

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



Deutscher Akademischer Austausch Dienst
German Academic Exchange Service



MinCyT-DAAD
IIB UNSAM



UNIVERSIDAD
NACIONAL DE
SAN MARTÍN

I I B - I N T E C H

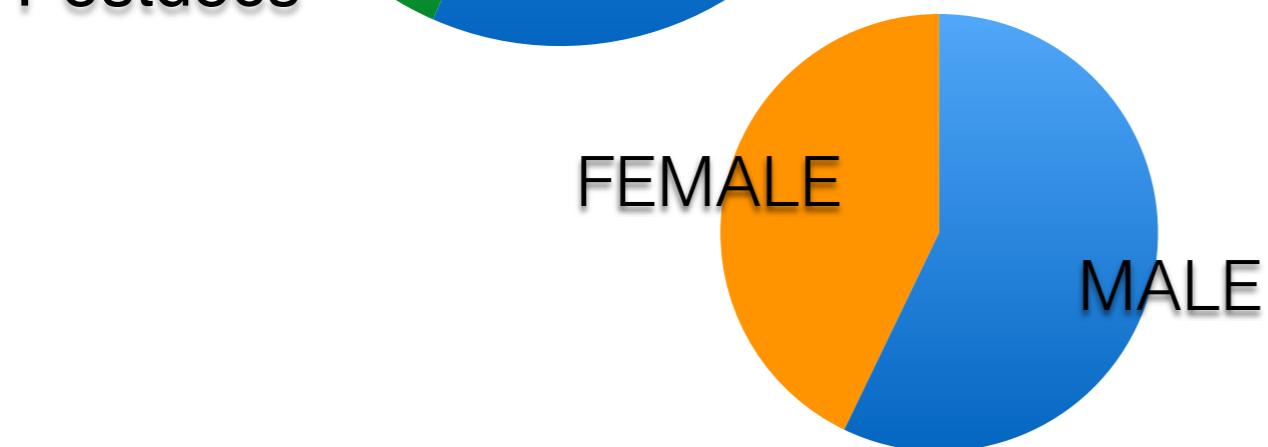
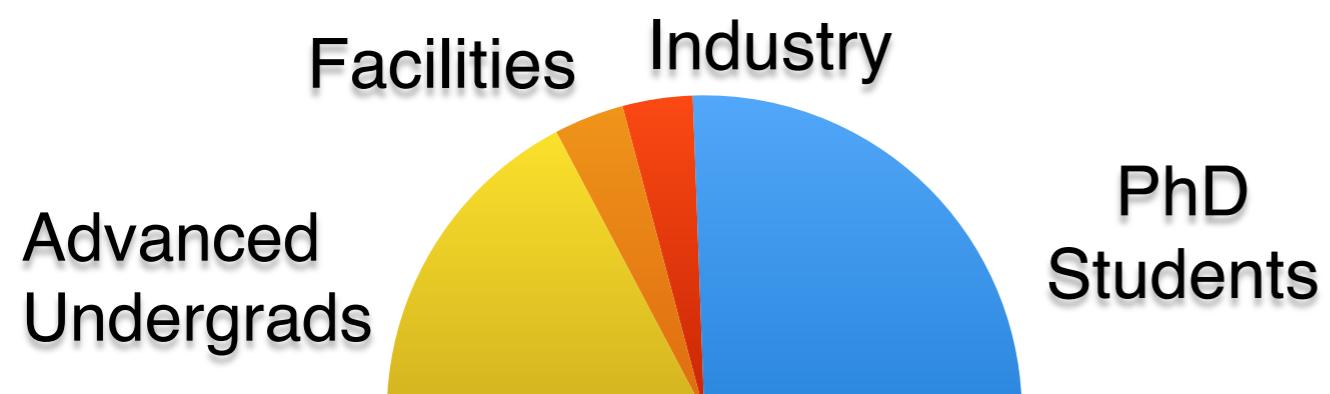


Ministerio de
Ciencia, Tecnología
e Innovación Productiva

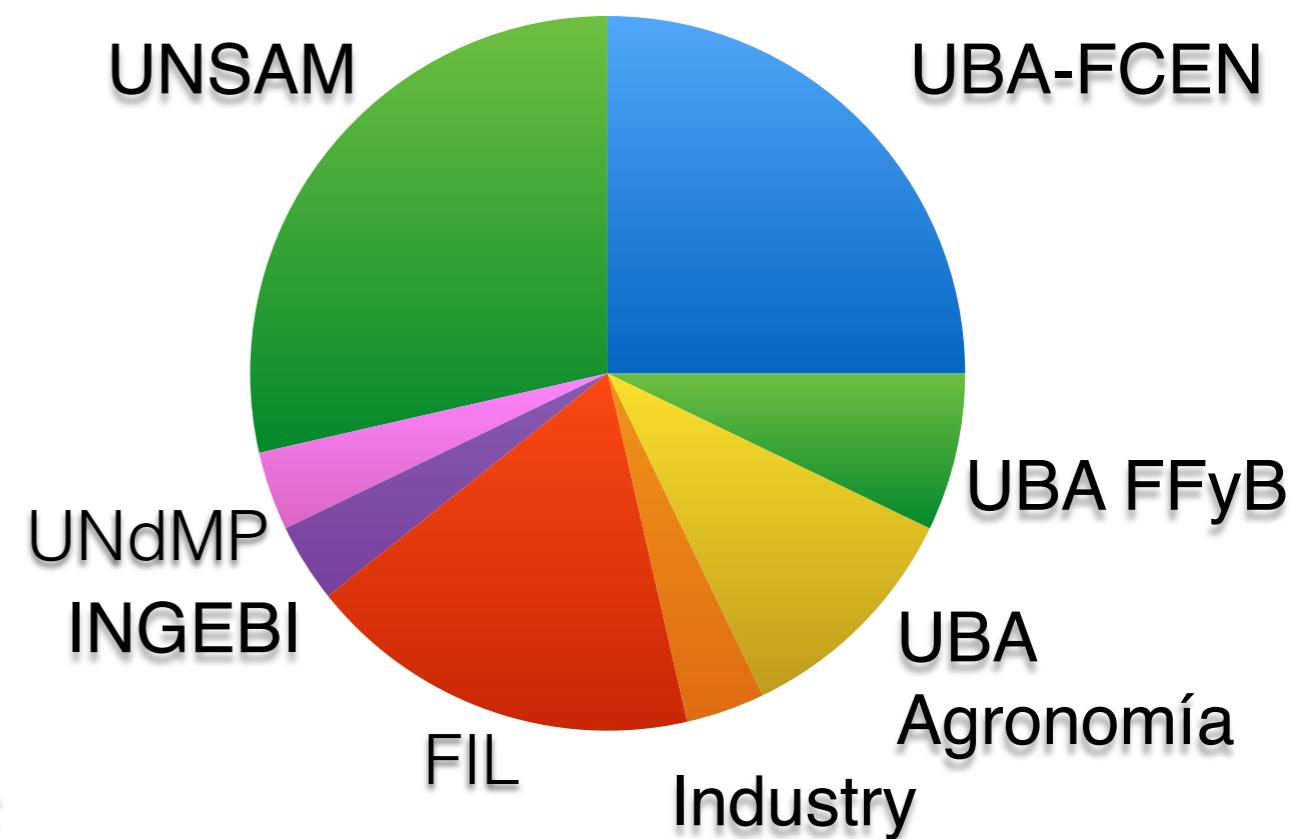
Presidencia de la Nación

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

S. cerevisiae, S. Pombe

Yeast genetics/signaling

A. thaliana, others

Plant physiology

P. aeruginosa, B. abortus

Bacterial Pathogens

HRSV, Adenovirus

Viral Pathogens

T. cruzi, Leishmania

Eukaryotic Pathogens

What is your motivation for taking the course?

Considero investigación tanto complejos resultados estructura
poder entender otras comprender bioinformática
plantas redes Las formación datos resulta
conocimiento experiencia tema analizar estudio
brindará ayudará Creo nuevas muchas gran función
motivación mejor tal desarrollo
este
doctoral sobre doctorado cómo podría útiles importante
nuestro factores aprender mediante grupo
posible mismas técnicas posibles llevan trabajando esto
sistema tesis proteína manejo intrínsecamente regiones
interacciones Para células bioinformáticas
interesante realizar cruzi laboratorio
desordenadas estructural serían

Dynamic protein assemblies

Intrinsically Disordered regions

Protein-protein interaction networks

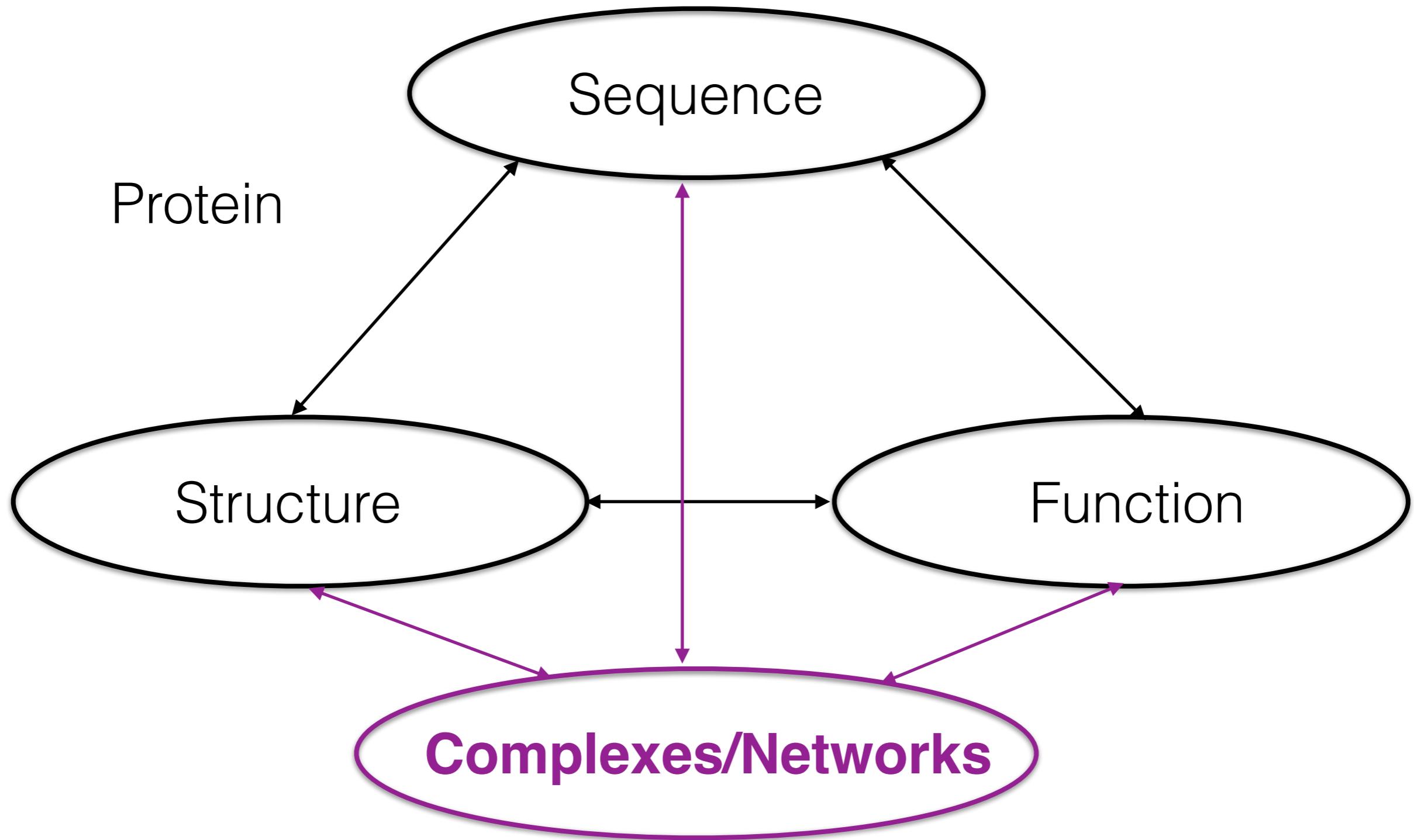
Genomics/Proteomics

Structural biology

How should this course help you?

Challenge:

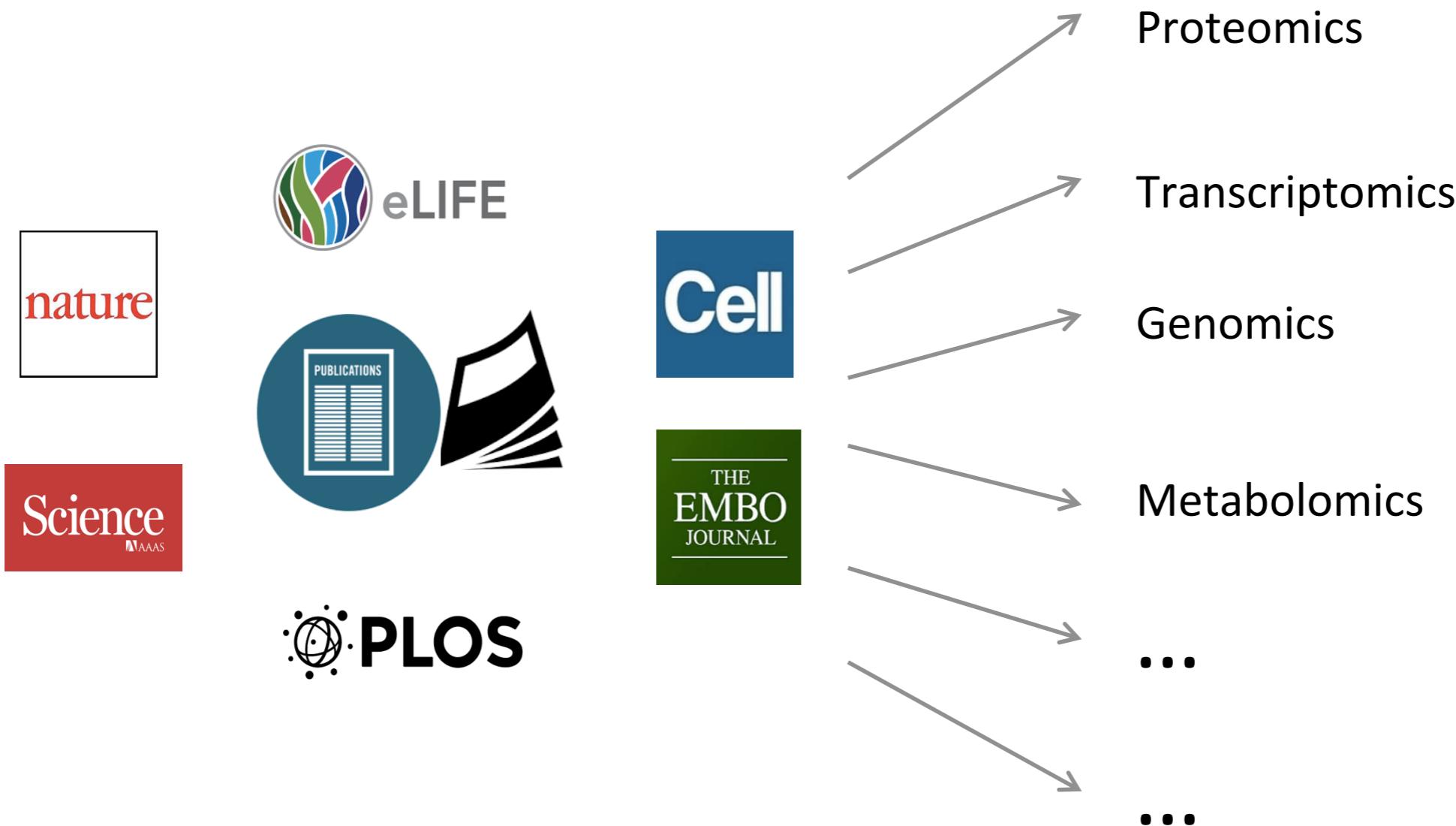
**Understanding the complexity of protein
sequence-structure-function-interaction
relationships**



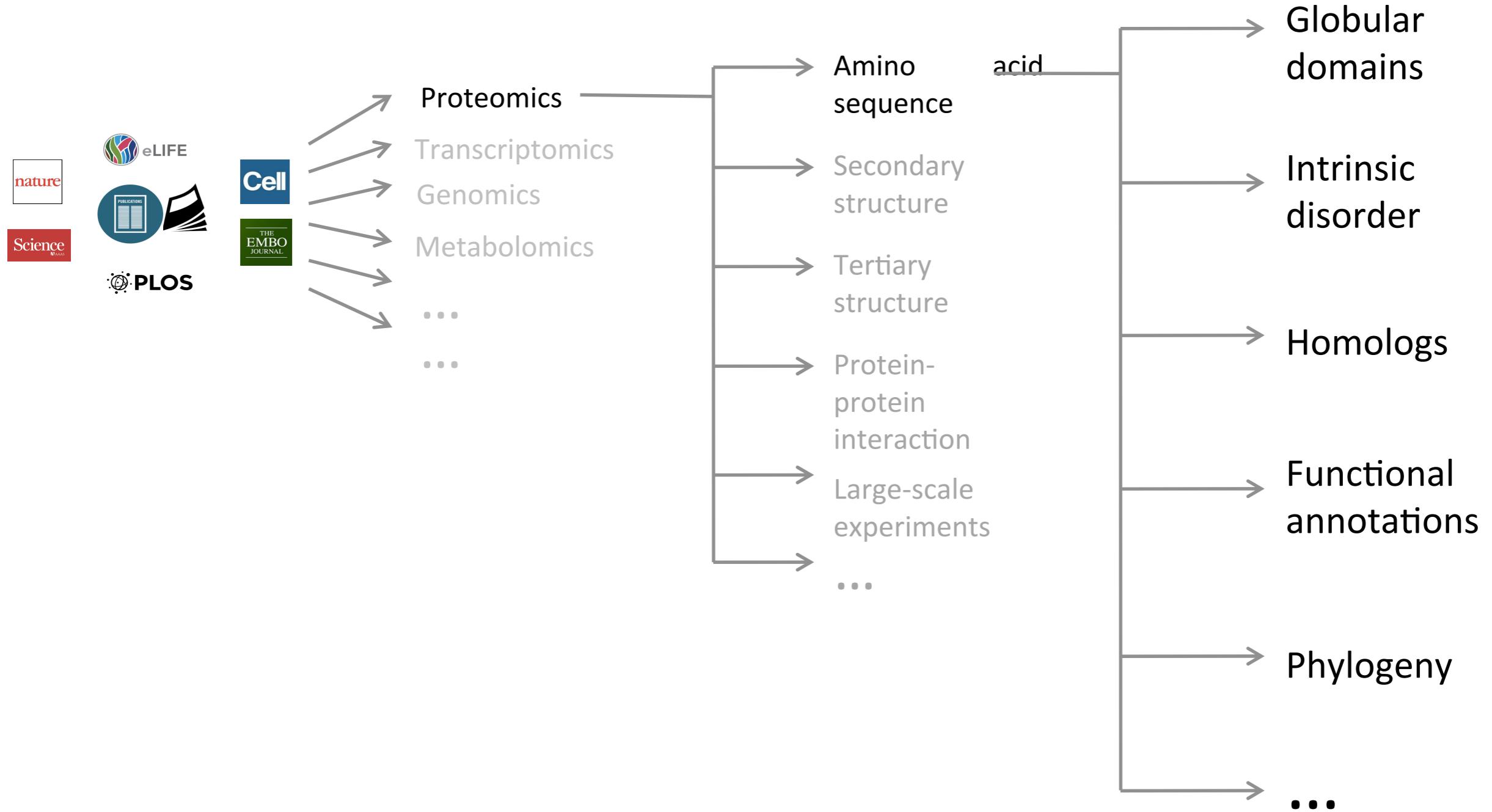
Protein sequence-structure-function-complexes-networks are all interrelated !!!!!

BIOINFORMATICS CAN HELP UNDERSTAND PROTEIN FUNCTION
OK, but HOW?

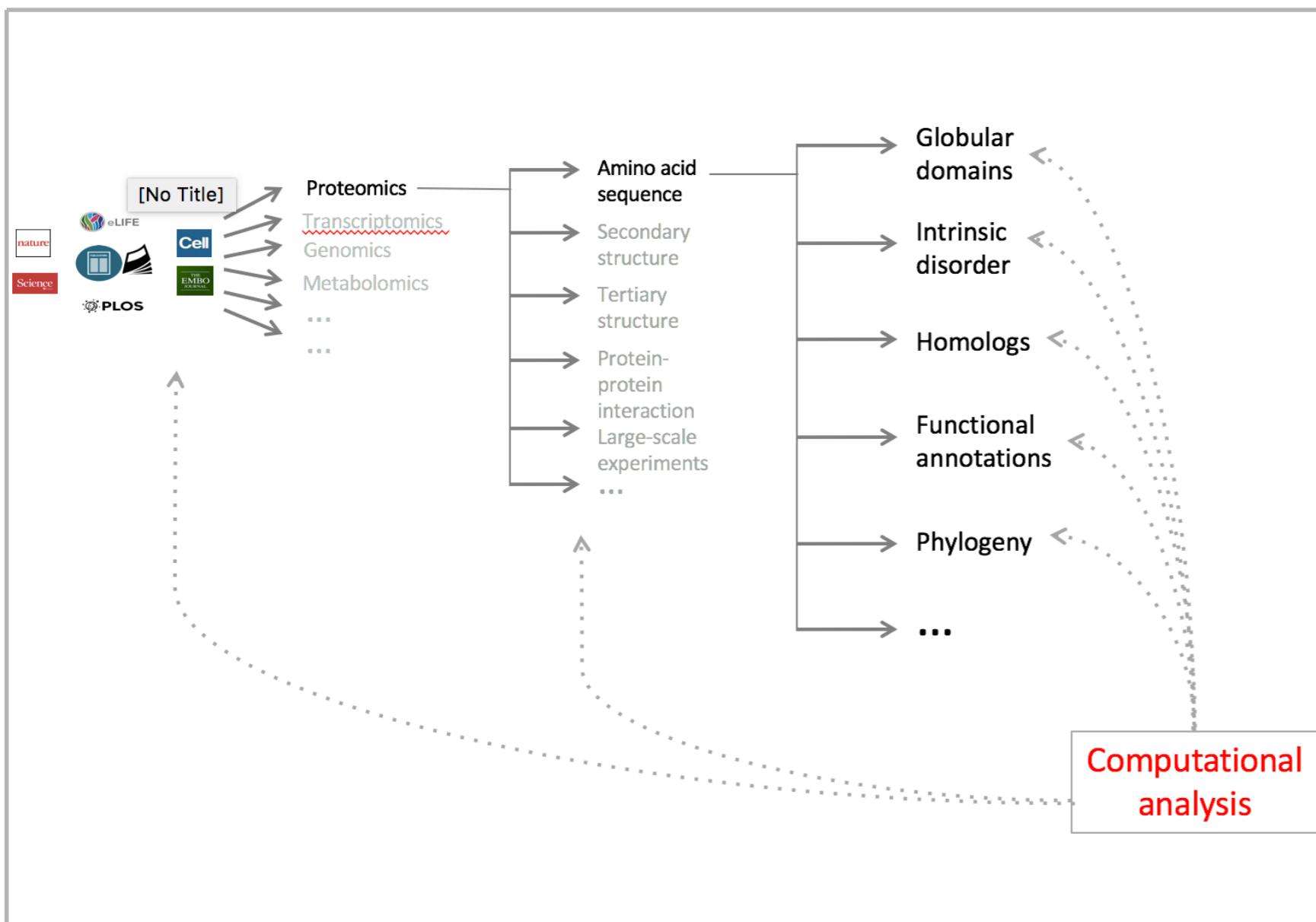
Biological Data



Protein Data



Bioinformatics Database



Organizing particular type of dataset



Databases

Reference databases for proteins

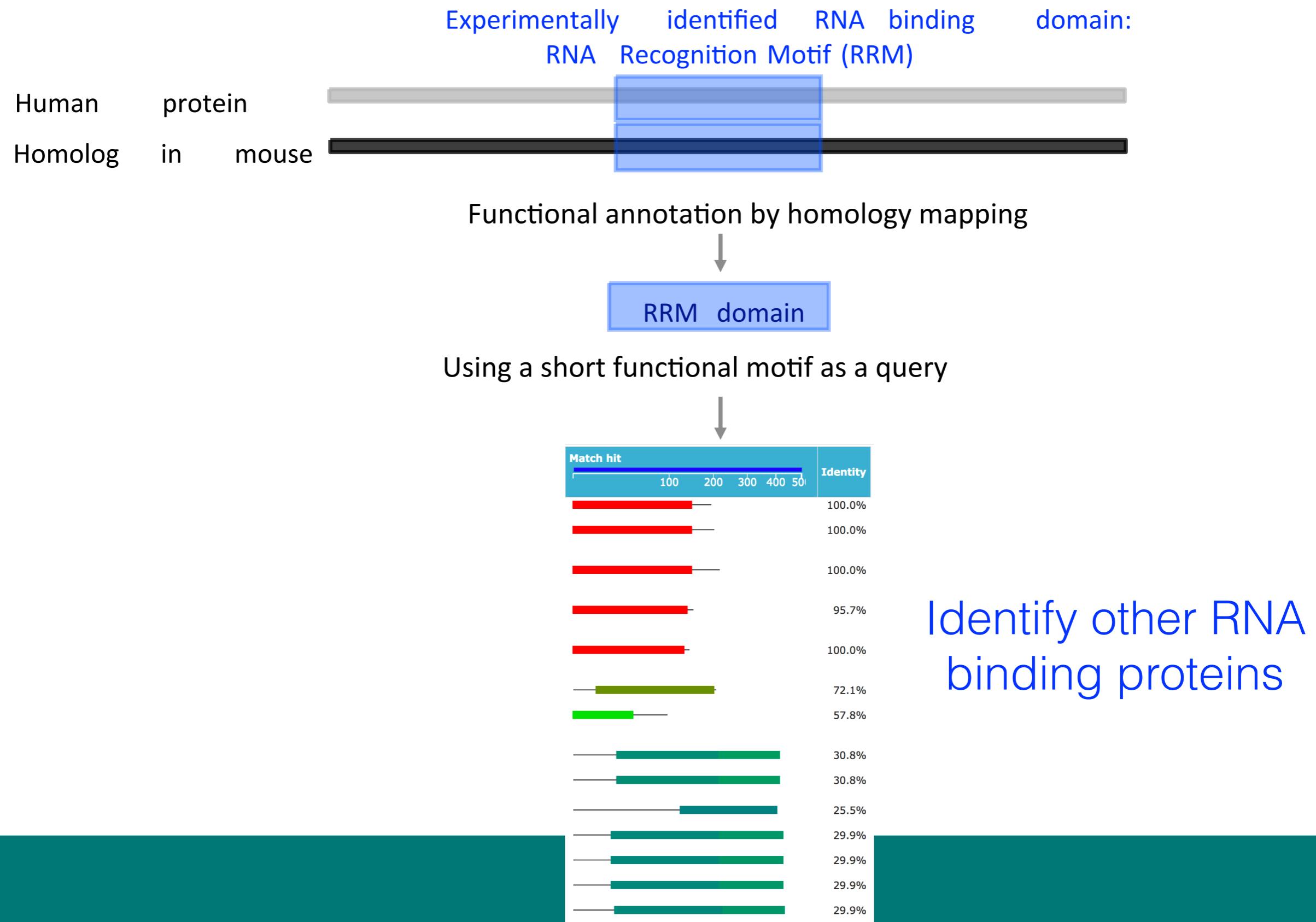
Uniprot, Interpro

Specialized databases

PFAM (domains)
RCSB-PDB (structures)

Sessions 1 & 2 (Tuesday)

Why do we need to compare data?

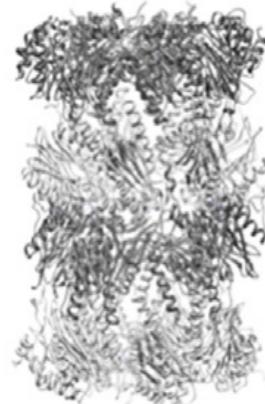


Structure

Challenging the protein Structure-Function Paradigm:

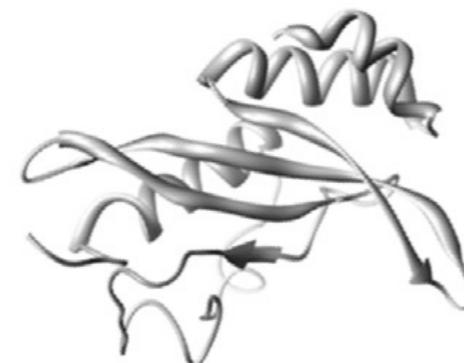
Proteins have Structured and Disordered Domains
Function is present in BOTH!!!

Protein complexes



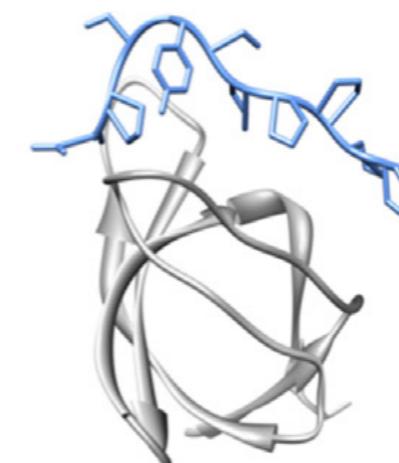
>1,000

Globular domains



~100

Binding motifs



~10

PTM sites



~1

Typical size (residues)

~600-1,000

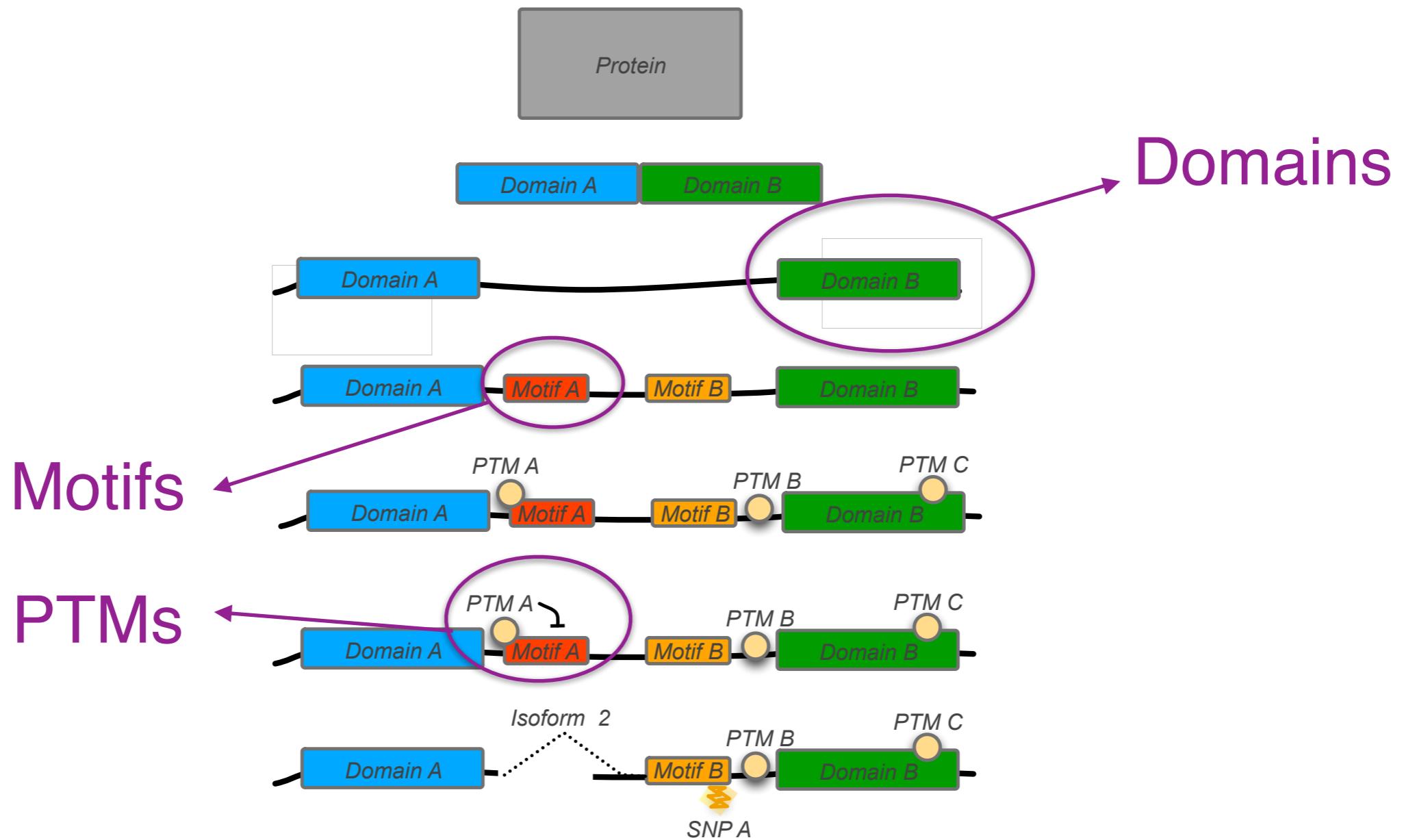
~35,000

~100,000

~1,000,000

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*

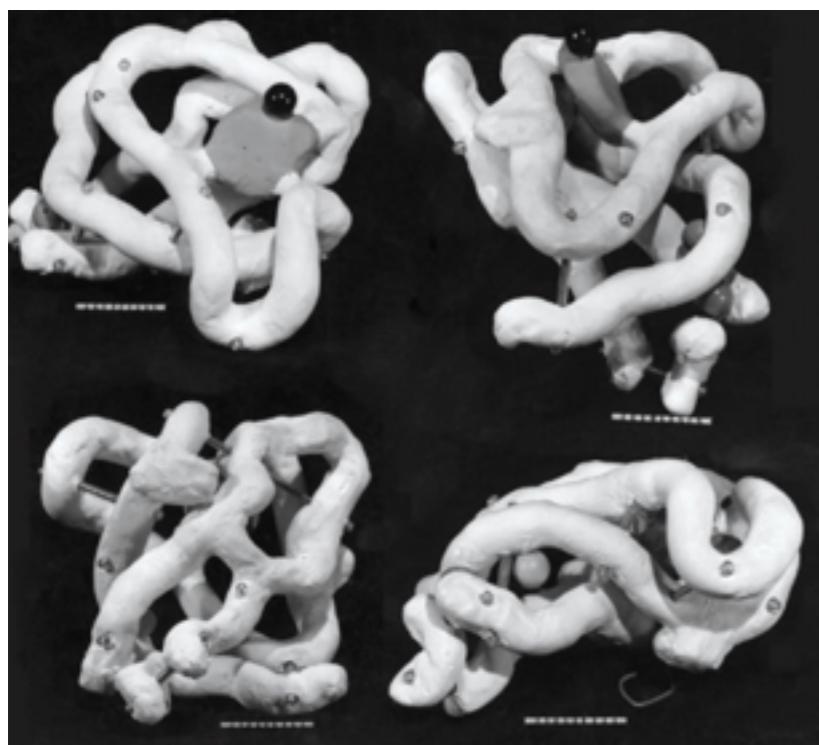
They are not “artifacts” !!!!

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

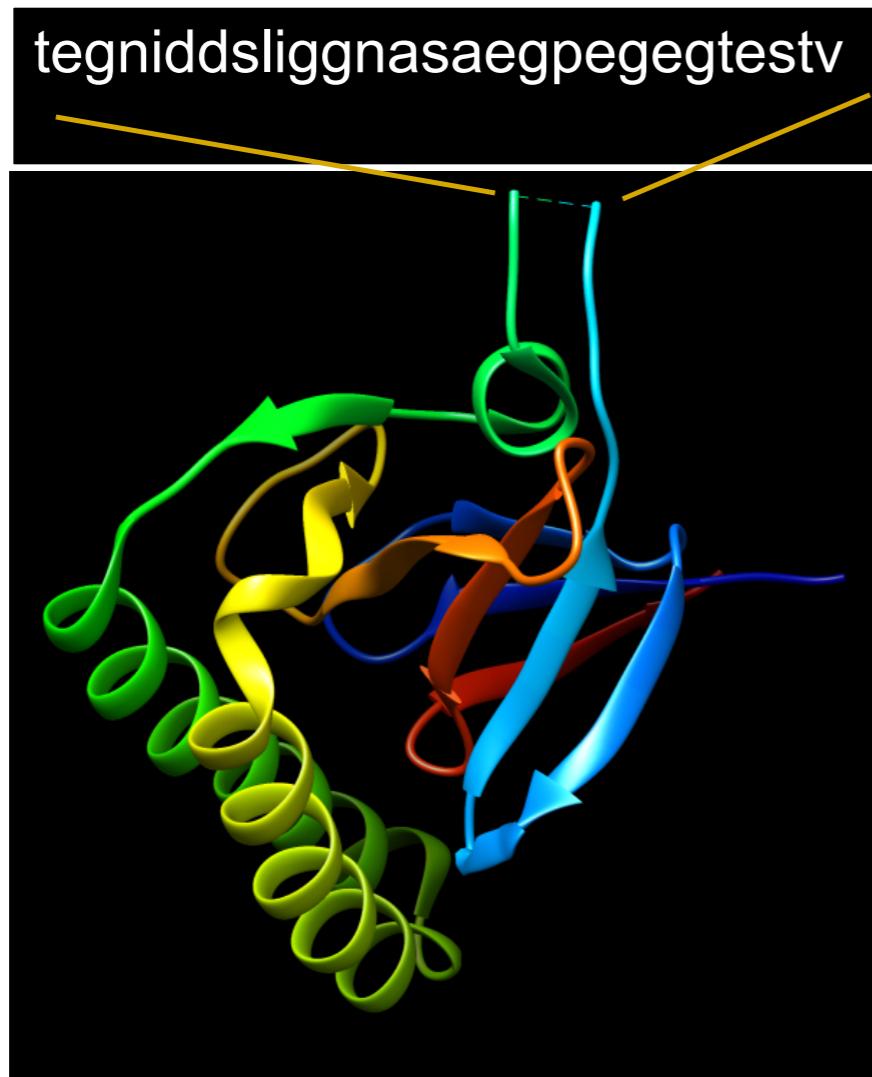
“Freaky” sequences ...

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



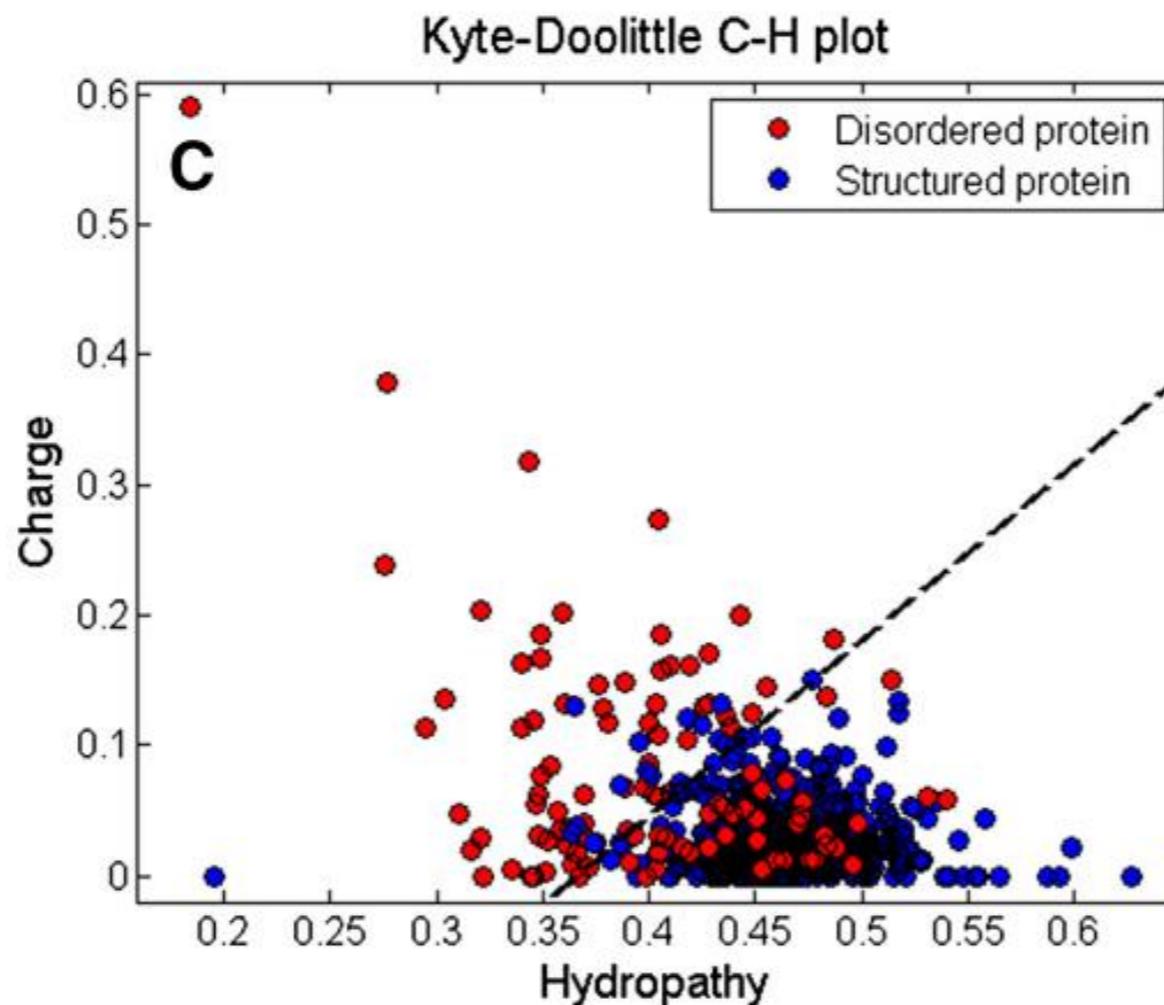
Missing electron density regions from
the PDB



NMR structures with large structural
variations

Sequence properties of IDPs

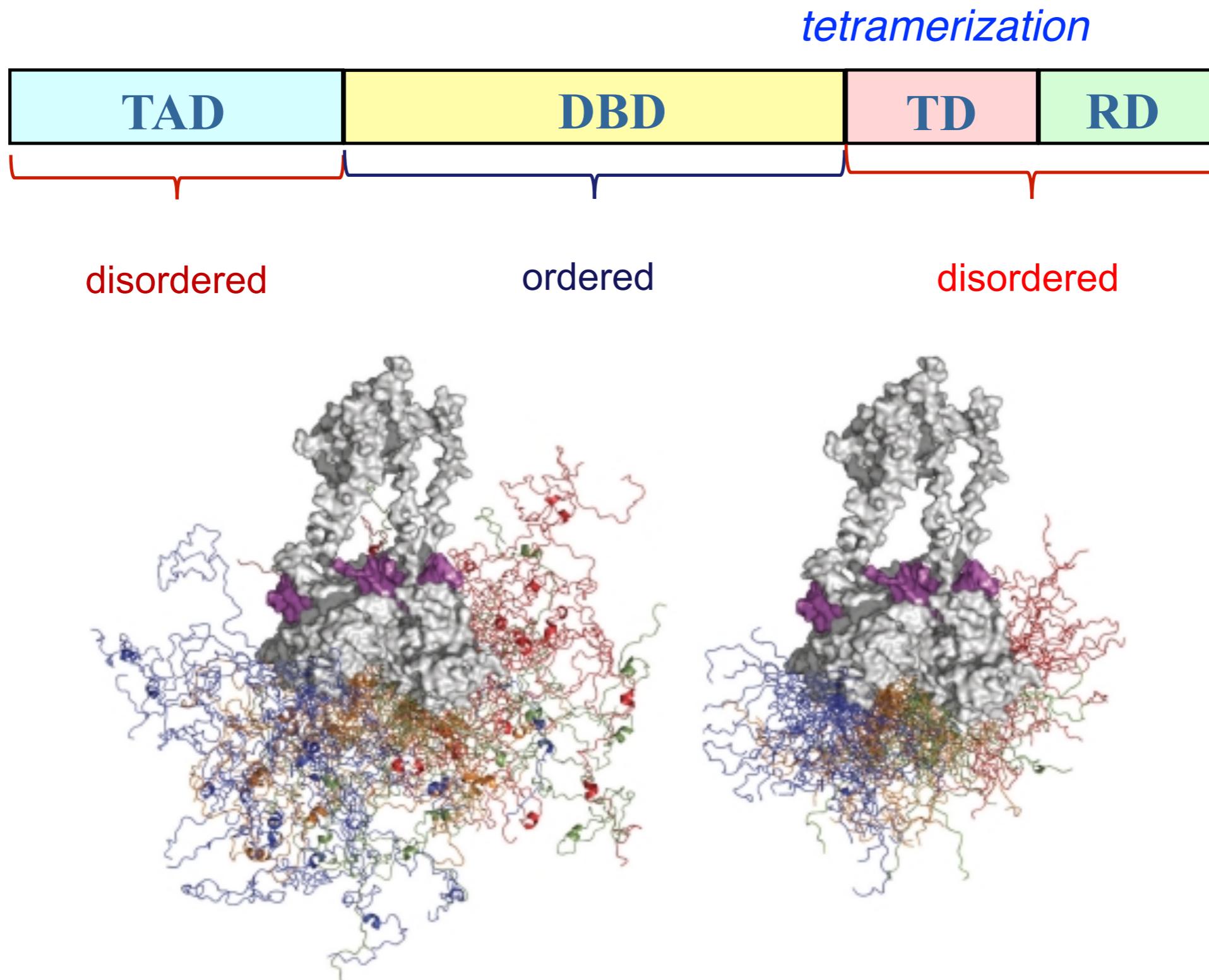
Charge-hydropathy plot



- 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



Where do we learn about “disordered regions”?



Prediction

*Experimental annotation:
803 proteins; 2167 regions*

MobiDB

a database of protein disorder and mobility annotations

About News Help Contact "P04637" Search

[MobiDB](#) [Disorder](#) [Uniprot](#) [Pfam](#) [String](#)

Protein overview

Names and details

UniProt ID	P04637	Seq. Length	393
Entry Name	P53_HUMAN	Organism	Homo sapiens
Protein Name	Cellular tumor antigen p53	Subcellular Loc.	Cytoplasm > Nucleus > Nucleus-PML body > Endoplasmic reticulum > Mitochondrion matrix

Tissue and function [-]

Tissue specificity	Ubiquitous. Isoforms are expressed in a wide range of normal tissues but in a tissue-dependent manner. Isoform 2 is expressed in most normal tissues but is not detected in brain, lung, prostate, muscle, fetal	Function	Acts as a tumor suppressor in many tumor types; induces growth arrest or apoptosis depending on the physiological circumstances and cell type. Involved in cell cycle regulation as a trans-activator that acts to negatively
--------------------	------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	----------	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Sequence annotations

Long Disorder			83.89 %
Disorder Sources [+]			83.89 %

Other [-]

Secondary Structure		
Pfam		

Legend [-]

Color Code	Disorder	Structure	Predicted Disorder	Predicted Structure	Ambiguous
	Helix	Strand	Tum		

Detailed disorder annotations [-]

DisProt [-]

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

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

Welcome to DisProt



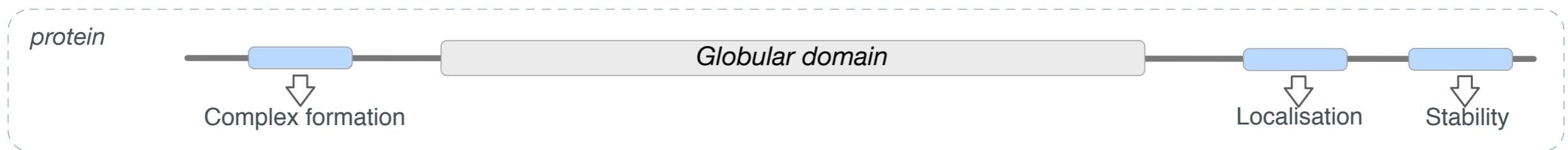
DisProt is a community resource annotating protein sequences for intrinsically disorder regions from the literature.

DISPROT: <http://www.disprot.org>

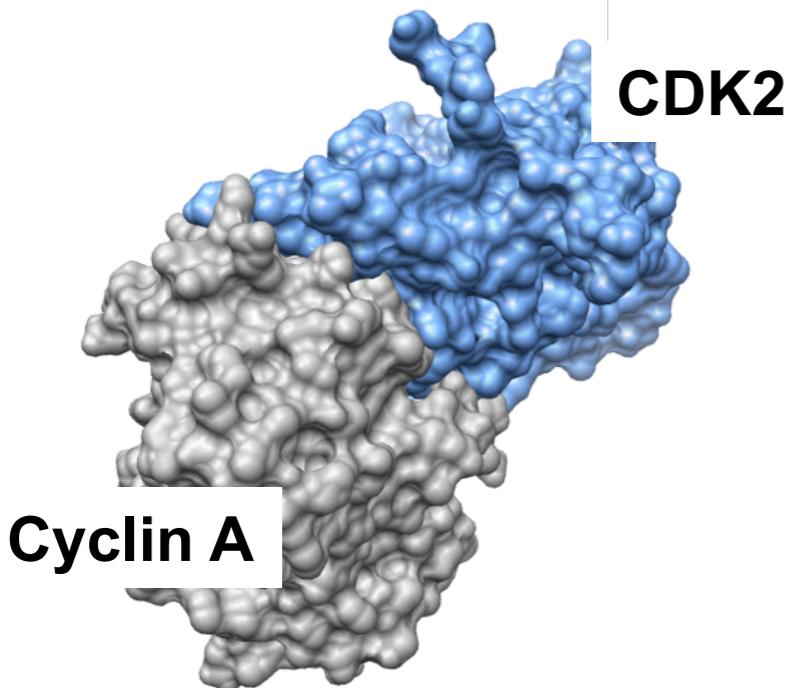
Session 3: Experimental investigation and prediction of intrinsically disordered regions

Function encoded within IDP regions: IDRs and SLiMs

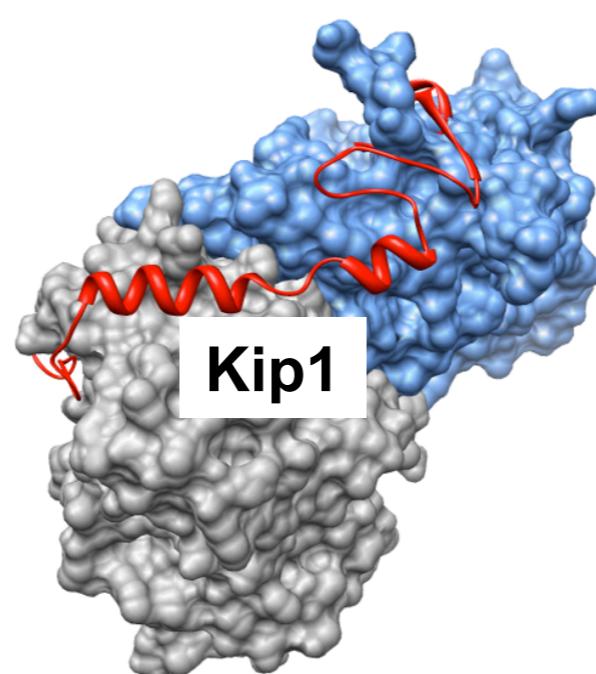
Protein Module



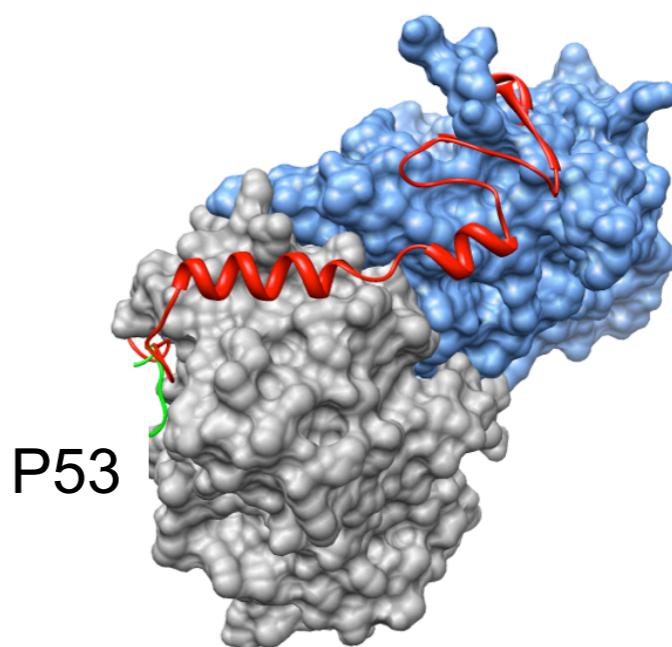
Globular Domain-Globular
Domain Interaction



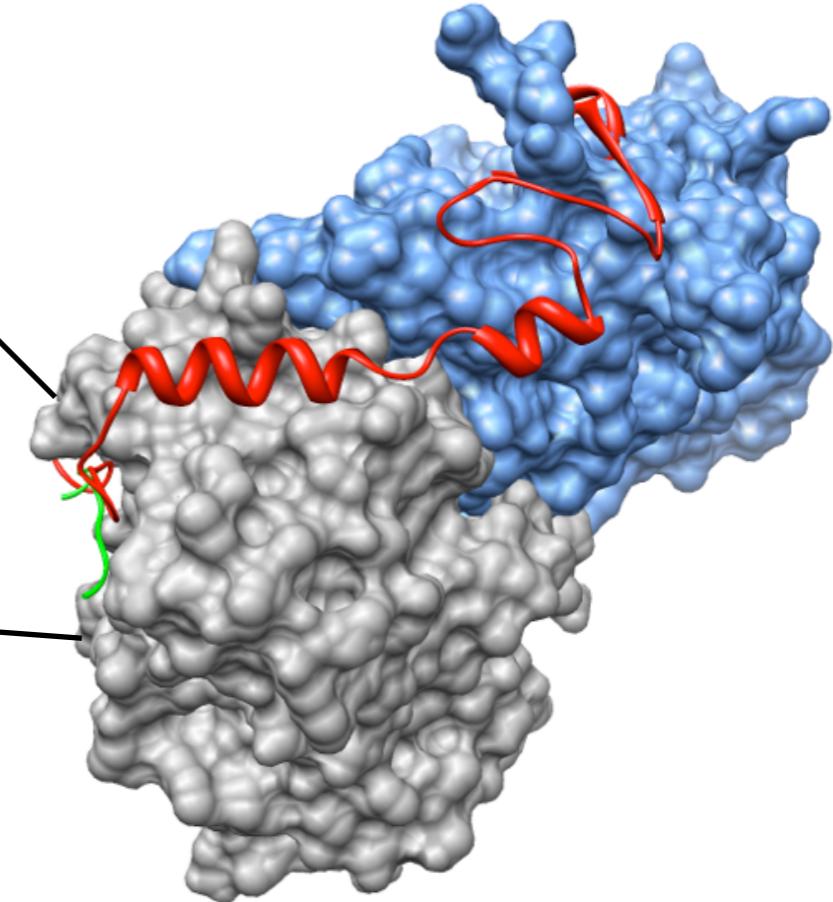
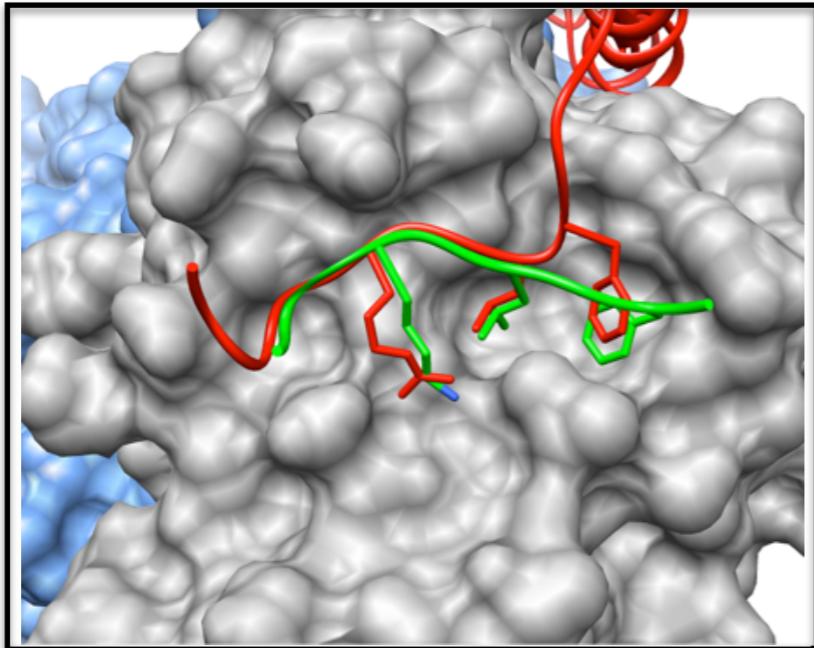
Globular Domain-Disordered
Domain Interaction



Globular Domain-Linear
Motif Interaction



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



- Motifs are usually defined as regular expressions

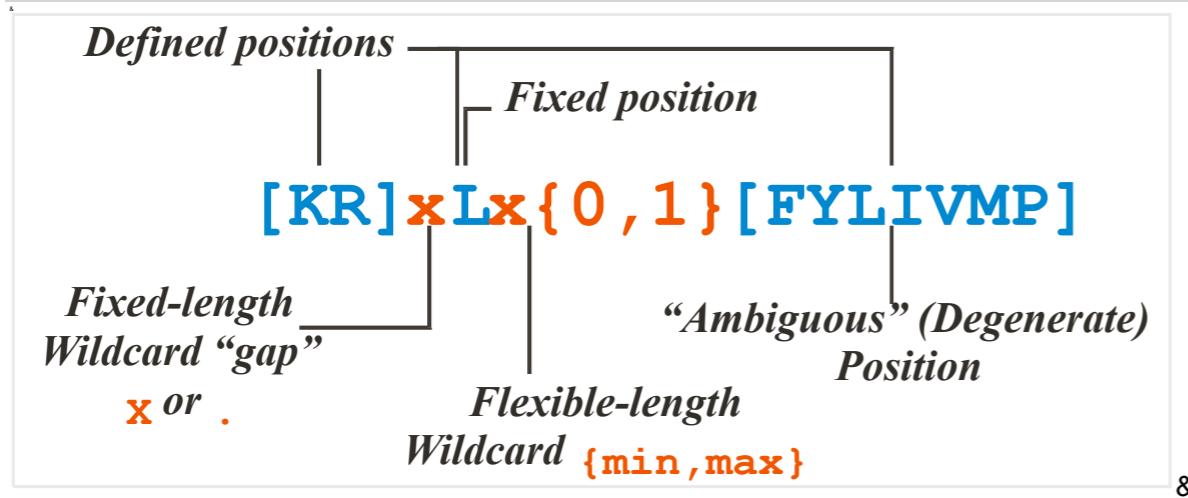


Figure 8.3. Anatomy of a SLiM.

Definitions of different properties of SLiM have been marked on the example ELM, LIG_CYCLIN_1 (Punnett 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

The Eukaryotic Linear Motif resource for Functional Sites in Proteins

search ELM Database

Help

ELM Home ELM Prediction ELM DB ELM Candidates ELM Information ELM downloads

»DEG_Kelch_KLHL3_1« »DEG_Nend_Nbox_1«

DEG_MDM2_SWIB_1

Accession: ELME000184

Functional site class: MDM2 binding motif

Functional site description: A degron motif found within the N-terminal p53 transactivation domain (TAD) ([PF08563](#)) and its relatives. The degron binds into a hydrophobic cleft in the N-terminal SWIB domain ([PF02201](#)) of the MDM2 E3 ubiquitin ligase. The sides of this pocket are formed by two helices, the bottom by two shorter helices and the ends are capped each by a three-stranded β -sheet (Kussie,1996). The p53 degron forms an amphipathic helix projecting a pair of aromatic residues deep into the MDM2 binding pocket. Regulation of p53 protein stability by Mdm2 is a key part of p53 function.

ELM Description: The MDM2-binding degron motif is located in the N-terminal transactivation domain (TAD) of p53 family members, so-called BOX-I ([PF08563](#)). The motif peptide folds as an amphipathic α -helix of about 2.5 turns, which binds in the hydrophobic cleft of the MDM2 SWIB domain (Kussie,1996). The three hydrophobic amino acids Phe-19, Trp-23 and Leu-26 are all found on the same side of the p53 degron helix and are critical for binding to MDM2 since they insert deeply into the binding pocket ([iYCR](#)). For example, substitution of residues Leu-22 and Trp-23 with Gln and Ser abolishes p53-MDM2 interaction, which leads to constitutively increased p53 levels ([Chehab,2000](#)). Since the motif adopts the α -helical fold when bound, proline residues are excluded from the non-conserved positions in the motif. When looking at p53 sequence alignments, there is sometimes an extra non-conserved residue between the Trp and Leu residues: This is likely to be due to underwinding of the end turn of the helix.

Pattern: F{^P}{3}W[^P]{2,3}[VIL]

Pattern Probability: 0.0000212

Present in taxon: Metazoa

Interaction Domain: SWIB / PRRs > SWIB / MDM2 domain (Structinmate: x + x)

PDB Structures: [iYCR](#)

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



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

Sequence

Sequence conservation can help identify *domains*

Sequence composition can help predict *disordered regions*

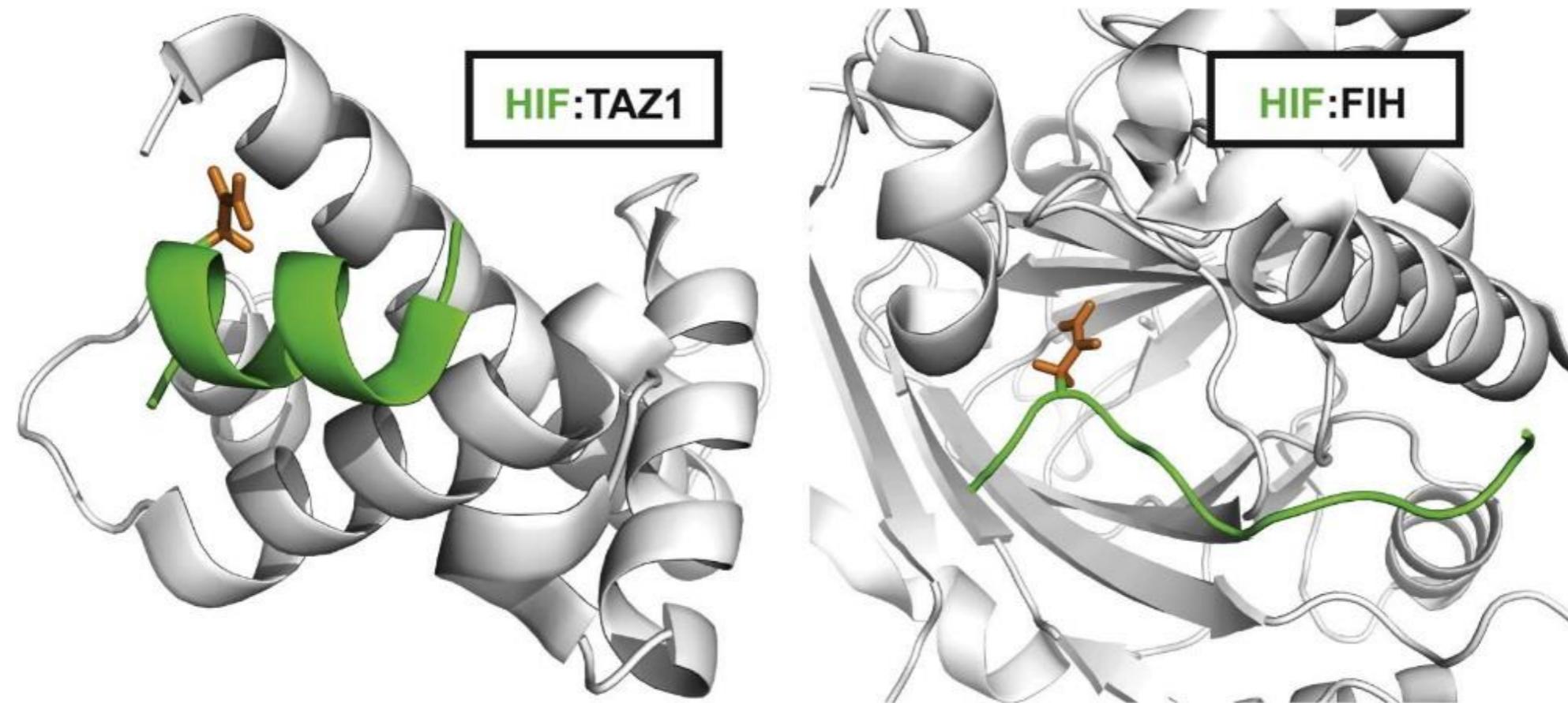
Sequence alignment can help identify other functional elements such as *Linear Motifs (SLiMs)*

Sequences carry valuable *evolutionary information*



<http://www.jalview.org>

Conformational plasticity



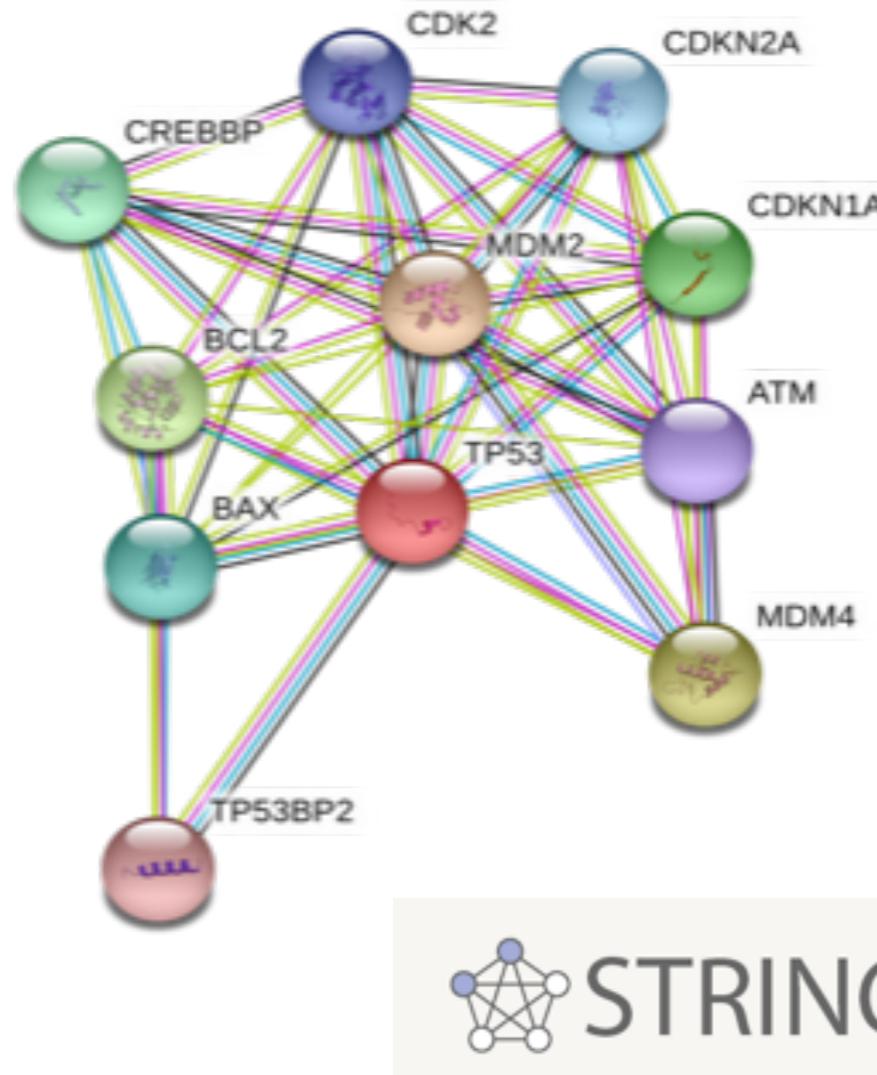
C-terminal transactivation domain (CTAD) of the hypoxia inducible factor-1 α

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

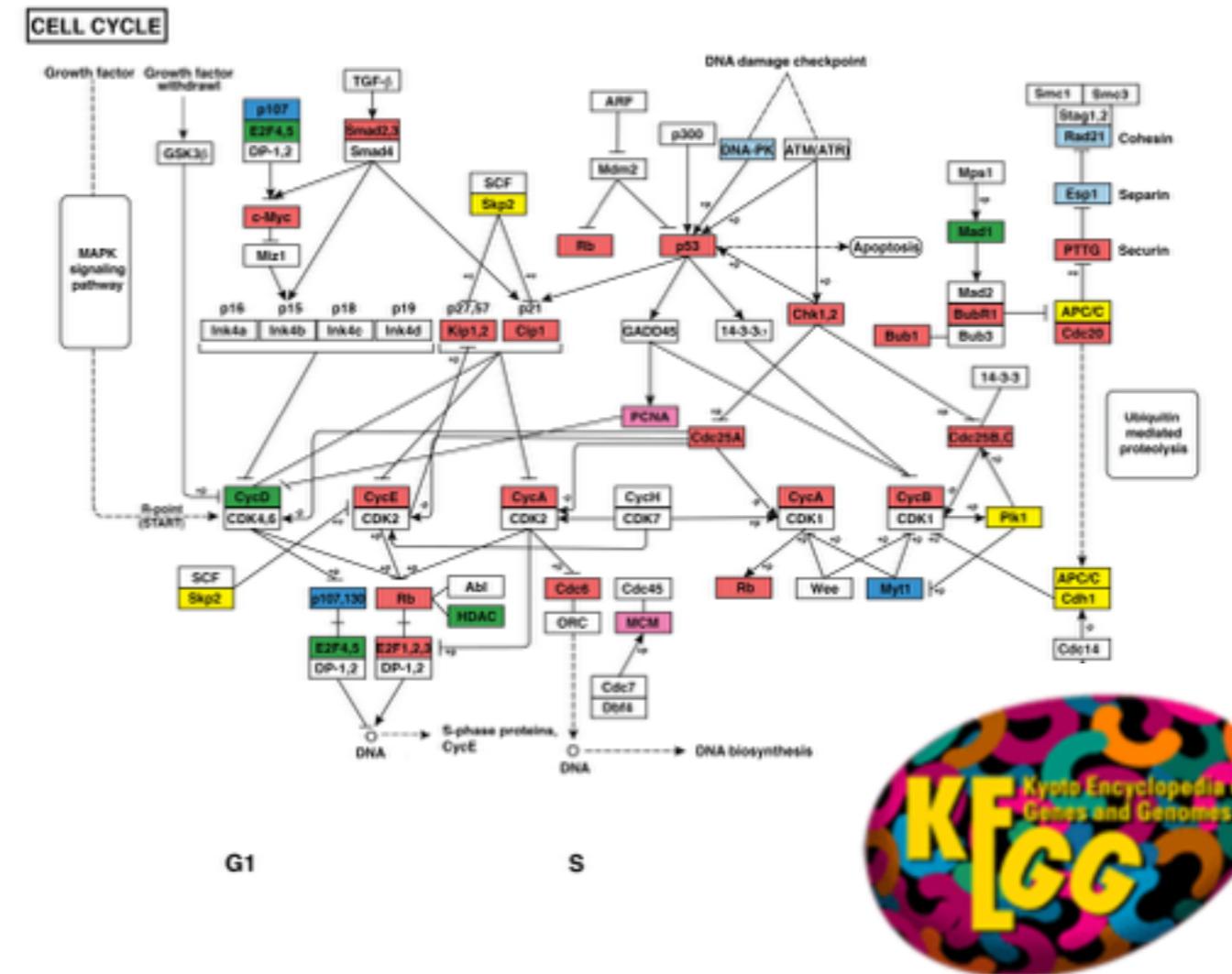
Session 6: Structural analysis using Chimera

Proteins work within highly connected “interaction networks”

“Molecular switching” changes the functional state of the cell



<https://string-db.org>



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

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