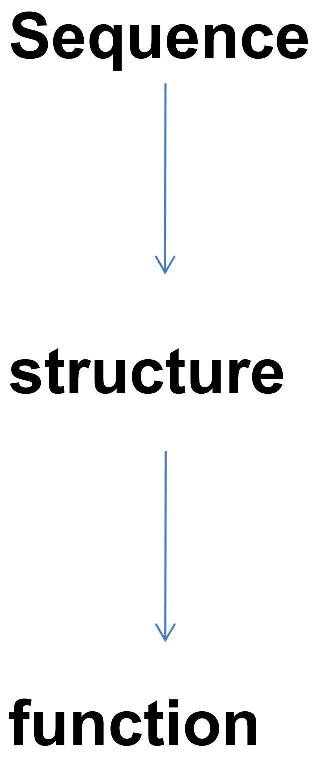
Protein linear motifs

Main Ref:

Frontiers in Bioscience 13:6580-6603.

Structure determines function?



Sequence

structure

function

Disorder proteins

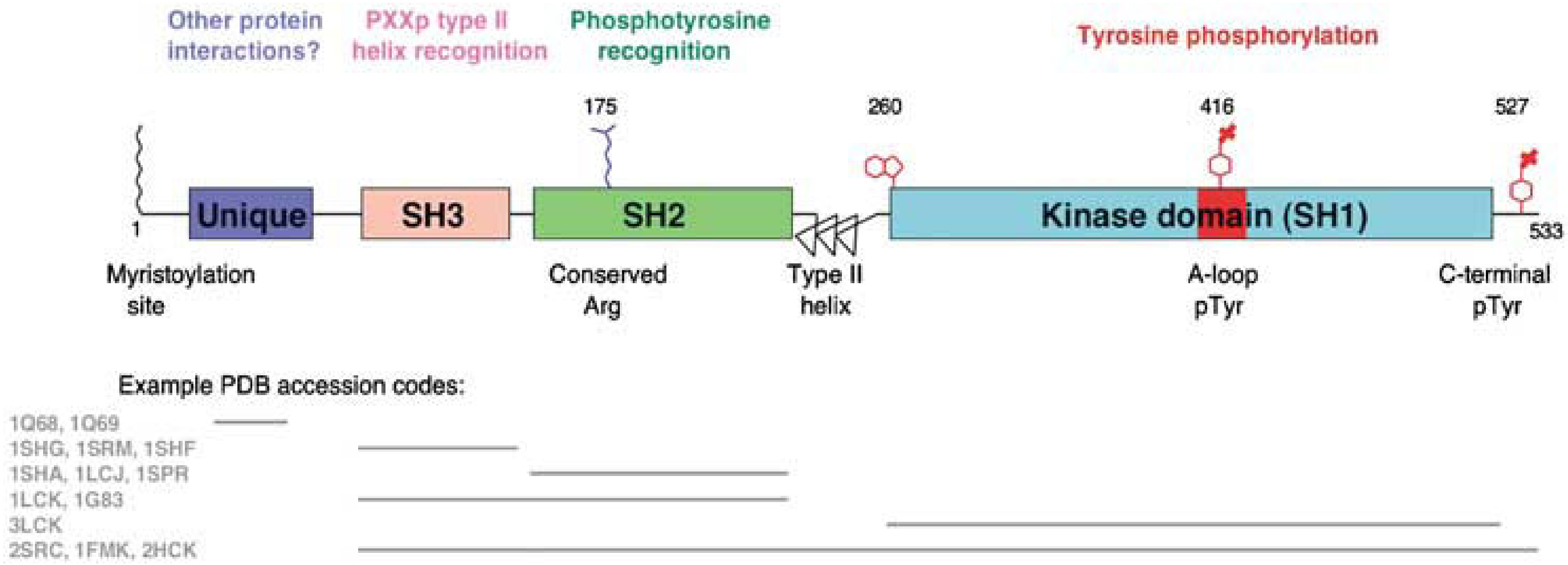
Intrinsically unstructured proteins, often referred to as *naturally unfolded proteins* or *disordered proteins*, are proteins characterized by lack of stable tertiary structure when the protein exists as an isolated polypeptide chain (a subunit) under physiological conditions in vitro.

The discovery of intrinsically unfolded proteins challenged the traditional protein structure paradigm, which states that a specific well-defined structure was required for the correct function of a protein and that the structure defines the function of the protein. This is clearly not the case for intrinsically unfolded proteins that remain functional despite the lack of a well-defined structure. Such proteins adopt fixed three dimensional structure only after binding to other macromolecules.

Disordered regions

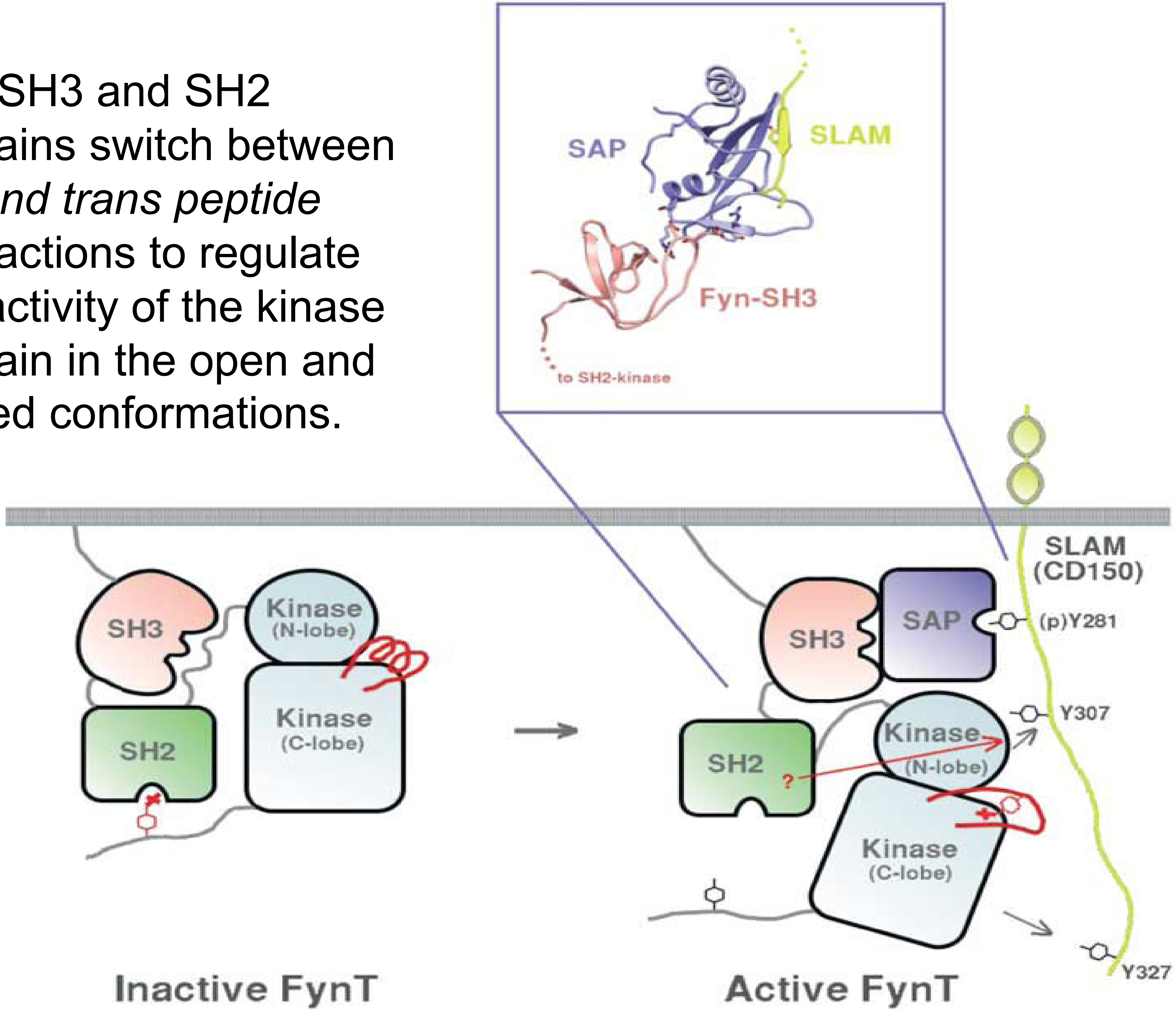
Disordered regions are often found as flexible linkers connecting two globular domains. Linker sequences vary greatly in length and amino acid sequence, but are similar in amino acid composition (rich in polar uncharged amino acids). Flexible linkers allow the connecting domains to freely twist and rotate through space to recruit their binding partners.

Src family kinase



The kinase domain acts on linear motifs containing a phosphorylatable tyrosine, the SH3 domain binds to proline-rich peptides, and the SH2 domain binds to phosphotyrosines.

The SH3 and SH2 domains switch between *cis and trans peptide* interactions to regulate the activity of the kinase domain in the open and closed conformations.

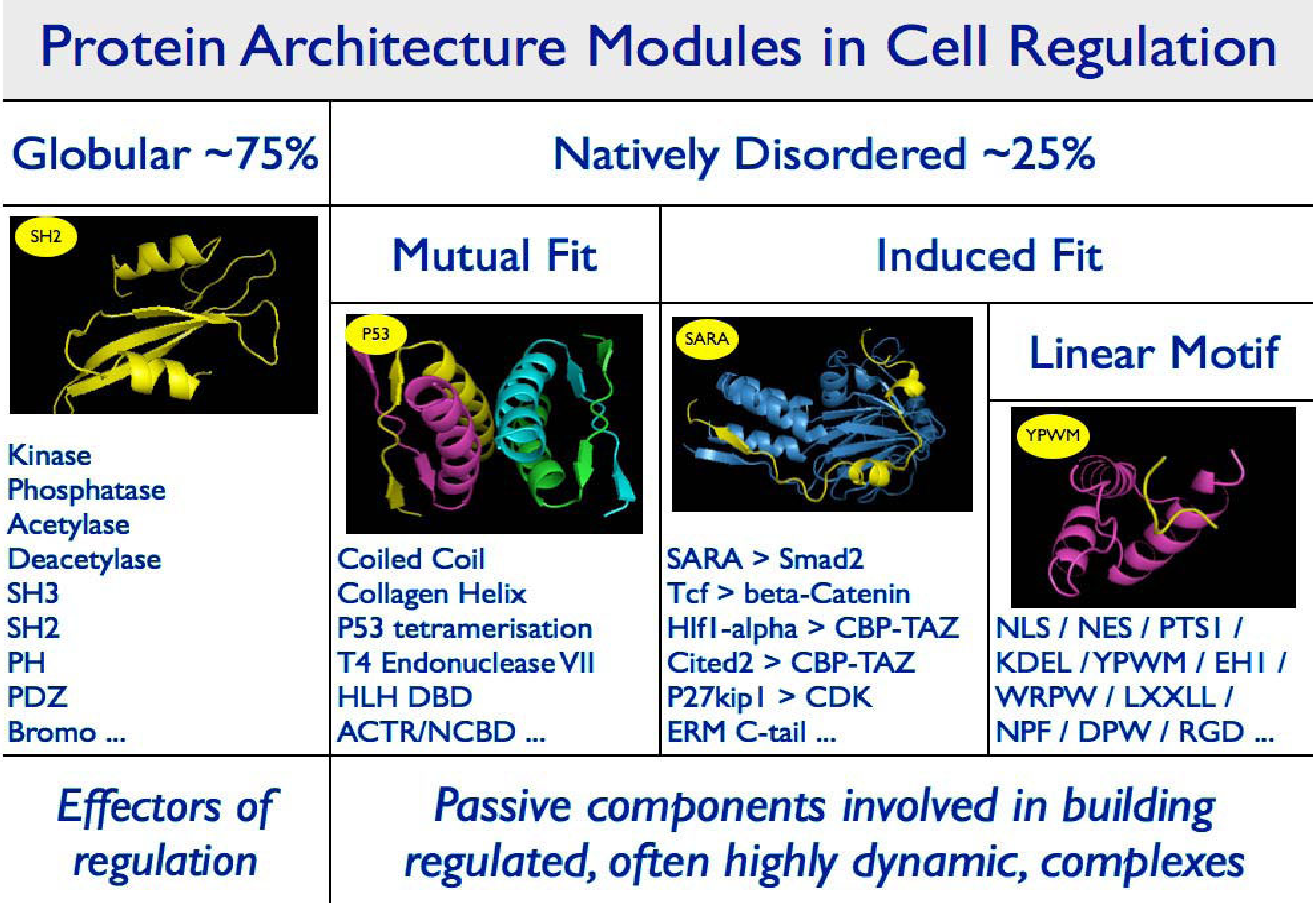


DisProt
http://www.disprot.org/

The Database of Protein Disorder (DisProt) is a curated database that provides information about proteins that lack fixed 3D structure in their putatively native states, either in their entirety or in part.

DisProt Release 7, 2017-11-05
DisProt Release 7.05 consists of 803 proteins entries and 2167 disordered regions.

Structural components of regulatory proteins



Protein Architecture Modules in Cell Regulation

Globular ~75%	Natively Disordered ~25%		
	Mutual Fit	Induced Fit	
			Linear Motif
Kinase Phosphatase Acetylase Deacetylase SH3 SH2 PH PDZ Bromo ...	Coiled Coil Collagen Helix P53 tetramerisation T4 Endonuclease VII HLH DBD ACTR/NCBD ...	SARA > Smad2 Tcf > beta-Catenin HIf1-alpha > CBP-TAZ Cited2 > CBP-TAZ P27kip1 > CDK ERM C-tail ...	NLS / NES / PTS1 / KDEL / YPWM / EH1 / WRPW / LXXLL / NPF / DPW / RGD ...
Effectors of regulation	*Passive components involved in building regulated, often highly dynamic, complexes*		

Protein Linear motif

- Linear motifs are short segments of multi-domain proteins that provide regulatory functions independently of protein tertiary structure.
- The essence of their function is embodied in the linear amino acid sequence and is not dependent on the tertiary structural context.

Another definition by Tim Hunter 1990

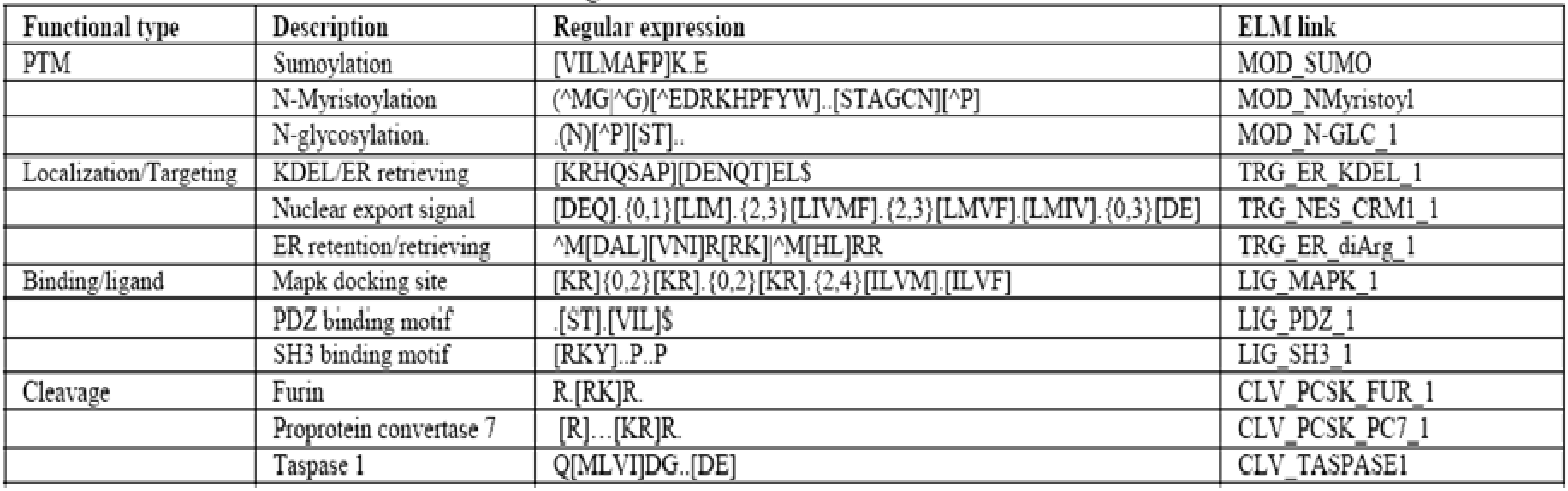
- The sequences of many proteins contain short, conserved motifs that are involved in recognition and targeting activities, often separate from other functional properties of the molecule in which they occur. These motifs are linear, in the sense that three-dimensional organization is not required to bring distant segments of the molecule together to make the recognizable unit. The conservation of these motifs varies: some are highly conserved while others, for example, allow substitutions that retain only a certain pattern of charge across the motif.

Properties of linear motifs

- typically 3 to 10 amino acids long
- predominantly found in regions of protein sequence that are obviously natively disordered
- linear motifs bind with lower affinity, usually between 1.0 and 150 micromolar e.g.
- Interaction with linear motifs in general is transient. as a consequence of low affinity binary binding interactions, they usually act in a concerted and cooperative manner, enabling regulatory decisions to be made on the basis of multiple inputs.
- in some cases a single linear motif interaction seems sufficient to mediate a given function, e.g. peroxisomal targeting via PTS1 (22), more often cooperativity among several motifs is required.
- a protein may contain different modification linear motifs that target the same amino acid residue for different PTMs
- Often many LMs are clustered within one segment of native disorder
- LMs quite frequently overlap, providing the potential for switch-like mutually exclusive functionality.

Classification of linear motifs

Table 2. Classification of linear motifs according to the ELM database.



Functional type	Description	Regular expression	ELM link
PTM	Sumoylation	[VILMAFP]K.E	MOD_SUMO
	N-Myristoylation	(^MG|^G)[^EDRKHPFYW]..[STAGCN][^P]	MOD_NMyristoyl
	N-glycosylation.	.(N)[^P][ST]..	MOD_N-GLC_1
Localization/Targeting	KDEL/ER retrieving	[KRHQSAP][DENQT]EL$	TRG_ER_KDEL_1
	Nuclear export signal	[DEQ].{0,1}[LIM].{2,3}[LIVMF].{2,3}[LMVF].[LMIV].{0,3}[DE]	TRG_NES_CRM1_1
	ER retention/retrieving	^M[DAL][VNI]R[RK]|^M[HL]RR	TRG_ER_diArg_1
Binding/ligand	Mapk docking site	[KR]{0,2}[KR].{0,2}[KR].{2,4}[ILVM].[ILVF]	LIG_MAPK_1
	PDZ binding motif	.[ST].[VIL]$	LIG_PDZ_1
	SH3 binding motif	[RKY]..P..P	LIG_SH3_1
Cleavage	Furin	R.[RK]R.	CLV_PCSK_FUR_1
	Proprotein convertase 7	[R]...[KR]R.	CLV_PCSK_PC7_1
	Taspase 1	Q[MLVI]DG..[DE]	CLV_TASPASE1

Pattern syntax

- Regular expression (REGEXP) language:
  - Each position is separated by a dash « - »
  - amino acids are represented by single letter code
  - « x » represent any amino acid
  - [] group of amino acid acceptable for a position
  - {} group of amino acid not acceptable for a position
  - () multiple or range e.g., A(1,3) means 1 to 3 A
  - < anchor at beginning of sequence
  - > anchor at end of sequence

Patterns / regular expression

- Pattern: <A-x-[ST](2)-x(0,1)-{V}
- Regexp: ^A.[ST](2).?[^V]
- Text: The sequence must start with an alanine, followed by any amino acid, followed by a serine or a threonine, two times, followed by any amino acid or nothing, followed by any amino acid except a valine.
- Simply the syntax differ…

Why use regexp

- Many LMs have short indels in the pattern. HMM software does not (yet) provide for variable gaps with exactly bounded ranges while ANNs do not account for gaps at all: a motif such as the NES with multiple short indels is hard to represent with these algorithms.
- The scoring of presence/absence matches for LM RegExps simplifies statistical analyses of motif searches.

Drawback of regexp

- RegExps cannot use position-weighting to capture weaker preferences. They are over-determined and can only capture exactly what is specified (whereas the more probabilistic HMMs and ANNs can rank near misses too).
- They do not support searching for an exact number of a given amino acid character within a specified range.

Conservation of linear motifs

- The ELM database contains experimentally validated motifs, the majority of annotations being drawn from human, mouse and other vertebrates. The majority of these (66%) are conserved only within vertebrates. However, a significant proportion (22%) is conserved across vertebrates, plants and yeast.
- Overall, the evolutionary stability of a linear motif is a function of its importance in the cellular network, and how that network has been evolving.

Multiple alignments to find LM

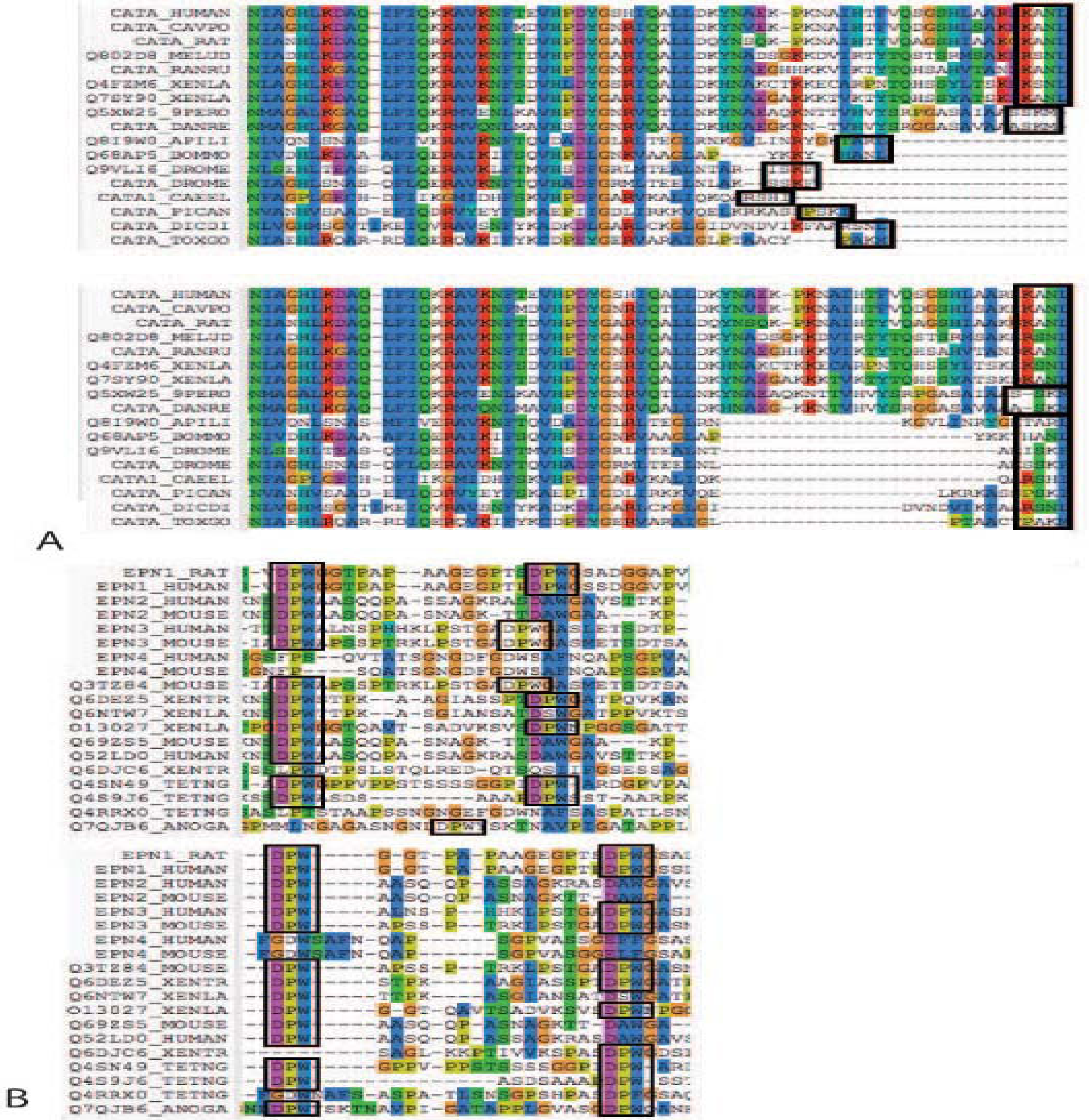


Table 4. List of methods to predict new instances and *de novo* motifs.

	Server	Method	Advantage	Disadvantage	url or reference
METHODS TO IDENTIFY NEW INSTANCES OF KNOWN MOTIFS	**PROSITE**	Regular expression matching	First attempt to catalogue LM	Stopped adding motifs due to high number of false positive matches. Currently the main focus is on globular domains	http://www.expasy.ch/prosite
	SCANSITE	Profile-based methods that uses data coming from oriented peptide library technique	>60 motifs. Quantitative representation of patterns is suitable for measuring features like motif specificity	Restricted to phosphorylation sites and motifs involved in signaling	http://scansite.mit.edu/
	ELM	Regular expression matching plus contextual filtering	Context-based rules and logical filters reduce the amount of false positives. >130 motifs are manually curated	Incomplete coverage of known motifs	http://elm.eu.org/
	Minimotif Miner	Regular expression matching plus contextual filtering	Large number of regular expressions	Motifs have little extra annotation	http://sms.engr.uconn.edu
	QuasiMotiFinder	Matching of patterns similar to PROSITE signatures plus evolutionary filtering	Evolutionary filtering reduces number of false predictions	Restricted to the set of motifs in PROSITE	(92)
	AutoMotifServer	Prediction of motifs based on trained support vector machine (SVM). Each type of PTM trained separately	The server predicts a good number of PTMs not present in other resources	The score assigned to the predicted instances is not biologically significant	http://ams2.bioinfo.pl/
	SIRW	Combine Regular expression with keyword search	Very intuitive method for prediction of new instances. Enrichment with GO terms can provide significant support	Low throughput interactive method	http://sirw.embl.de/
DE NOVO MOTIF DISCOVERY	**DILIMOT**	Identification of over-represented motifs in a set of proteins interacting with a target protein	First attempt at de novo motif prediction. Authors themselves found and tested new motifs	Only applicable to proteins present in interaction datasets. Only returns identities at motif conserved positions	http://dilimot.embl.de/
	SLiMFinder	Identification of over represented motifs in set of proteins, typically the set is an interaction dataset	Is able to retrieve motif matches with semi-conserved positions	Mainly applicable to proteins present in interaction datasets	http://bioinformatics.ucd.ie/shields/software/slimfinder/
	D-MOTIF/D-STAR algorithm	Detection of correlated (co-occurring) short sequence motifs	Improve detection from sparse and noisy interaction data	Rather stringent	(105)

Resources of linear motifs

- PROSITE (http://www.expasy.ch/prosite/) made the first systematic attempt to catalogue known motifs. The PROSITE database has collected a number of linear protein motifs, representing them as regular expression patterns. PROSITE patterns have been very useful, but also suffer from severe over prediction problems and more recently the database has focused on protein signature and domain annotation.
- SCANSITE (http://scansite.mit.edu/) (99) is a web-accessible tool that predicts motifs important in cellular signaling such as phosphorylation motifs or peptides binding to SH2 domains, 14-3-3 domains or PDZ domains. Each sequence motif is represented as a positionspecific scoring matrix (PSSM) based on results from oriented peptide library and phage display experiments.
- The Eukaryotic Linear Motif (ELM) resource (http://elm.eu.org/) (46) stores manually curated information about known linear motifs: it combines the use of regular expressions with logical filters (or rules), based on contextual information, to discriminate between likely true and false positives in order to improve the predictive value of ELM. The currently implemented context filters are a) taxonomic range filter, b) cell compartment filter, c) globular domain filter. In addition known ELM instances and predictions in sequences similar to ELM instance sequences, where the motif is positionally conserved, are identified and displayed.
- The Minimotif database (http://sms.engr.uconn.edu) contains 312 minimotifs extracted from the literature and other online resources and a web-based simple motif search system for identifying linear motifs in proteins. Homology analysis, surface prediction and frequency scores in complete proteomes are used to estimate the probability that the identified minimotifs are iologically functional.

5th in-class question

Please list at least two reasons that one may want to study RNA secondary structure. Please list at least two reasons that one should not study RNA secondary structure.