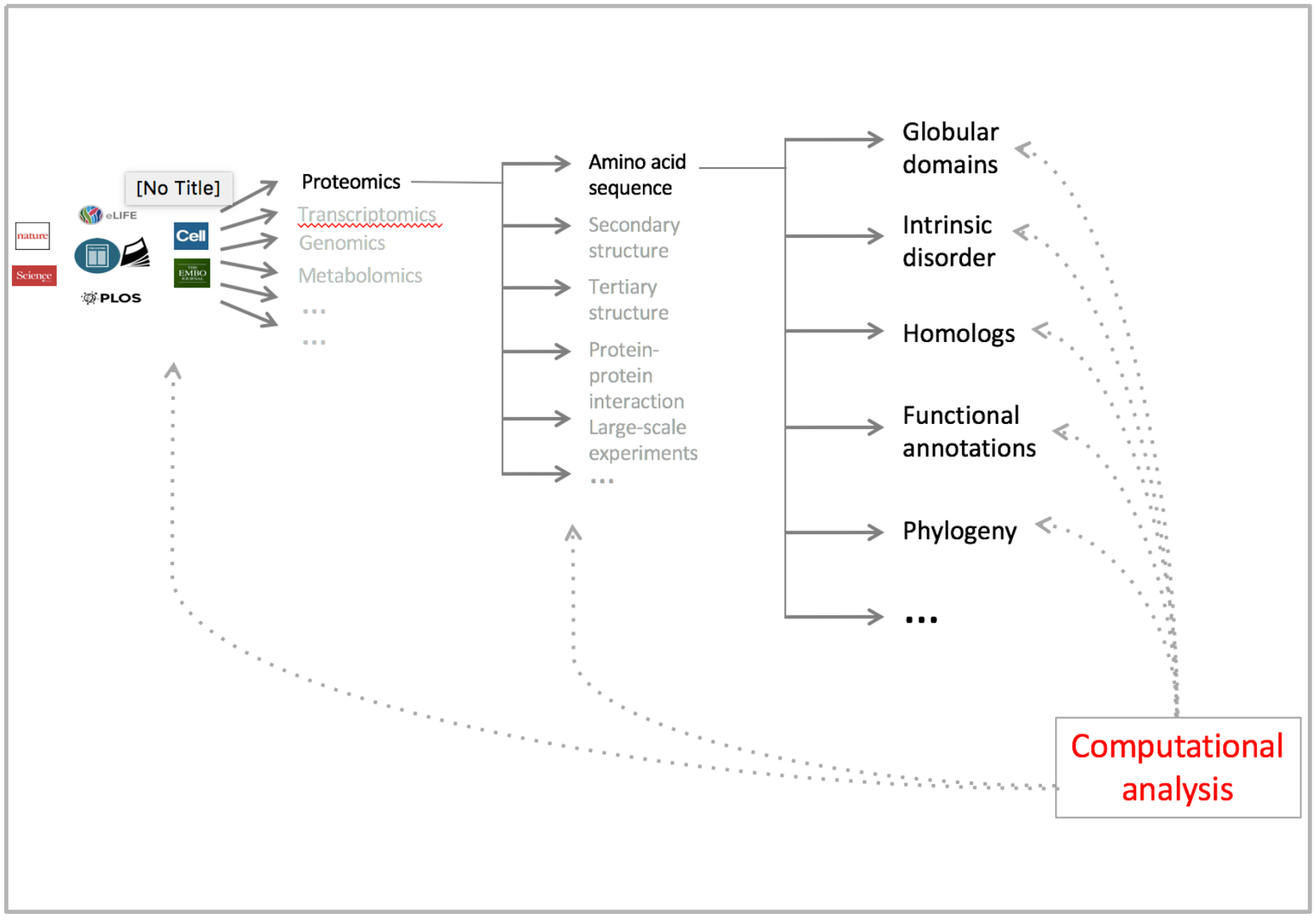
Biological Data



Protein Data



Bioinformatics Database



Organizing particular type of dataset

Databases

Reference databases for proteins

Uniprot, Interpro

Specialized databases

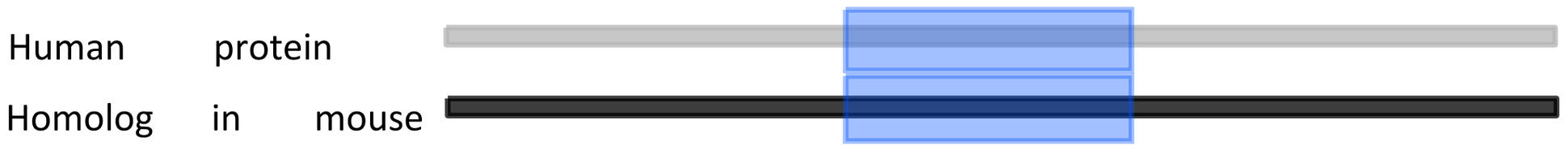
PFAM (domains)

RCSB-PDB (structures)

Sessions 1 & 2 (Tuesday)

Why do we need to compare data?

Experimentally identified RNA binding domain:
RNA Recognition Motif (RRM)



Functional annotation by homology mapping

Using a short functional motif as a query

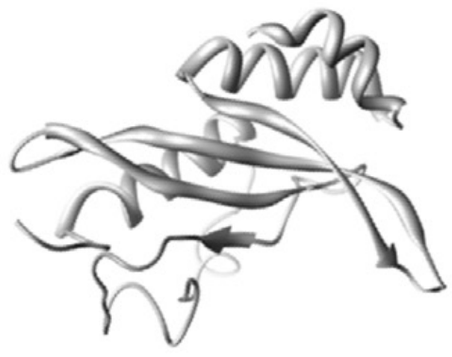
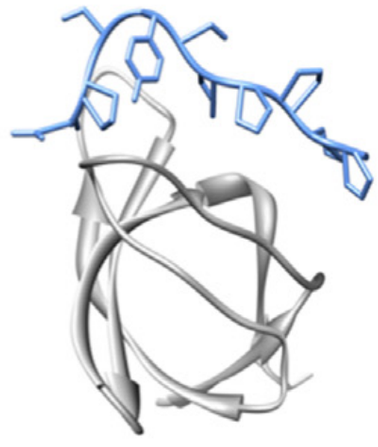
Identify other RNA binding proteins

Structure

Challenging the protein Structure-Function Paradigm:

Proteins have Structured and Disordered Domains

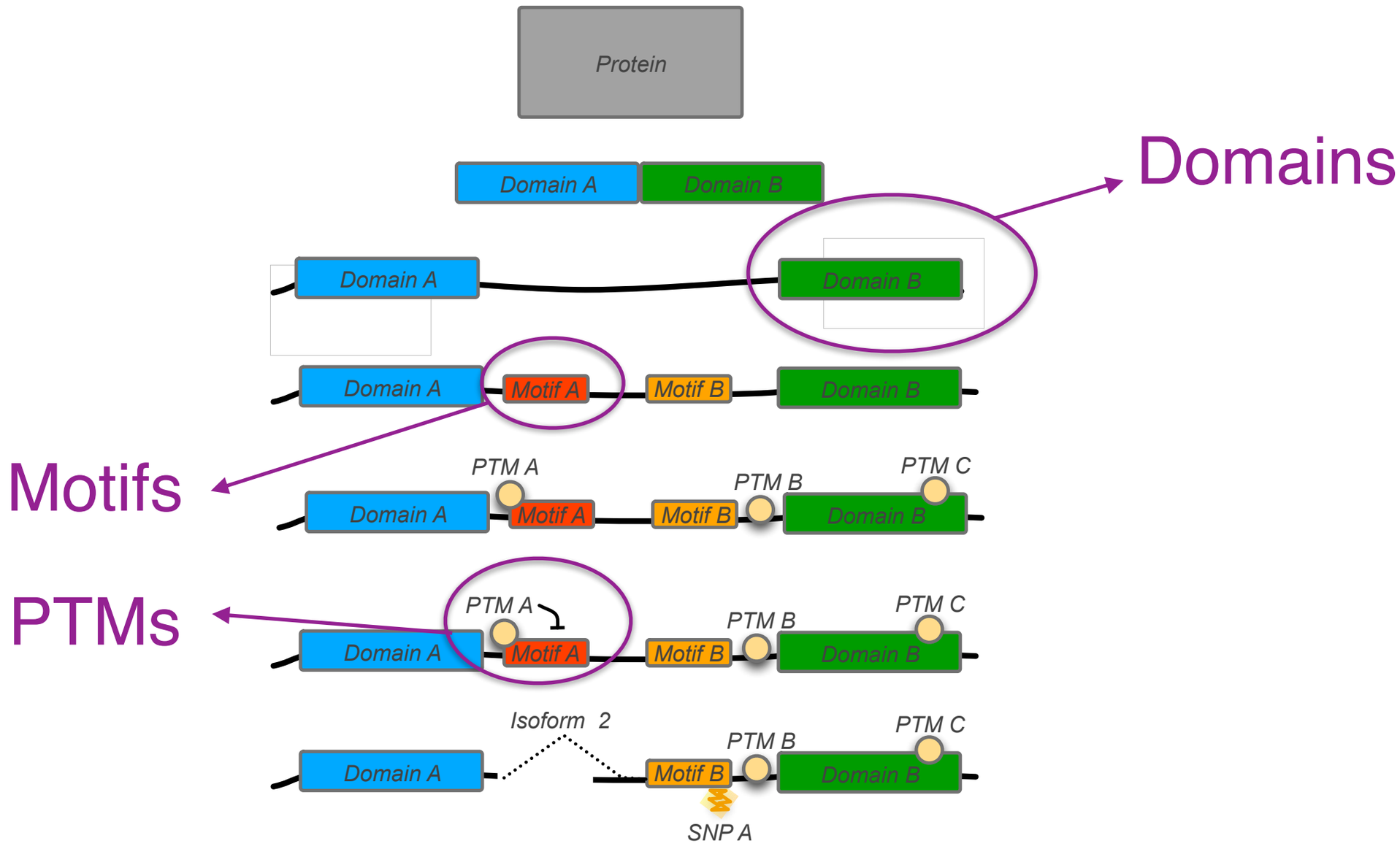
Function is present in BOTH!!!

Protein complexes	Globular domains	Binding motifs	PTM sites
			
>1,000	~100	~10	~1
~600-1,000	~35,000	~100,000	~1,000,000

Typical size (residues)

Estimated instances

Modularity in Proteins



Functional elements can be found in Domains and motifs and Function can be further modulated by post translational modification, degradation, sub-cellular targeting, etc…

What are Intrinsically Disordered Proteins or "IDPs"?

- Intrinsically disordered proteins/regions (IDPs/IDRs)
- Do not adopt a well-defined structure in isolation under "native-like" conditions
- Highly flexible ensembles
- Involved in various diseases
- *Used by pathogens to hijack the cell*

They are not "artifacts" !!!!

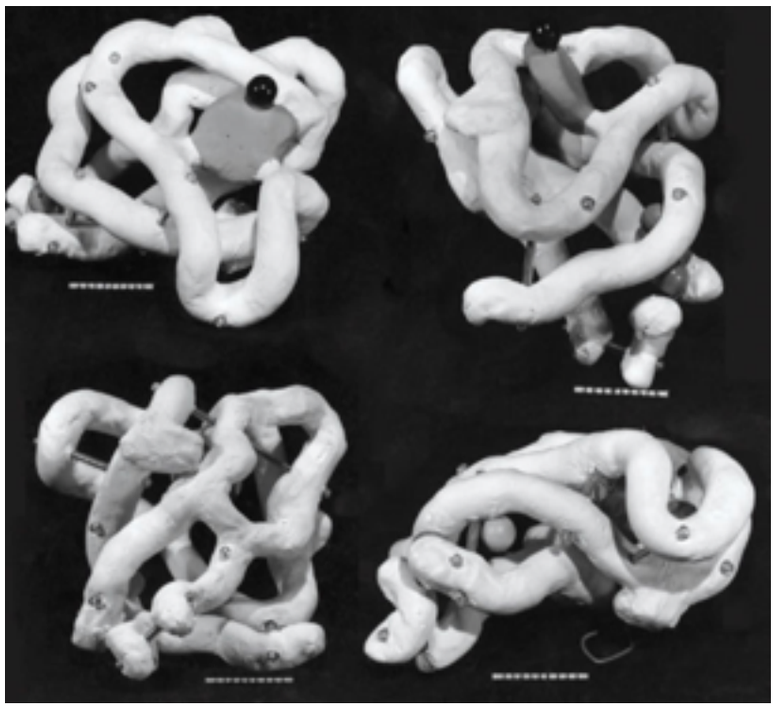
Functions of IDPs

- IDRs harbor binding, targeting, and PTM motifs
- Flexible "linkers" between domains
- Can bind to a large number of protein partners
- They can evolve quickly…motifs can appear *de novo*

convergent evolution of functions

"IDPs" have only recently been recognized

1958: John Kendrew & Max Perutz
First Protein Structure: myoglobin



1999: *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-

Where can we find disordered proteins?

In the literature

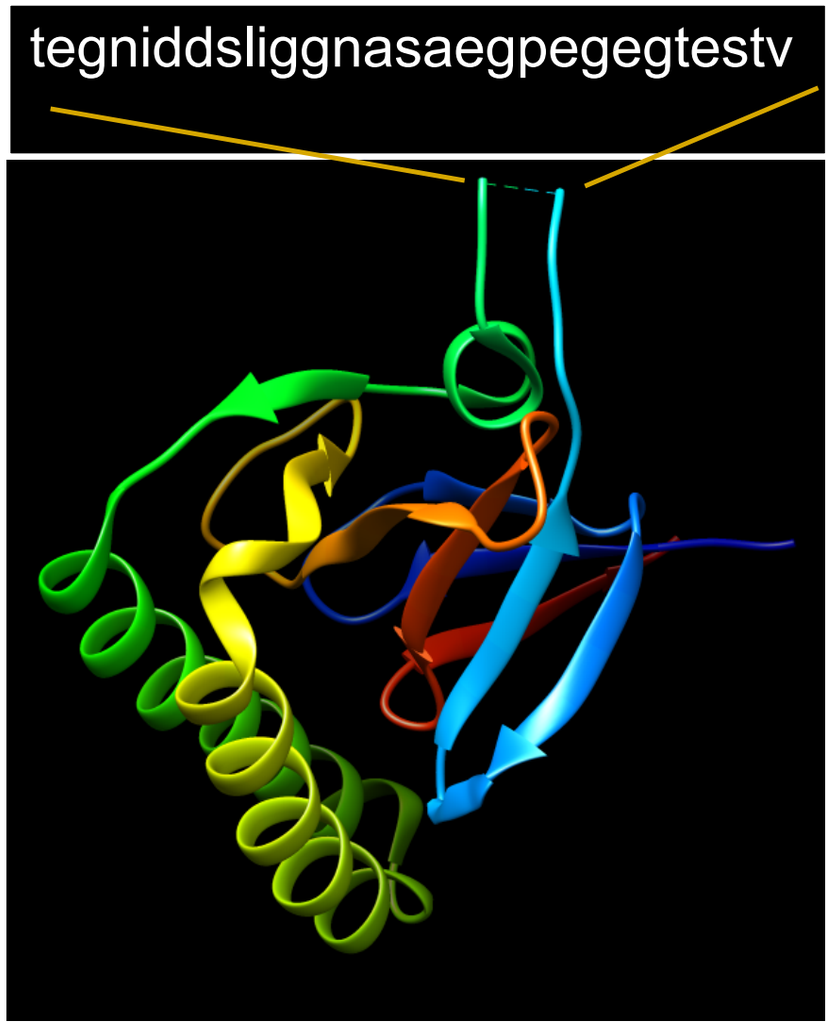
Failed attempts to crystallize
Lack of NMR signals
Heat stability
Protease sensitivity
Increased molecular volume
“Freaky” sequences …

In the Lab, the protein…

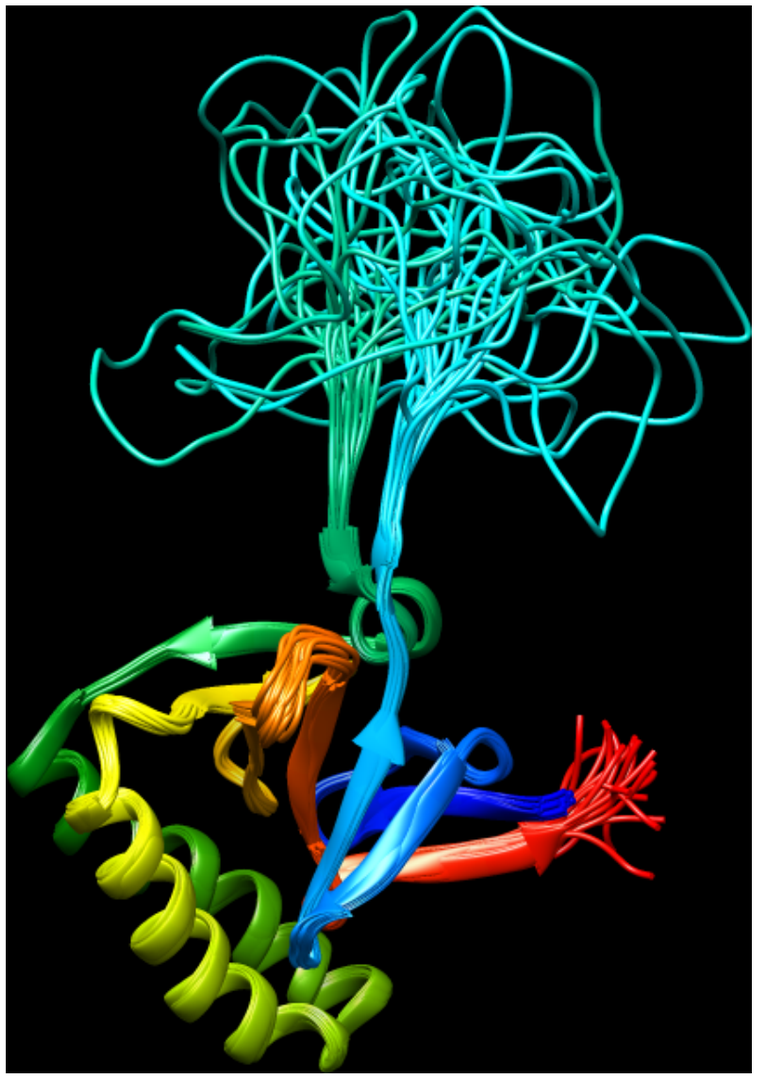
- “doesn’t behave well”
- “doesn’t crystallize”
- “has weird migration on SDS-PAGE”
- “has a large elution volume on SEC”
- “degrades easily”
- “has no secondary structure (CD)”
- “aggregates” easily

Where can we find disordered proteins?

In the PDB



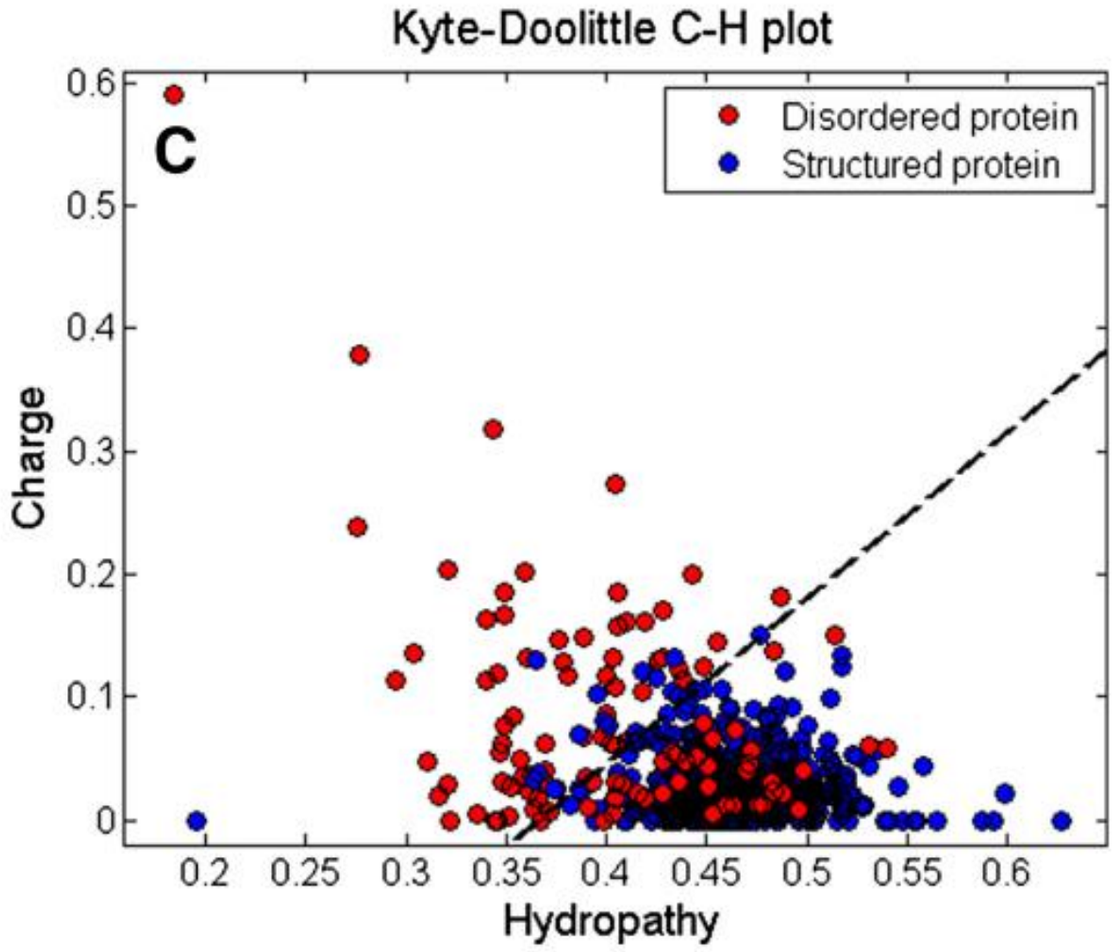
Missing electron density regions from the PDB



NMR structures with large structural variations

Sequence properties of IDPs

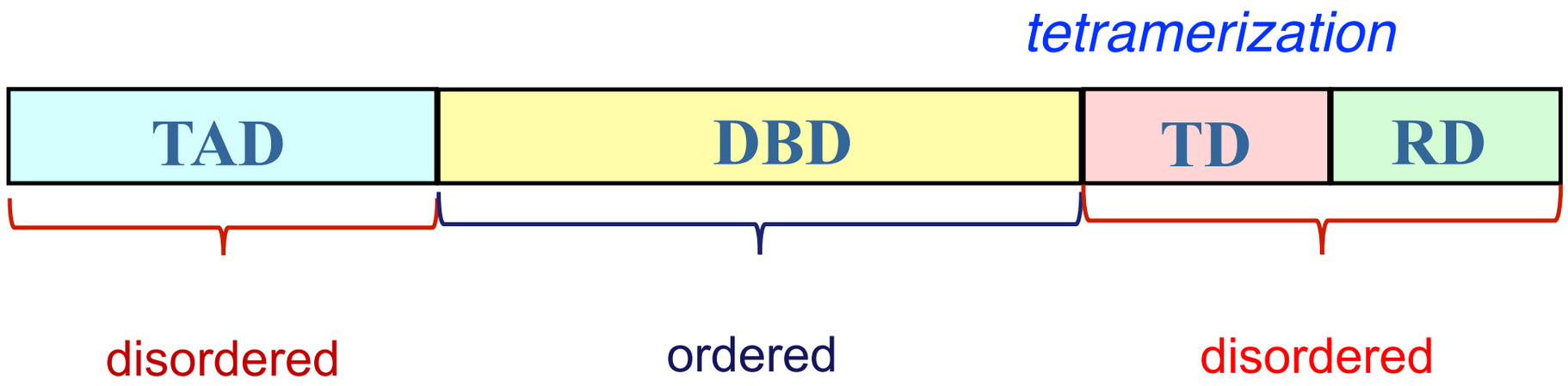
Charge-hydropathy plot



- Sequence compositional bias (low hidrophobicity, high net charge)
- High proline content
- Low “sequence complexity”

A "Famous" IDP: the p53 tumor suppressor

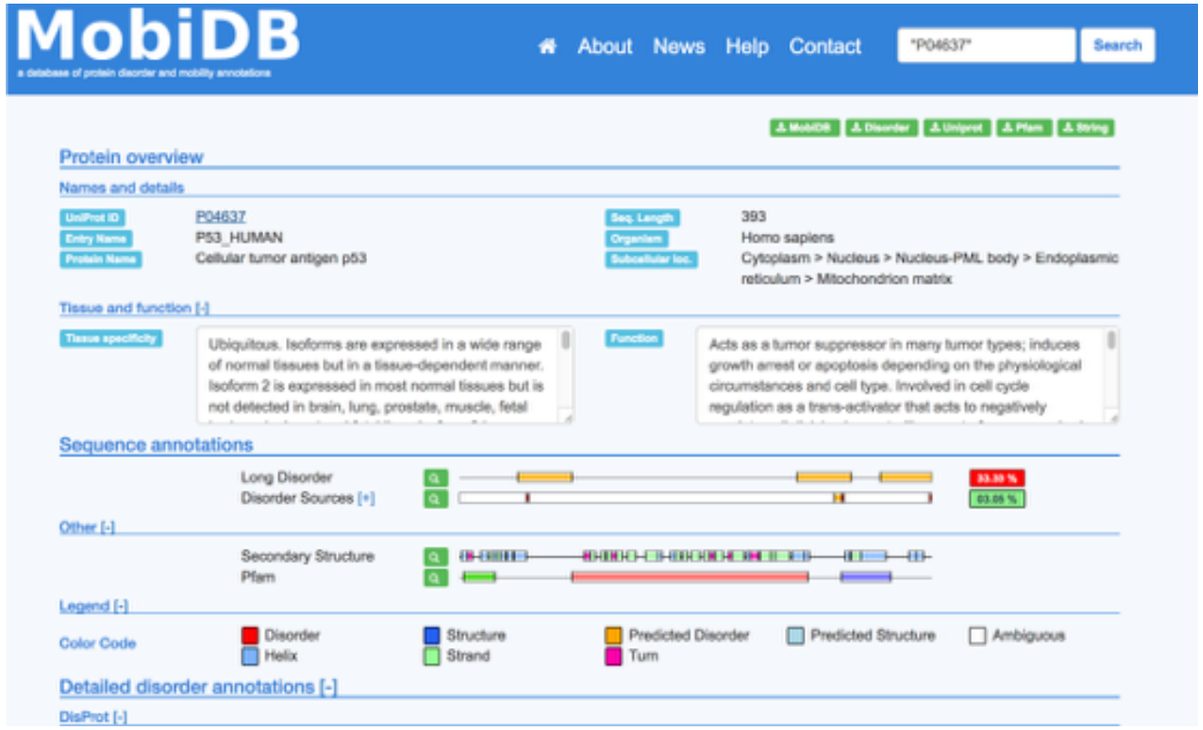
disordered regions mediate multiple protein-protein interactions



Wells et al. PNAS 2008; 105: 5762

Where do we learn about “disordered regions”?

Prediction



IUPred - http://iupred.enzim.hu/

MobiDB - http://mobidb.bio.unipd.it/

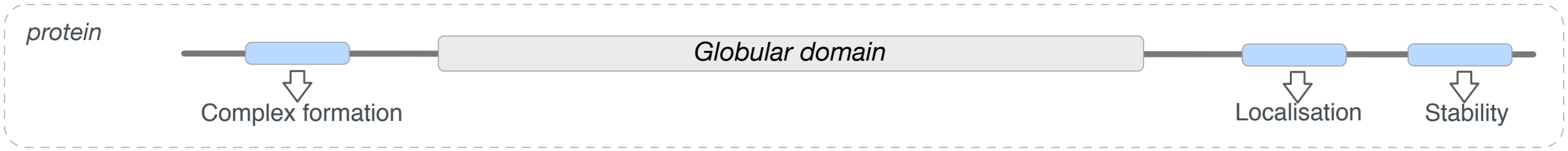
Experimental annotation:
803 proteins; 2167 regions

DISPROT: http://www.disprot.org

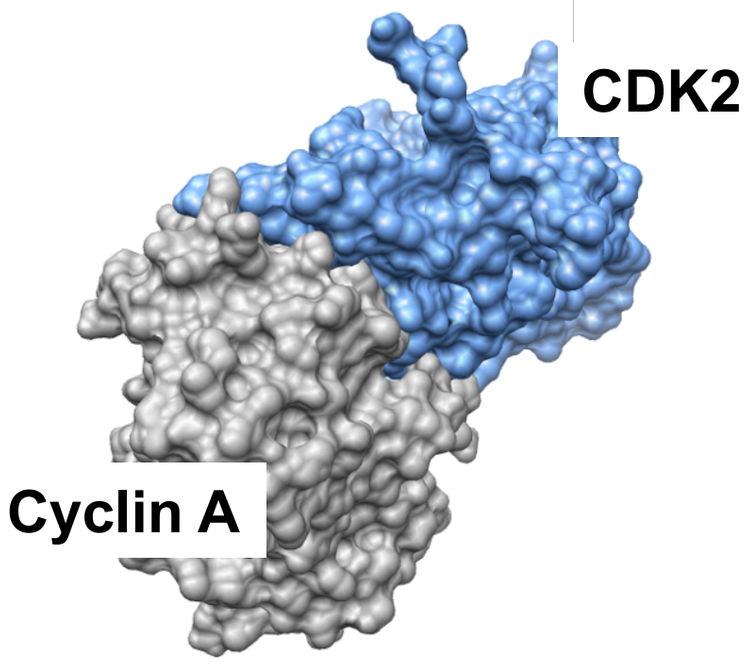
Session 3: Experimental investigation and prediction of intrinsically disordered regions

Function encoded within IDP regions: IDRs and SLiMs

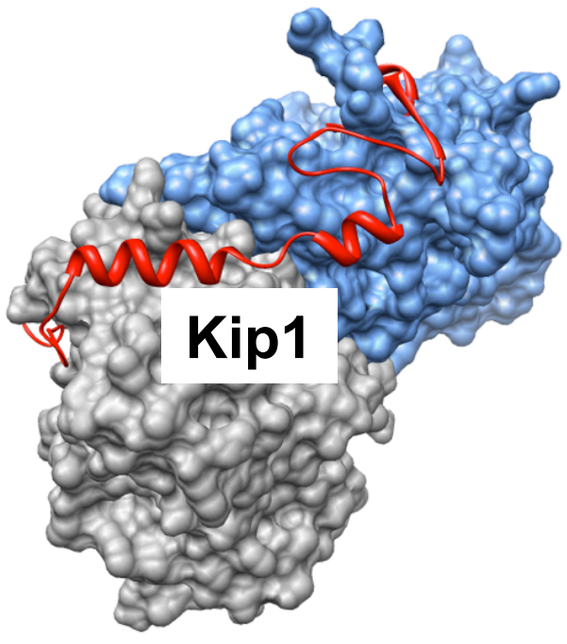
Protein Module



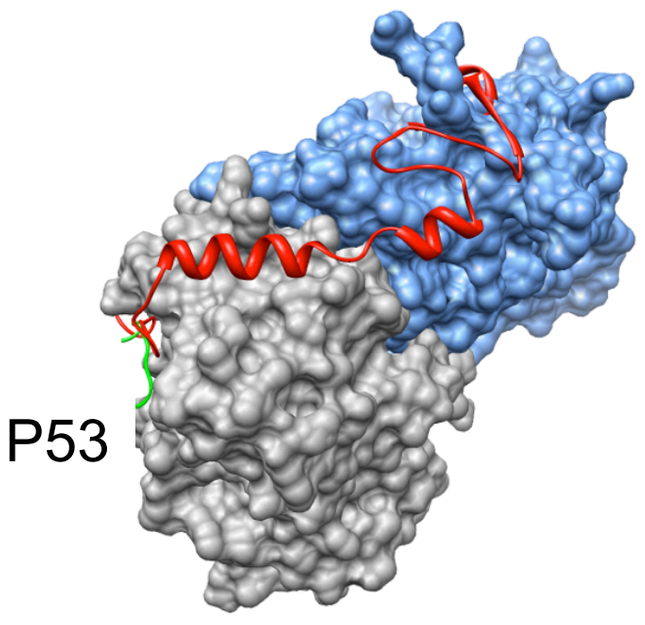
Globular Domain-Globular Domain Interaction



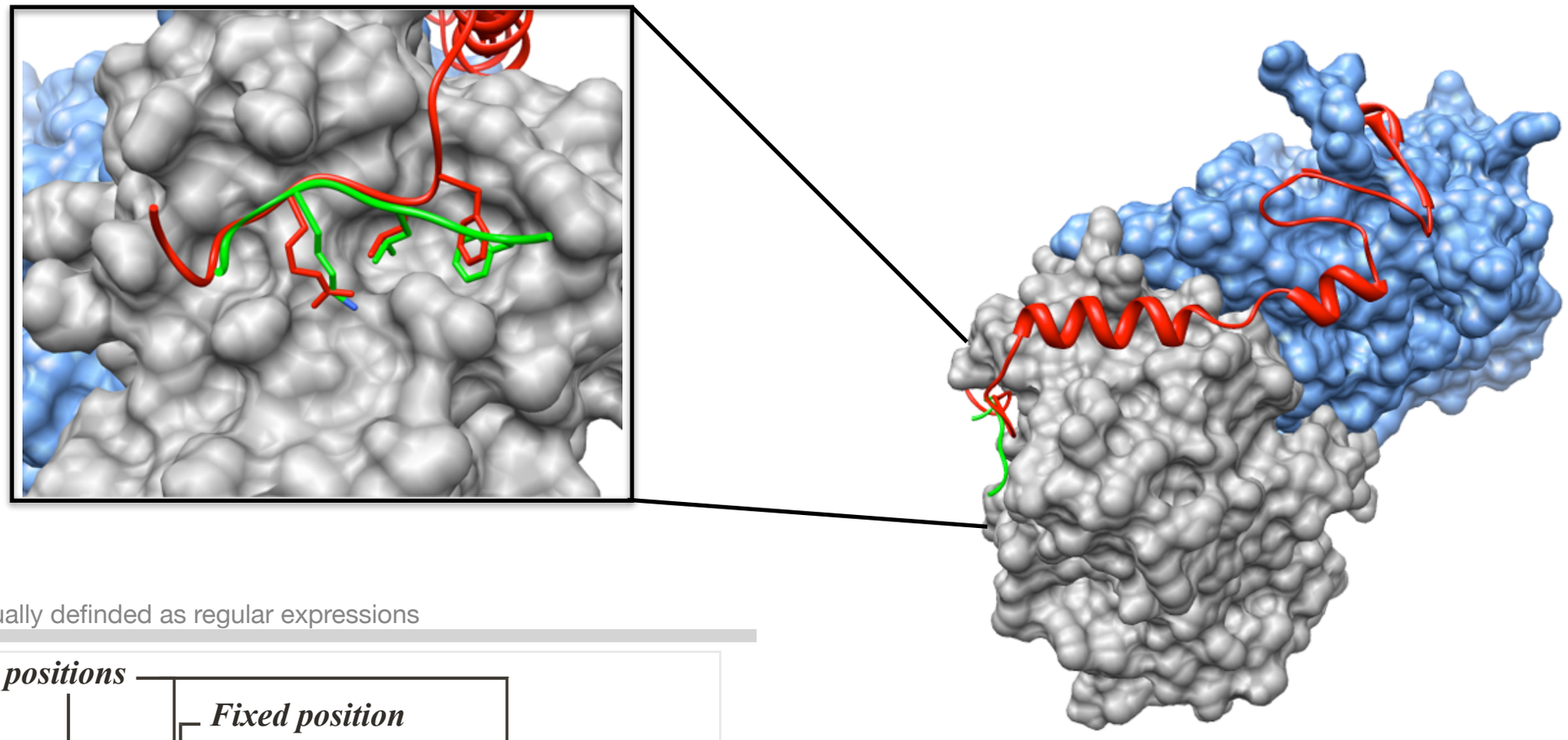
Globular Domain-Disordered Domain Interaction



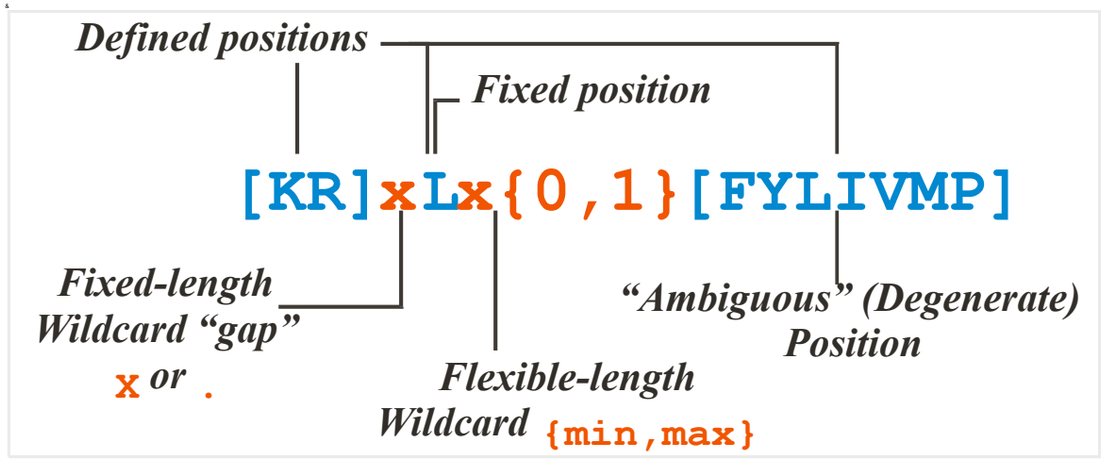
Globular Domain-Linear Motif Interaction



A SLiM can be defined by “regular expressions” which describe its sequence determinants



- Motifs are usually definded as regular expressions



&

Figure 3.3. Anatomy of a SLiM.
Definitions of different properties of SLiM have been marked on the example ELM, LIG_CYCLIN_1 (Puntervoll et al. 2003). This motif has three defined positions (one fixed and two degenerate) and two wildcard spacers (one fixed, one flexible-length) for a total length of 4-5aa.

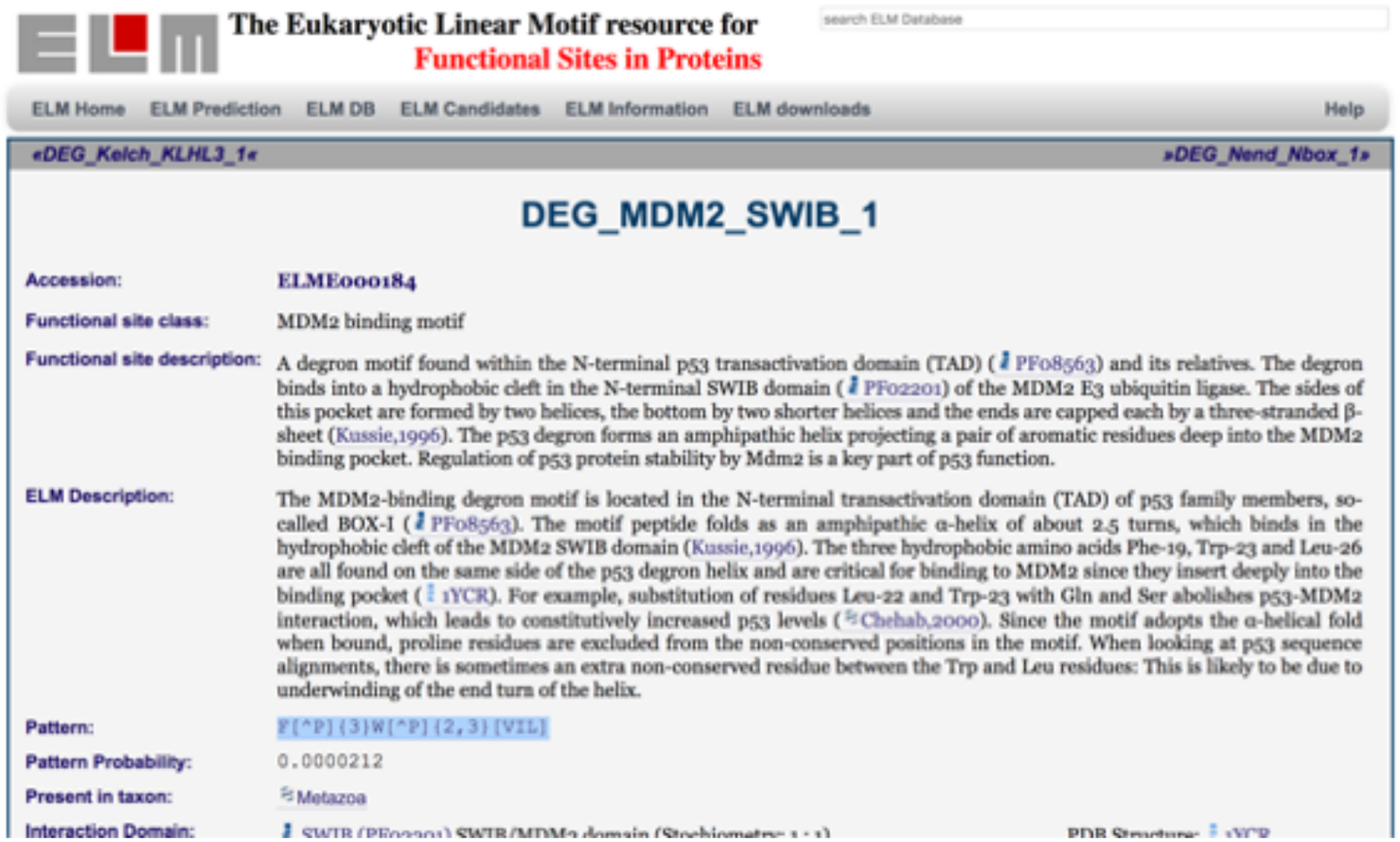
For example…

- The “KDEL” ER retention signal
- The “RGD” Integrin binding motif
- The “RxL” Cyclin Box motif

Short Linear Motifs

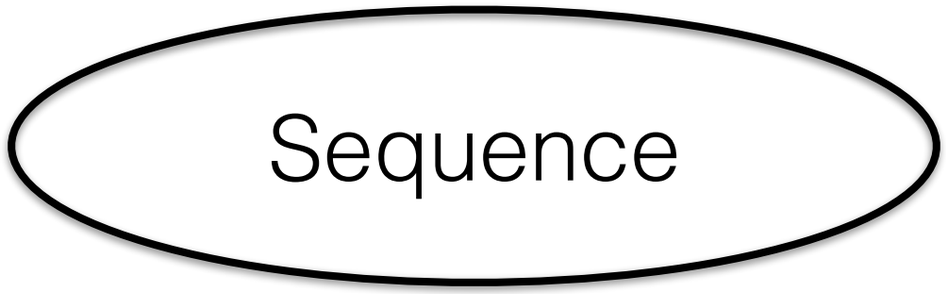
Annotation of eukaryotic SLiMs

- Binding
- Degradation
- Subcellular Targeting
- Modification



ELM - http://elm.eu.org/

Session 4: Annotation and discovery of small linear motifs (SLiMs). The ELM database



Sequence conservation can help identify *domains*

Sequence composition can help predict *disordered regions*

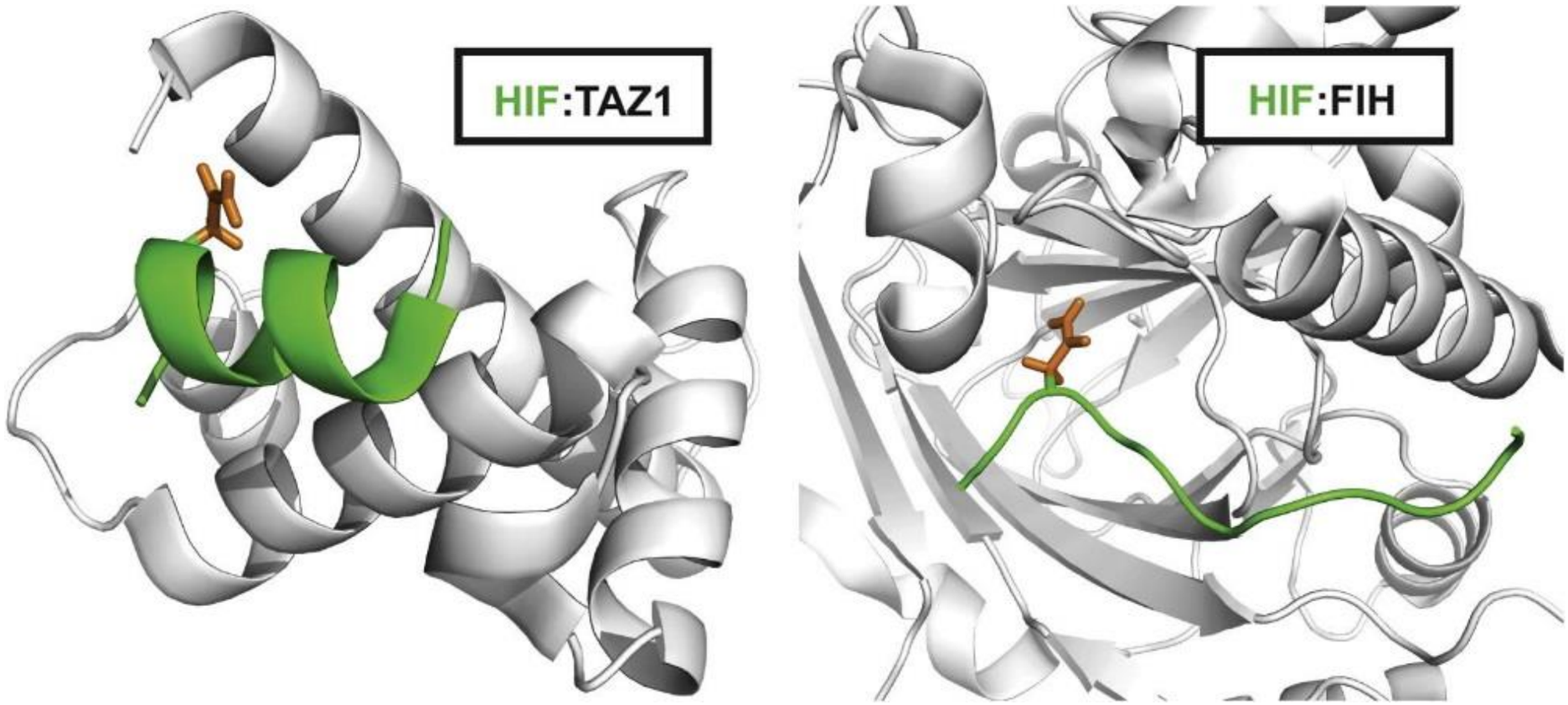
Sequence alignment can help identify other functional elements such as *Linear Motifs (SLiMs)*

Sequences carry valuable *evolutionary information*

Jalview

http://www.jalview.org

Conformational plasticity

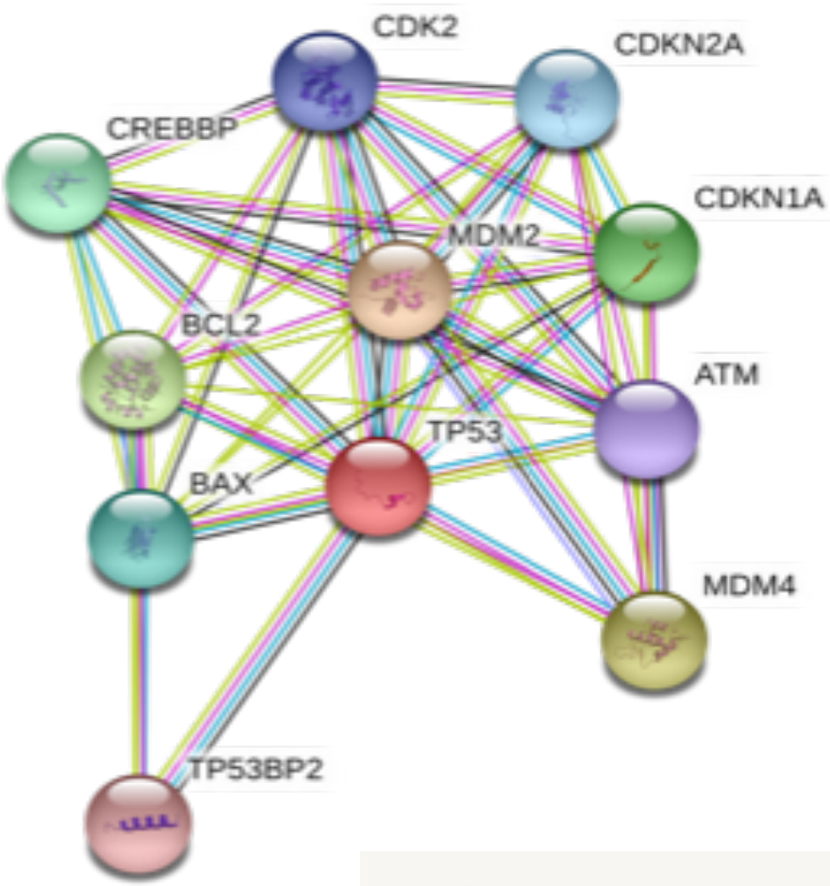


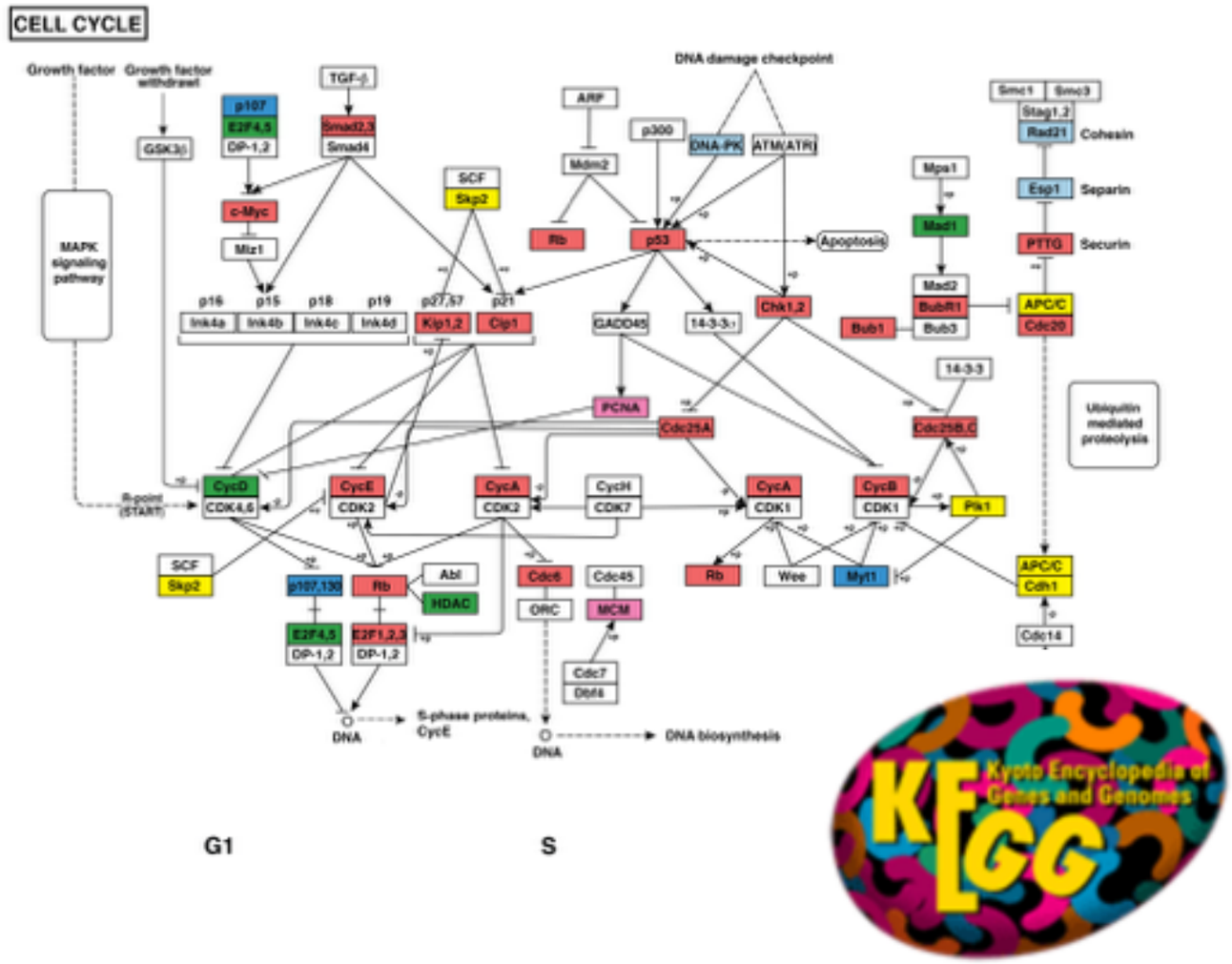
C-terminal transactivation domain (CTAD) of the hypoxia inducible factor-1α

Berlow et al. FEBS Lett. 2015;589:2433

Proteins work within highly connected “interaction networks”

“Molecular switching” changes the functional state of the cell





STRING

https://string-db.org

http://www.genome.jp/kegg/

Session 7: Protein-Protein Interaction Networks

And there is still more complexity to protein function…

- **Proteins are highly localized**
- **Proteins can function within scaffolds**
- **Compartmentalization and crowding: Proteins are not “freely diffusing” in the cytoplasm**

BUT LET’s START WITH SOME BASIC TOOLS TO UNDERSTAND THEM…

always keeping the complexity in mind