

Chapter 12

Protein analysis and proteomics

Outline

Introduction

Techniques for identifying proteins

Four perspectives on proteins

Perspective 1: Protein Domains and Motifs

Perspective 2: Physical Properties of Proteins

Introduction to Perspectives 3 and 4: Gene Ontology

Perspective 3: Protein Localization

Perspective 4: Protein Function

Learning objectives

Upon completing this material you should be able to:

- describe techniques to identify proteins including Edman degradation and mass spectrometry;
- define protein domains, motifs, signatures, and patterns;
- describe physical properties of proteins from a bioinformatics perspective;
- describe how protein localization is captured by bioinformatics tools; and
- provide definitions of protein function.

Protein databases

UniProt is a key database that includes UniProtKB/Swiss-Prot (~500,000 reviewed protein entries).

InterPro (<http://www.ebi.ac.uk/interpro/>) from the European Bioinformatics provides functional classification of proteins.

You can access UniProt, InterPro and many other protein databases through BioMart (web-based at www.ensembl.org) or the R package biomaRt.

biomaRt example 1: Given a list of gene symbols, what are the InterPro database identifiers and descriptions?

```
> getwd() # Confirm which directory you are working in
> source("http://bioconductor.org/biocLite.R")
> biocLite("biomaRt") # the package is now installed

> library("biomaRt") # load the package
> listMarts() # This displays >60 available databases
biomart
1 ensembl
2 snp
3 functional_genomics
4 vega
# additional Marts from this list of 60 are truncated.
> ensembl = useMart("ensembl")
> listDatasets(ensembl)
      dataset                                description
1  oanatinus_gene_ensembl  Ornithorhynchus anatinus genes (OANA5)
2  cporcellus_gene_ensembl  Cavia porcellus genes (cavPor3)
# This list is truncated.
```

biomaRt example 1: Given a list of gene symbols, what are the InterPro database identifiers and descriptions?

```
> ensembl = useDataset("hsapiens_gene_ensembl", mart=ensembl)
> filters = listFilters(ensembl)
> filters
> attributes = listAttributes(ensembl)
> attributes
# Browse the attributes to find protein-related topics!
# Let's select a small set of globin gene symbols
> globinsymbols <- c(HBB,HBA2,HBE,HBFB)
# Next let's do the search, sending the results to a file
# called myinterpro:
> myinterpro <-
getBM(attributes=c("interpro","interpro_description"),
filters="hgnc_symbol",values=globinsymbols, mart=ensembl)
> myinterpro # we print the results
```

	interpro	interpro_description
1	IPR000971	Globin
2	IPR002338	Haemoglobin, alpha
3	IPR002339	Haemoglobin, pi
4	IPR009050	Globin-like
5	IPR002337	Haemoglobin, beta

biomaRt example 2: Given a region of interest (e.g., 100,000 base pairs on chromosome 11) what are the gene symbols? For the genes that are protein-coding, which have predicted transmembrane regions?

```
> getBM(c("hgnc_symbol", "transmembrane_domain"),  
filters=c("chromosome_name", "start", "end"),  
values=list(11, 5200000, 5300000), mart=ensembl)
```

	hgnc_symbol	transmembrane_domain
1	OR52A1	Tm1hmm
2	OR51V1	Tm1hmm
3	HBB	
4	HBD	
5	HBD	Tm1hmm
6	HBG1	
7	HBG2	
8	HBE1	

The Human Proteome Organisation (HUPO) Proteomics Standards Initiative (PSI)

Goals: defining standards for proteomic data representation to facilitate the comparison, exchange, and verification of data

The Human Proteome Organisation (HUPO) Proteomics Standards Initiative (PSI)

Work groups

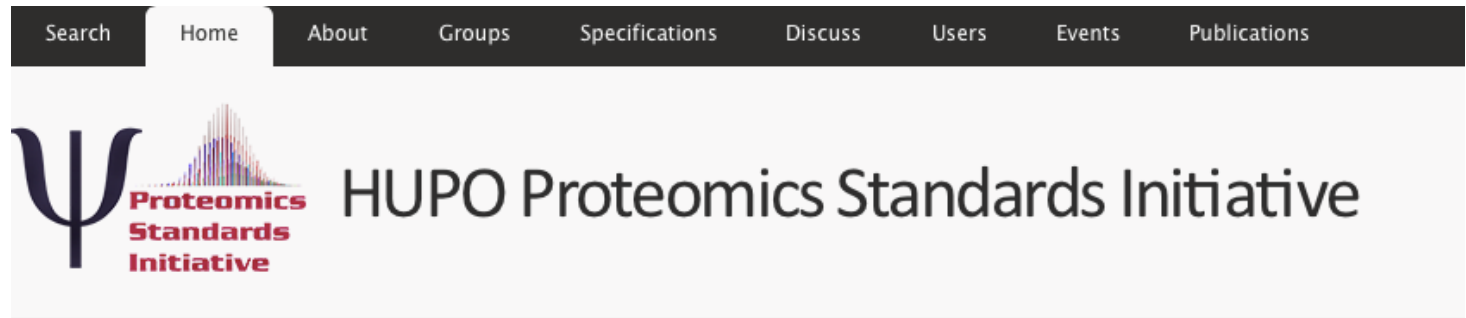
- # Gel Electrophoresis
- # Mass Spectrometry
- # Molecular Interactions
- # Protein Modifications
- # Proteomics Informatics
- # Sample Processing

Themes

- # Controlled vocabularies
- # MIAPE: Minimum information about a proteomics experiment

The Human Proteome Organisation (HUPO) Proteomics Standards Initiative (PSI)

<http://www.psidev.info/>



The HUPO Proteomics Standards Initiative defines community standards for data representation in proteomics to facilitate data comparison, exchange and verification.

HUPO-PSI Working Groups and Outputs

Working Groups	Guidelines	v.	Formats	v.	Controlled Vocabularies	v.
Molecular Interactions	MIMIx	1.1.2	PSI-MI XML (incl. MITAB)	2.5.4	PSI-MI CV	2.5.0
	MIABE	1.0.0				
	MIAPAR	1.0.0	PSI-PAR	1.0.0	PAR CV	n/a
Mass Spectrometry			mzML	1.1.0		
	Mass spectrometry (MIAPE_MS)	2.98	TraML	1.0.0		
			mzData	1.0.0		

Outline

Introduction

Techniques for identifying proteins

Four perspectives on proteins

Perspective 1: Protein Domains and Motifs

Perspective 2: Physical Properties of Proteins

Introduction to Perspectives 3 and 4: Gene Ontology

Perspective 3: Protein Localization

Perspective 4: Protein Function

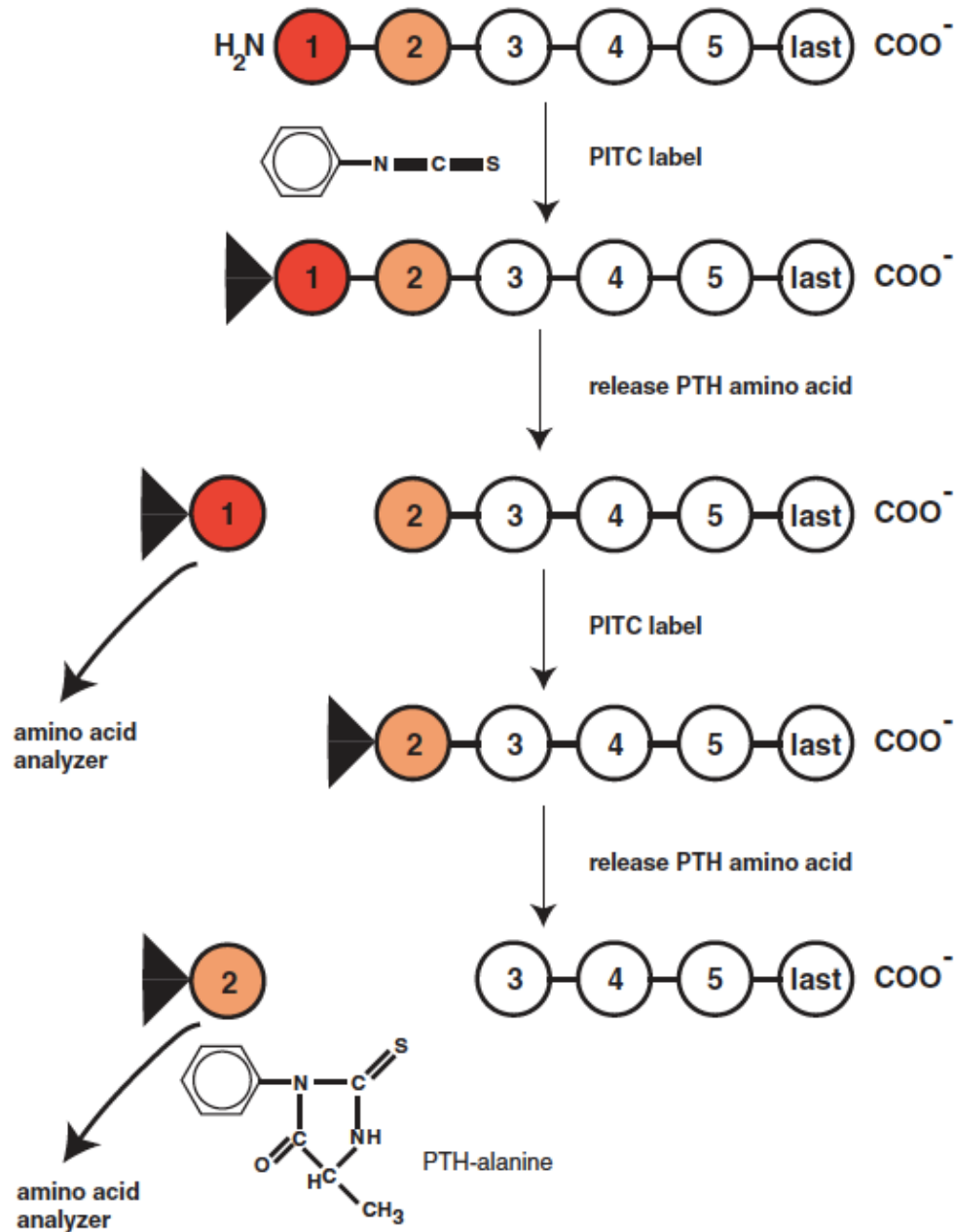


Protein sequencing by Edman degradation

Beginning in the 1940s Pehr Edman developed a method to determine the amino-terminal amino acid sequence of a peptide (protein).

The method involves modification of the N-terminal amino acid of a purified protein by phenylisothiocyanate, cleavage, and identification of the residue.

Protein sequencing by Edman degradation



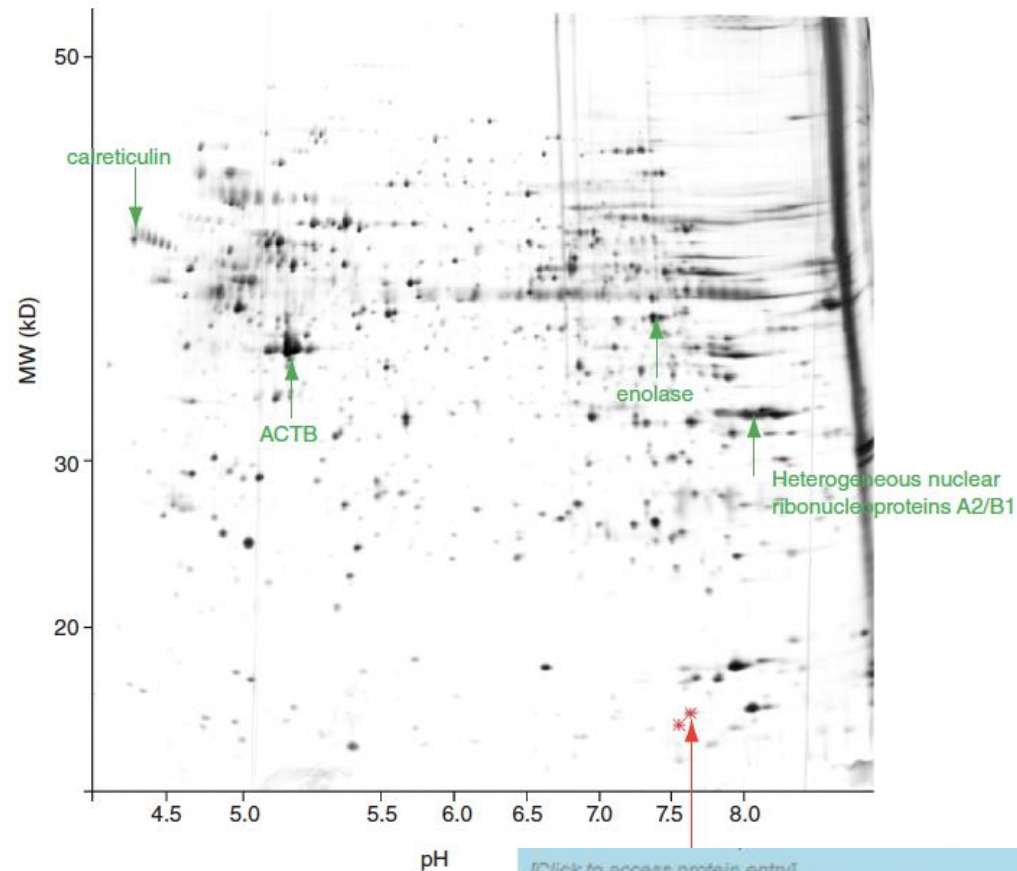
Polyacrylamide gel electrophoresis (PAGE)

Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) is useful to separate proteins based on molecular mass.

Two dimensional SDS-PAGE includes a second separation of proteins in the basis of charge: a protein migrates in an electric field to its isoelectric point, the pH at which the net charge is neutral.

Proteins on 1D or 2D SDS-PAGE can be visualized with dyes, identified with an antibody (Western blotting), sequenced by Edman degradation, or identified by mass spectrometry (MS).

Polyacrylamide gel electrophoresis (PAGE)



See 2D gels (SDS-PAGE, isoelectric focusing) at the ExPASy website. Mouse over a spot for information.

[Click to access protein entry]

Spot: **2D-001YG0** (lymphocyte_human)

pI: 7.63 Mw: 16594

%vol: 0.227604 %od: 0.198777

HBB_HUMAN

accession n°: P68871

Identification Methods:

*NORMAL LEVEL, MAPPING (PMF)

peptide masses: { (TRYPsin)

31126.6138 (0), 1274.7705 (0), 1314.7063 (0), 1378.7344 (0),

1669.9086 (0), 1778.9884 (0), 1797.9838 (0), 2074.9416 (0)

}

ExPASy offers many proteomics resources



ExPASy
 Bioinformatics Resource Portal

[Home](#)
[About](#)
[Contact](#)

Query all databases

Visual Guidance

Categories

proteomics

- protein sequences and identification
- mass spectrometry and 2-DE data
- protein characterisation and function
- families, patterns and profiles
- post-translational modification
- protein structure
- protein-protein interaction
- similarity search/alignment

genomics

structural bioinformatics

systems biology

phylogeny/evolution

population genetics

transcriptomics

biophysics

imaging


IT infrastructure

drug design

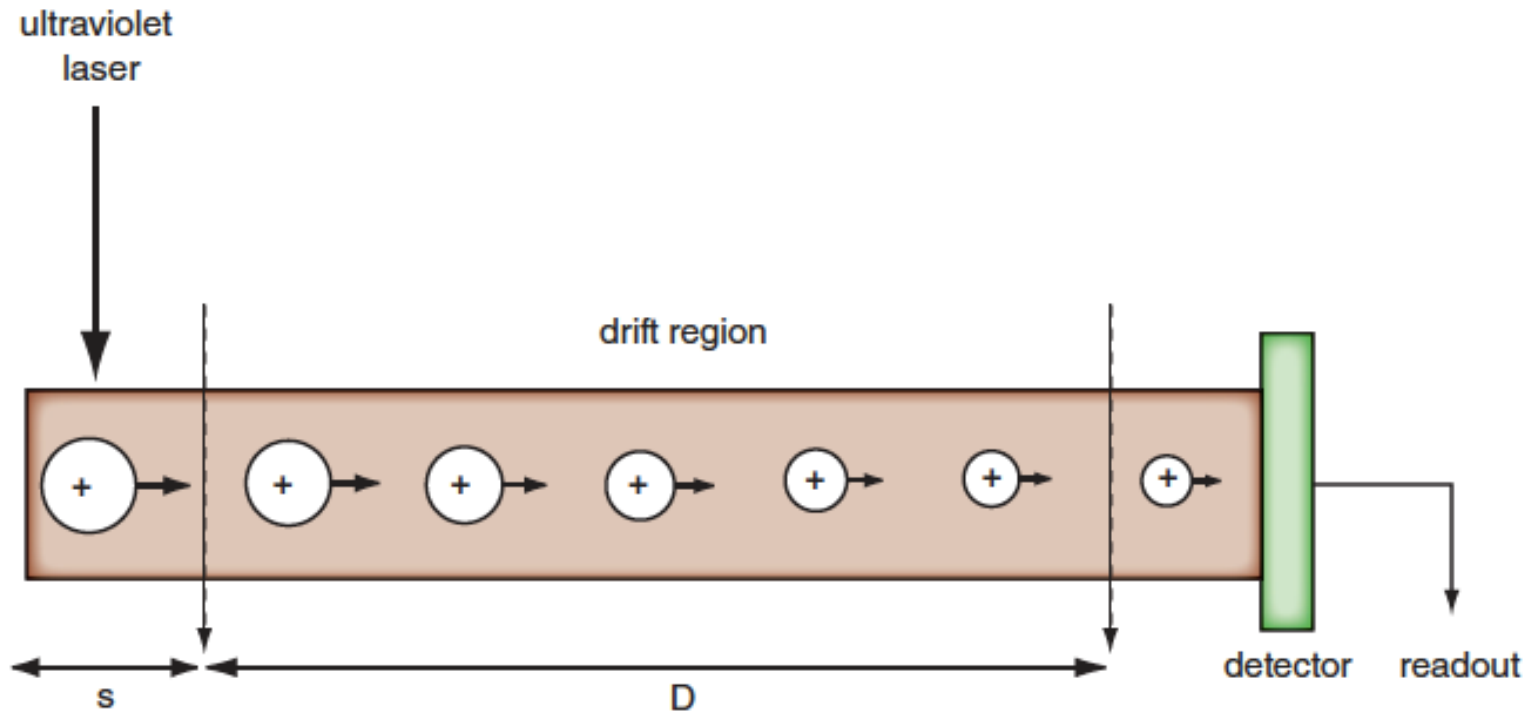
Resources A..Z

Examples: [\[show\]](#)

! Detected query type: text

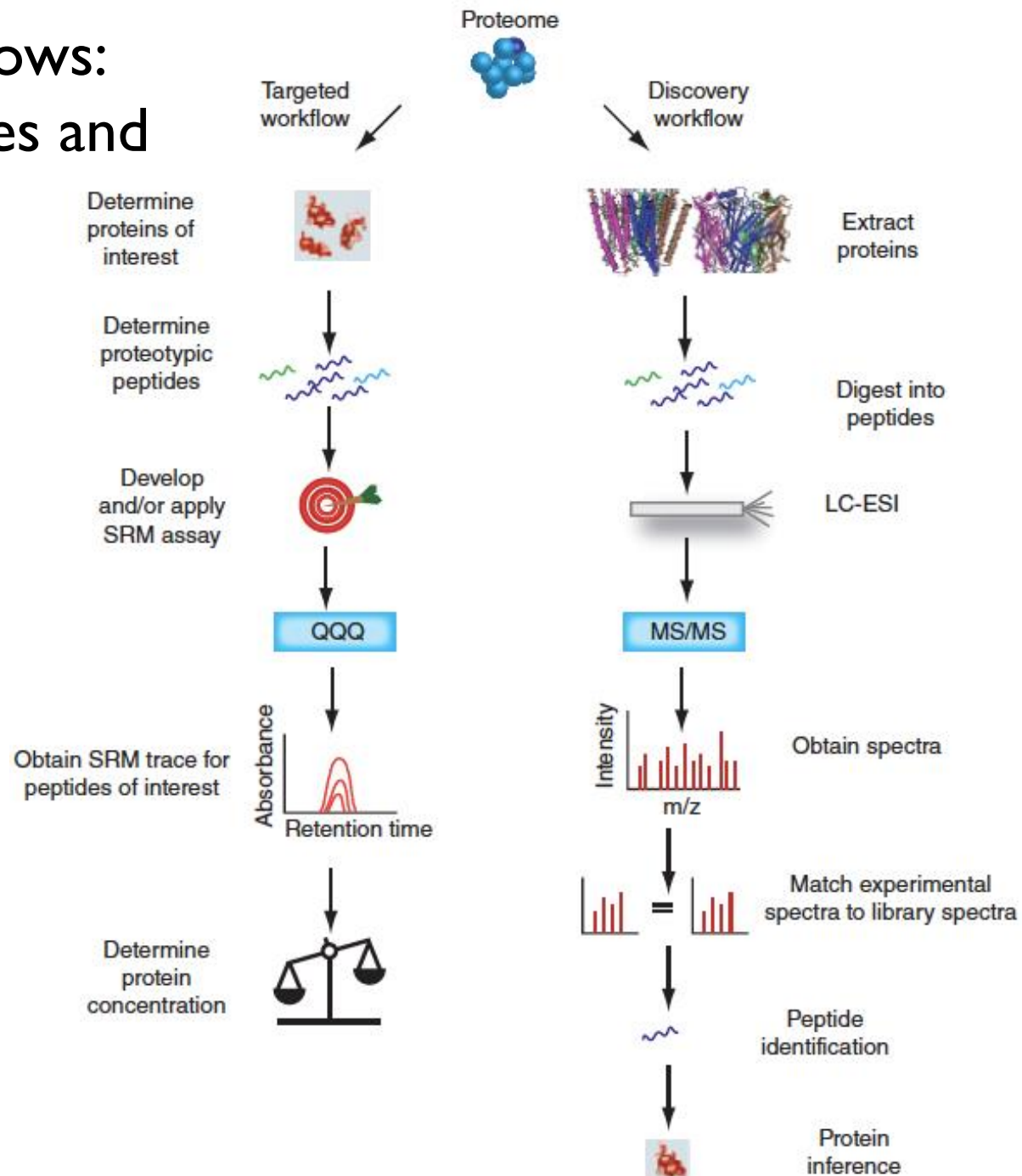
Resource	Hits	Category	Comment
 OMA	657 hits	ge, ph	
 PROSITE	7 hits	pr	PROSITE documentation entries
 STRING	415 hits	pr	
 SWISS-MODEL Repository	1348 hits	bi, pr, st	
 UniProtKB		pr	No valid response from server.
 ViralZone	0 hits	pr	ViralZone pages (for given virus name/family)
 ENZYME	7 hits	pr	ENZYME entries
 GPSDB	78 hits	ge, pr	gene synonyms
 HAMAP		pr	No valid response from server.
 miROrtho	0 hits	ge, ph	orthologous group(s) returned by your request
 MyHits	1000 hits	pr	Protein found: >1000; Motif found: 0; Other node found: 0;
 OpenFlu	0 hits	ge	
 OrthoDB	43 hits	ge, ph	in Metazoa; 28 in Bacteria; 23 in Arthropods; 10 in Vertebrates; 5 in Fungi
 Protein Spotlight	5 hits	pr	Spotlight/Prolune articles
 Selectome	157 hits	ph	results in Selectome
 SWISS-2DPAGE	04 hits	pr	SWISS-2DPAGE entries
 SwissVar	13 hits	pr	Swiss-Prot protein - variant - disease groups

Matrix-assisted laser desorption/ionization time-of-flight spectroscopy (MALDI-TOF)



Mass spectrometry (MS) enables sensitive identification of proteins

Two MS workflows: targeted analyses and discovery



PRIDE at EBI: database for mass spectrometry

(a) PRIDE search results for mass spectrometry datasets including P68871 (beta globin)

Accession	Title	Species	Tissue	Cell Type	GO Term	Disease	Protein Count	Peptide Count	Spectra Count	Retrieve Details (View in web browser or download as XML file)
193	Plasma Proteome (GPM10100001689)	Homo sapiens (Human)	-	-	-	-	1	4	0	Web View PRIDE Inspector * Download
8959	Human Hep3B cells, untreated, cytoplasmic fraction	Homo sapiens (Human)	HEP-3B cell, liver	hepatocyte	cytoplasm	-	1	1	1	Web View PRIDE Inspector * Download
19112	Human Occipital Lobe (BA17)	Homo sapiens (Human)	-	-	-	-	1	26	22	Web View PRIDE Inspector * Download
25907	The proteome of mononuclear cells from human blood 2	Homo sapiens (Human)	mononuclear cell, blood	-	-	-	1	10	10	Web View PRIDE Inspector * Download

PRIDE at EBI: database for mass spectrometry

(b) PRIDE Inspector software 1.3.2

Overview Protein (222) Peptide (2554) Spectrum (55955) Quantification Summary Charts (8)

Protein: Type: Gel Free Obtain Protein Details Decoy Filter Shared Peptides Disclaimer

#	Submitted	Mapped	Protein Name	Status	Coverage	Score	Threshold	# Peptides	# Distinct Peptides	# PTMs	More
5	P04406	P04406	Glyceraldehyde-3-phosphate dehydr...	ACTIVE	75.2%	0.0	0.0	37	35	0	More
6	P68371	P68371	Tubulin beta-4B chain (Tubulin beta...	ACTIVE	46.5%	0.0	0.0	56	50	0	More
7	P06576	P06576	ATP synthase subunit beta, mitoch...	ACTIVE	68.1%	0.0	0.0	42	35	0	More
8	P07437	P07437	Tubulin beta chain (Tubulin beta-5 c...	ACTIVE	46.8%	0.0	0.0	50	49	0	More
9	Q13885	Q13885	Tubulin beta-2A chain (Tubulin beta...	ACTIVE	49.9%	0.0	0.0	53	48	0	More
10	P30086	P30086	Phosphatidylethanolamine-binding ...	ACTIVE	64.7%	0.0	0.0	18	16	0	More
11	Q16555	Q16555	Dihydropyrimidinase-related protein ...	ACTIVE	40.3%	0.0	0.0	41	37	0	More
12	P68871	P68871	Hemoglobin subunit beta (Beta-glob...	ACTIVE	67.3%	0.0	0.0	26	24	0	More
13	P02042	P02042	Hemoglobin subunit delta (Delta-glob...	ACTIVE	39.5%	0.0	0.0	22	20	0	More
14	P68892	P68892	Hemoglobin subunit gamma-2 (Ga...	ACTIVE	15.6%	0.0	0.0	5	4	0	More

Peptide [P68871] PTM: NONE

#	Peptide	Fit	Charge	Delta m/z	Precursor m/z	# PTMs	PTM List	# Ions	Length	Start	Stop	More
1	VHLTPEEK	Unknown	1	259.0914	1211.6	0		6	8	2	9	More
2	HLTPEEK	Unknown	1	2698.1764	1461.62	0		0	7	3	9	More
3	SVYALVQK	Unknown	1	1615.7408	2548.26	0		5	9	10	18	More
4	VNVDEVGCEALDR	Unknown	1	158.0462	1472.71	0		1	13	19	21	More
5	LLV	Unknown	1	2025.8735	2468.4	0		7	4	32	36	More
6	LLVYVPWTQR	Unknown	1	480.1745	1754.9	0		15	10	32	41	More
7	LLVYVPWTQR	Unknown	1	-691.4181	583.31	0		1	10	32	41	More
8	VYVPWTQR	Unknown	1	141.12	907.44	0		0	8	34	41	More
9	FFESFGDLSTPDVAVGNPK	Unknown	1	493.515	2552.46	0		7	19	42	60	More
10	FDVAVGNPK	Unknown	1	2135.9619	3054.42	0		0	9	52	60	More
11	VNVDEVGCEALDR	Unknown	1	933.0632	1024.04	0		7	13	17	21	More

Spectrum Fragmentation Table Sequence

Accession: P68871, Name: Hemoglobin subunit beta (Beta-globin) [Hemoglobin beta chain] [Cleaved into: LVV hemorphin 7; Spinothrin]

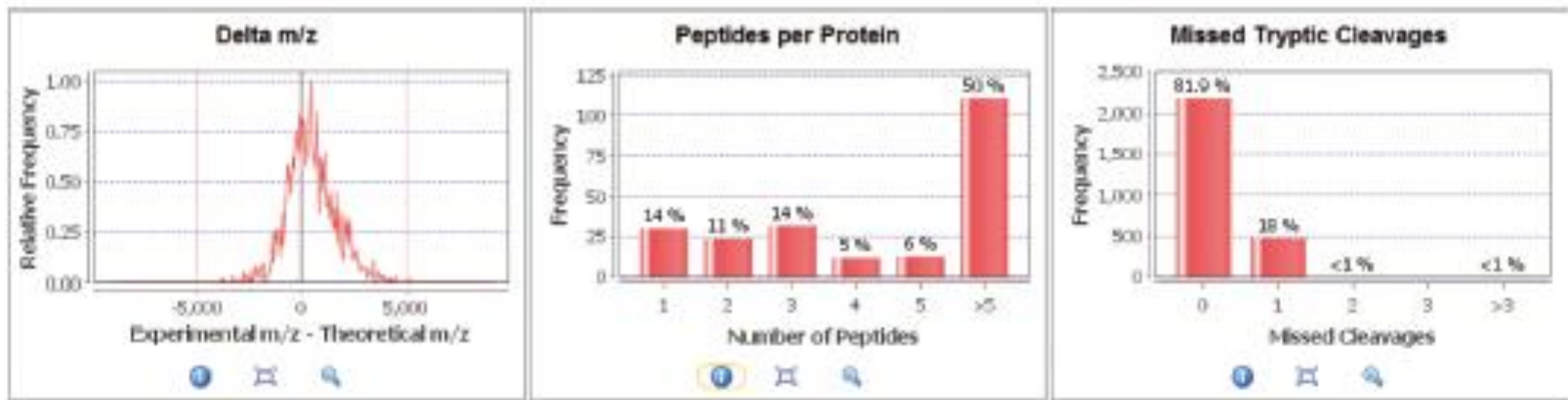
26 peptides (26 matched, 24 distinct), 99/147 amino acids (67.3% coverage)

MVHLTPEEKS	AVTALMCKVN	VDEVGCEALG	RLLVYVPWTQ	FFESFGDLR	TDDAVGCHDK	VKAKCKKVLG	RFSDGLAHLR	20
HLTPFATLR	ELKCKLHVG	PEVFLLGNV	LVCVLAKHFG	KEFTFPVQAA	YQKVVAGVAN	ALAKKYN		147

Legend: Selected (Yellow), PTM (Pink), Fit (Green), Fuzzy Fit (Orange), Overlap (Dark Green)

PRIDE at EBI: database for mass spectrometry

(c) PRIDE Inspector summary charts



Outline

Introduction

Techniques for identifying proteins

Four perspectives on proteins

Perspective 1: Protein Domains and Motifs

Perspective 2: Physical Properties of Proteins

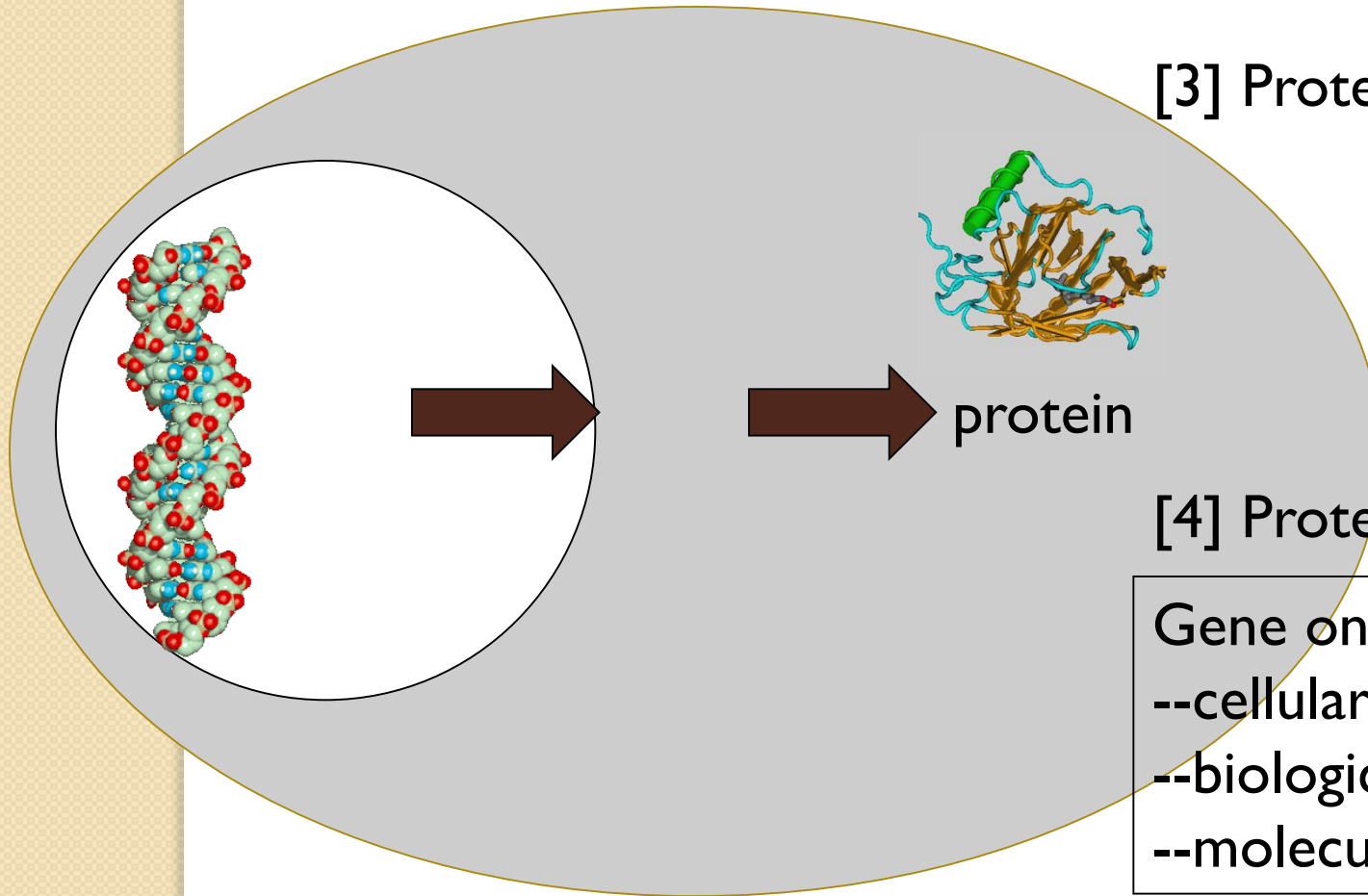
Introduction to Perspectives 3 and 4: Gene Ontology

Perspective 3: Protein Localization

Perspective 4: Protein Function

[1] Protein families

[3] Protein localization



[4] Protein function

Gene ontology (GO):

- cellular component
- biological process
- molecular function

[2] Physical properties

Perspective I: Protein domains and motifs

Definitions

Signature:

- a protein category such as a domain or motif

Definitions

Signature:

- a protein category such as a domain or motif

Domain:

- a region of a protein that can adopt a 3D structure
- a fold
- a family is a group of proteins that share a domain
- examples: zinc finger domain
 immunoglobulin domain

Zinc finger proteins are among the most abundant proteins in eukaryotic genomes. Their **functions** are extraordinarily diverse and include DNA recognition, RNA packaging, transcriptional activation, regulation of apoptosis, protein folding and assembly, and lipid binding.

Motif (or fingerprint):

- a short, conserved region of a protein
- typically 10 to 20 contiguous amino acid residues

Definitions from the InterPro database at EBI

Term	Definition
Family	A protein family is a group of proteins that share a common evolutionary origin reflected by their related functions, similarities in sequence, or similar primary, secondary or tertiary structure. A match to an InterPro entry of this type indicates membership of a protein family.
Domain	Domains are distinct functional, structural, or sequence units that may exist in a variety of biological contexts. A match to an InterPro entry of this type indicates the presence of a domain.
Repeat	A match to an InterPro entry of this type identifies a short sequence that is typically repeated within a protein.
Site	A match to an InterPro entry of this type indicates a short sequence that contains one or more conserved residues. The type of sites covered by InterPro are active sites, binding sites, post-translational modification sites, and conserved sites.

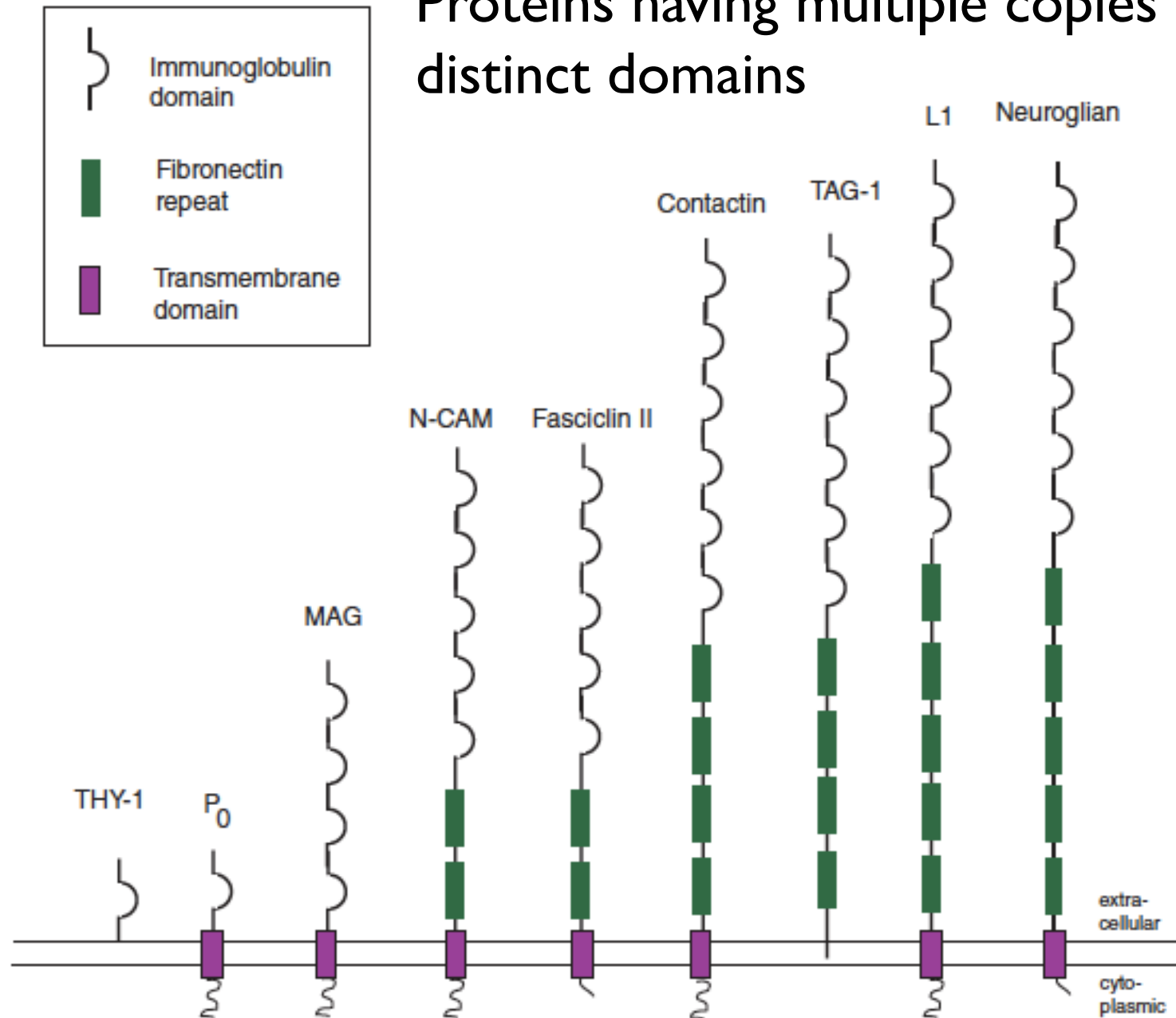
Source:  <http://www.ebi.ac.uk/interpro/>.

10 most common domains (human)

InterPro accession	Proteins matched	Name of domain
IPR027417	1022	P-loop containing nucleoside triphosphate hydrolase
IPR007110	1015	Immunoglobulin-like domain
IPR007087	806	Zinc finger; C2H2
IPR015880	801	Zinc finger; C2H2-like
IPR017452	796	GPCR; rhodopsin-like; 7TM
IPR000276	789	G protein-coupled receptor; rhodopsin-like
IPR003599	623	Immunoglobulin subtype
IPR013106	619	Immunoglobulin V-set
IPR011009	560	Protein kinase-like domain
IPR000719	513	Protein kinase; catalytic domain

Source: InterPro (2015)

Proteins having multiple copies of distinct domains



Definition of a domain

According to InterPro at EBI (<http://www.ebi.ac.uk/interpro/>):

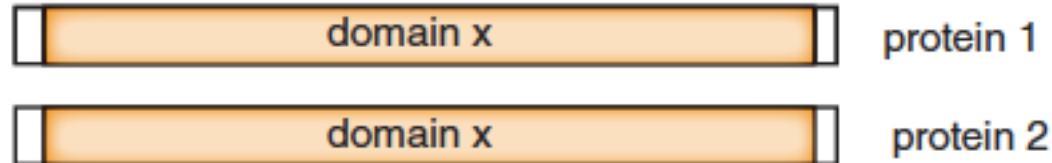
A domain is an independent structural unit, found alone or in conjunction with other domains or repeats.
Domains are evolutionarily related.

According to SMART (<http://smart.embl-heidelberg.de>):

A domain is a conserved structural entity with distinctive secondary structure content and a hydrophobic core.
Homologous domains with common functions usually show sequence similarities.

Varieties of protein domains

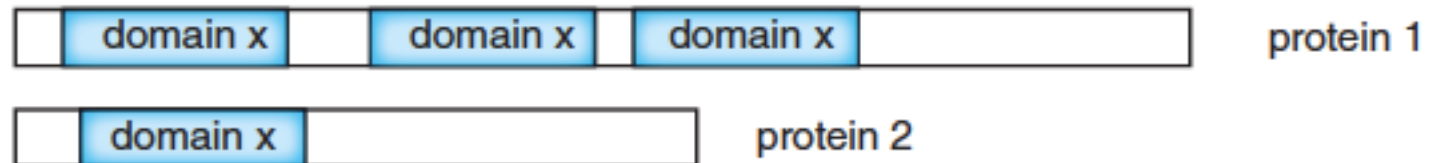
Extending along the length of a protein



Occupying a subset of a protein sequence



Occurring one or more times



Example of a protein with domains: Methyl CpG binding protein 2 (MeCP2)



The protein includes a methylated DNA binding domain (MBD) and a transcriptional repression domain (TRD). MeCP2 is a transcriptional repressor.

Mutations in the gene encoding MeCP2 cause Rett Syndrome, a neurological disorder affecting girls primarily.

MECP2 is a gene that encodes the protein MECP2. MECP2 appears to be essential for the normal function of nerve cells. The protein seems to be particularly important for mature nerve cells, where it is present in high levels. The MECP2 protein is likely to be involved in turning off several other genes.

MeCP2 486 aa

MBD1 605 aa

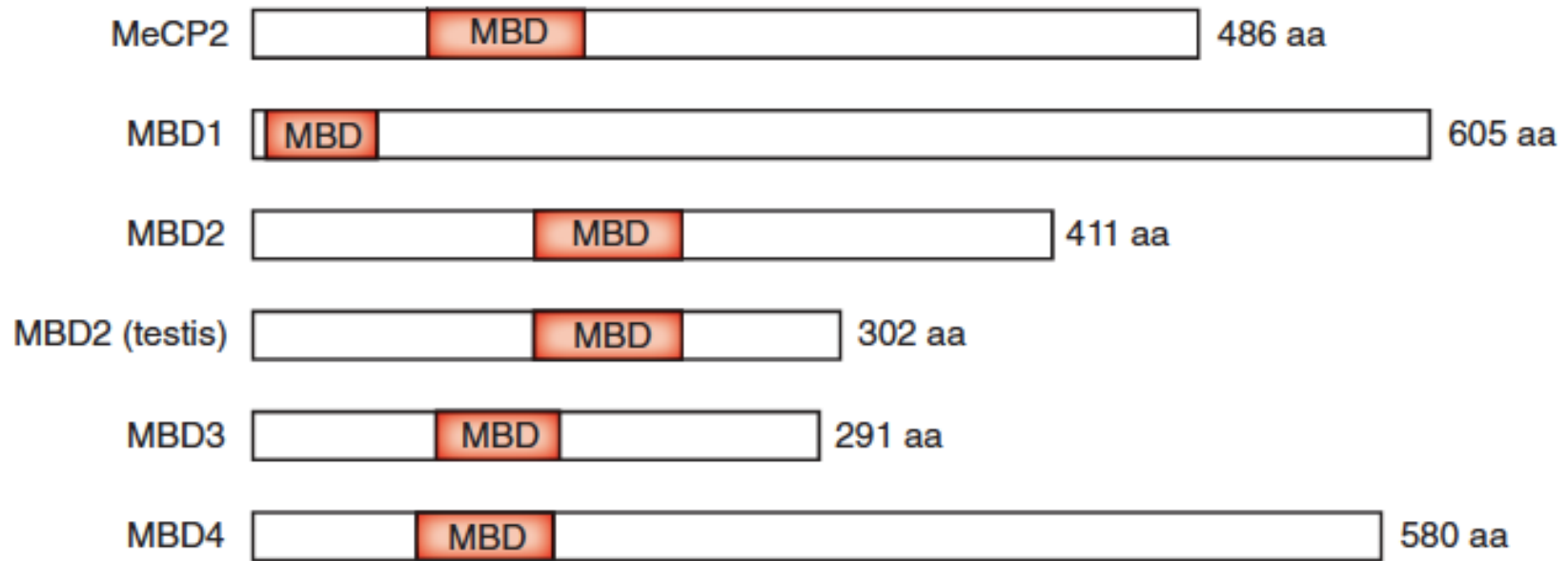
MBD2 411 aa

MBD2 (testis) 302 aa

MBD3 291 aa

MBD4 580 aa

Are proteins that share only a domain homologous?



- ◆ Definitely yes with respect to the domain
- ◆ Definitely no with respect to regions outside the shared domain
- ◆ Homology implies descent from a common ancestor, which only occurred with respect to the domain.
- ◆ Methyl-CpG-binding domain (**MBD**)

Example of a multidomain protein: HIV-1 pol

Pol (NP_789740), 995 amino acids long
Gag-Pol (NP_057849), 1435 amino acids

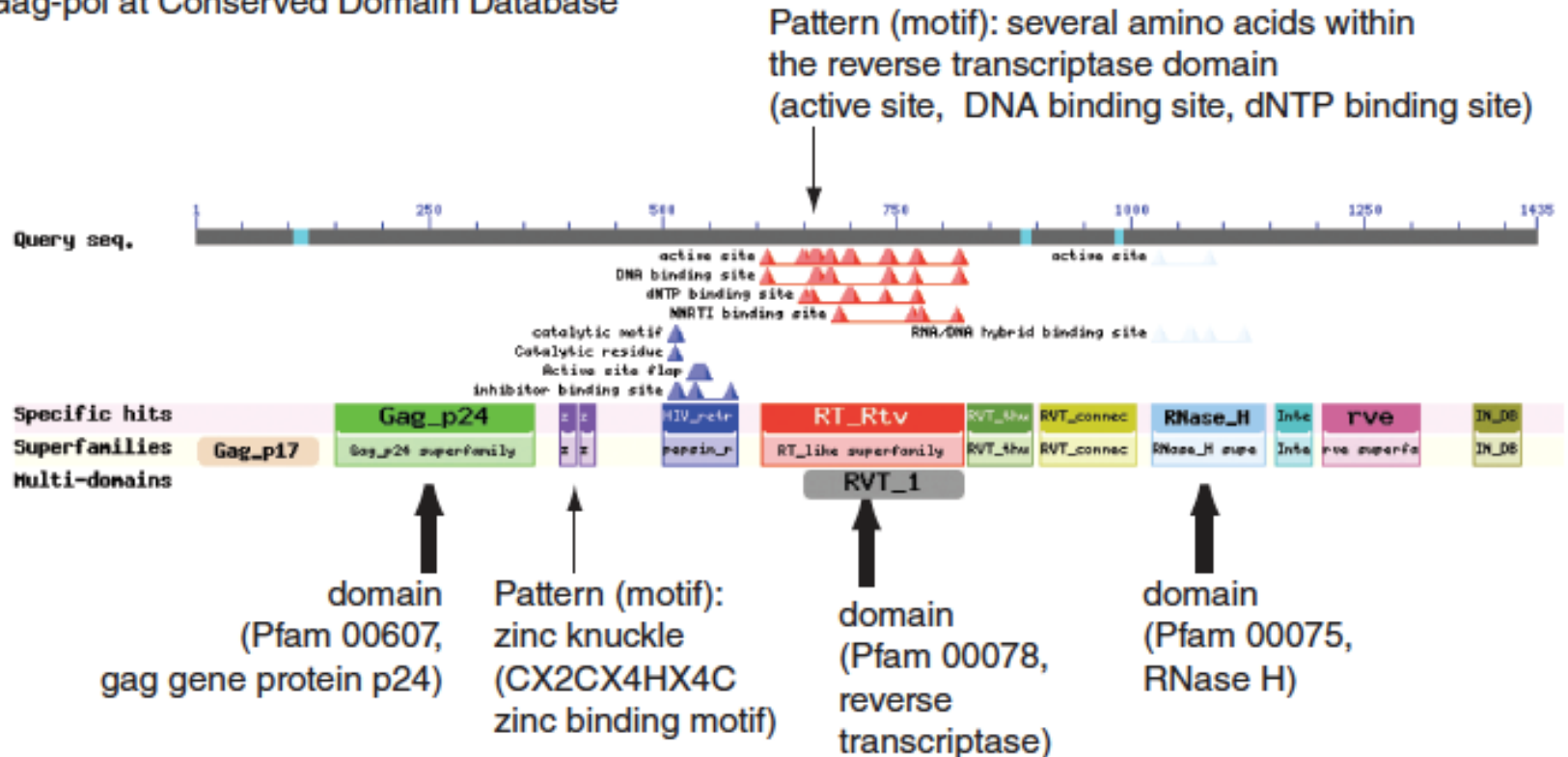
- cleaved into three proteins with distinct activities:
 - aspartyl protease
 - reverse transcriptase
 - integrase

We will explore HIV-1 pol and other proteins at the Expert Protein Analysis System (ExPASy) server.

Retroviral **integrase** (IN) is an enzyme produced by a retrovirus (such as HIV) that integrates—forms covalent links between—its DNA (genetic information) into that of the host cell it infects.

Searches for a multidomain protein: HIV gag-pol

Gag-pol at Conserved Domain Database



<https://www.uniprot.org/uniprot/P04585>

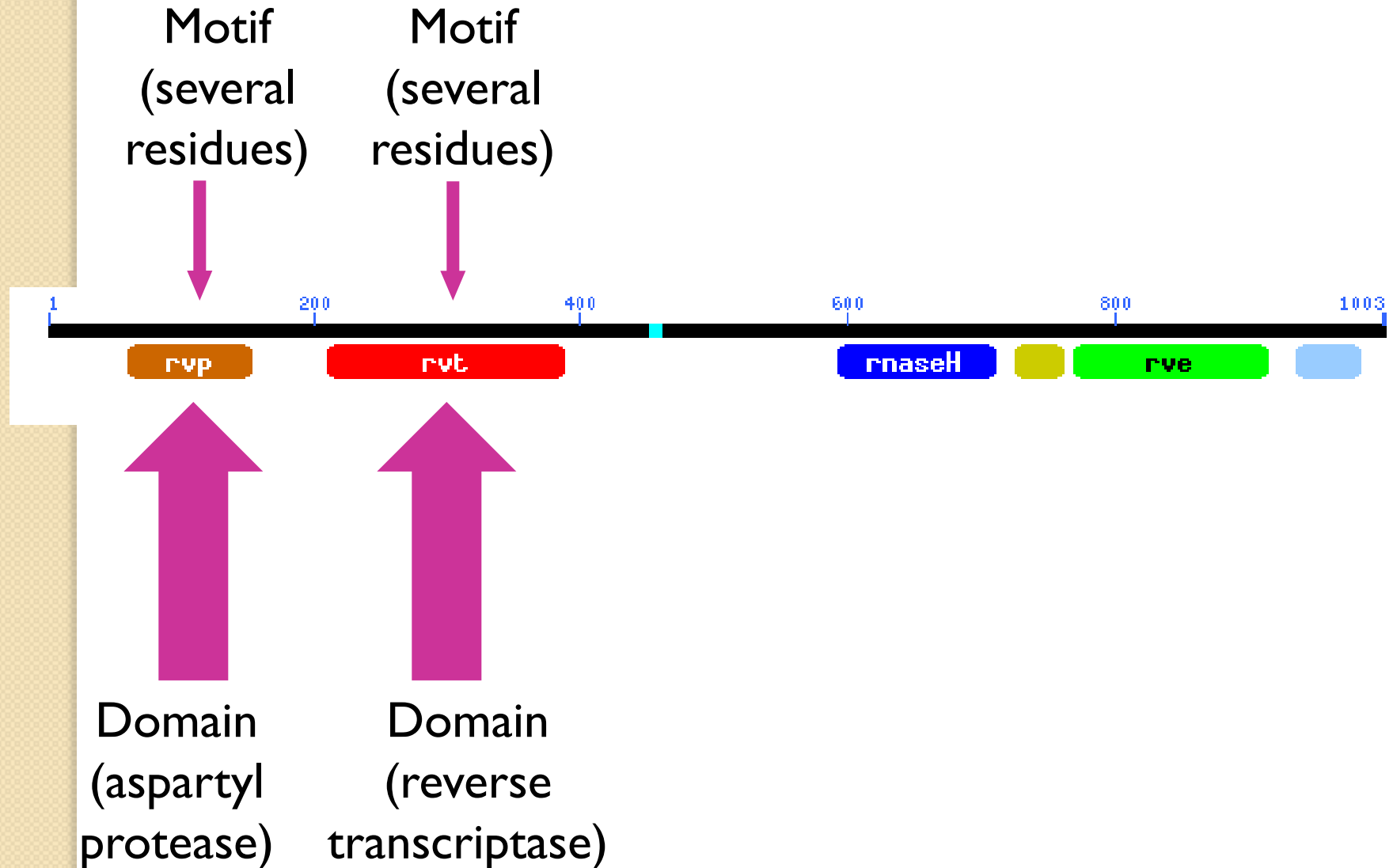
UniProt (www.uniprot.org): key proteomics database

Three protein databases recently merged to form UniProt:

- SwissProt
- TrEMBL (translated European Molecular Biology Lab)
- Protein Information Resource (PIR)

You can search for information on your favorite protein there; a BLAST server is provided.

Proteins can have both domains and motifs (patterns)



Eukaryotic and viral aspartyl proteases signature and profile

PROSITE cross-reference(s)	
PS00141; ASP_PROTEASE	Retrieve an alignment of Swiss-Prot true positive hits: [Clustal format, color, condensed view] [Clustal format, color] [Clustal format, pl]
PS50175; ASP_PROT_RETROV	Retrieve an alignment of Swiss-Prot true positive hits: [Clustal format, color, condensed view] [Clustal format, color] [Clustal format, pl]
Documentation	
<p>Aspartyl proteases, also known as acid proteases, (EC 3.4.23.-) are a widely distributed family of proteolytic enzymes [1,2,3] known to exist in vertebrates, fungi, plants, retroviruses and some plant viruses. Aspartate proteases of eukaryotes are monomeric enzymes which consist of two domains. Each domain contains an active site centered on a catalytic aspartyl residue. The two domains most probably evolved from the duplication of an ancestral gene encoding a primordial domain. Currently known eukaryotic aspartyl proteases are:</p> <ul style="list-style-type: none"> - Vertebrate gastric pepsins A and C (also known as gastricsin). - Vertebrate chymosin (rennin), involved in digestion and used for making cheese. - Vertebrate lysosomal cathepsins D (EC 3.4.23.5) and E (EC 3.4.23.34). - Mammalian renin (EC 3.4.23.15) whose function is to generate angiotensin I from angiotensinogen in the plasma. - Fungal proteases such as aspergillopepsin A (EC 3.4.23.18), candidapepsin (EC 3.4.23.24), mucoropepsin (EC 3.4.23.23) (mucor rennin), endothiapepsin (EC 3.4.23.22), polyporopepsin (EC 3.4.23.29), and rhizopuspepsin (EC 3.4.23.21). - Yeast saccharopepsin (EC 3.4.23.25) (proteinase A) (gene PEP4). PEP4 is implicated in posttranslational regulation of vacuolar hydrolases. - Yeast barrierpepsin (EC 3.4.23.35) (gene BAR1); a protease that cleaves alpha-factor and thus acts as an antagonist of the mating pheromone. - Fission yeast <i>ssa1</i> which is involved in degrading or processing the mating 	
Consensus pattern	[LIVMFGAC]-[LIVMTADN]-[LIVFSA]-D-[ST]-G-[STAV]-[STAPDENQ]-x-[LIVMFSTNC]-x-[LIVMFGTA] [D is the active site residue]
Sequences known to belong to this class detected by the pattern	ALL.
Other sequence(s) detected in Swiss-Prot	37.
Sequences known to belong to this class detected by the profile	ALL viral- type proteases.

Definition of a motif

A motif (or fingerprint) is a short, conserved region of a protein. Its size is often 10 to 20 amino acids.

Simple motifs include transmembrane domains and phosphorylation sites. These do not imply homology when found in a group of proteins.

PROSITE (www.expasy.org/prosite) is a dictionary of motifs (there are currently 1600 entries). In PROSITE, a pattern is a qualitative motif description (a protein either matches a pattern, or not). In contrast, a profile is a quantitative motif description. We will encounter profiles in Pfam, ProDom, SMART, and other databases.

Summary of Perspective I: Protein domains and motifs

A signature is a protein category such as a domain or motif.

You can learn about domains in databases such as InterPro and Pfam.

A motif (or fingerprint) is a short, conserved sequence. You can study motifs at Prosite at ExPASy.

Perspective 2: Physical properties of proteins

Post-translational modifications of proteins at InterPro

Accession	Post-translational modification site
IPR000152	EGF-type aspartate/asparagine hydroxylation site
IPR001020	Phosphotransferase system, HPr histidine phosphorylation site
IPR002114	Phosphotransferase system, HPr serine phosphorylation site
IPR002332	Nitrogen regulatory protein P-II, uridylation site
IPR004091	Chemotaxis methyl-accepting receptor, methyl-accepting site
IPR006141	Intein splice site
IPR006162	Phosphopantetheine attachment site
IPR012902	Prokaryotic N-terminal methylation site
IPR018051	Surfactant-associated polypeptide, palmitoylation site
IPR018070	Neuromedin U, amidation site
IPR018243	Neuromodulin, palmitoylation/phosphorylation site
IPR018303	P-type ATPase, phosphorylation site
IPR019736	Synapsin, phosphorylation site
IPR019769	Translation elongation factor, IF5A, hypusine site
IPR021020	Adhesin, Dr family, signal peptide

Physical properties of proteins

Many websites are available for the analysis of individual proteins. ExPASy and ISREC are two excellent resources.

The accuracy of these programs is variable. Predictions based on primary amino acid sequence (such as molecular weight prediction) are likely to be more trustworthy. For many other properties (such as posttranslational modification of proteins by specific sugars), experimental evidence may be required rather than prediction algorithms.

Access a variety of protein analysis programs from the ExPASy home page



ExPASy

Bioinformatics Resource Portal

Compute pI/Mw

Compute pI/Mw

Theoretical pI/Mw (average) for the user-entered sequence:

```
      10      20      30      40      50      60
MVHLTPEEKS AVTALWGKVN VDEVGGEALG RLLVVYPWTQ RFFESFGDLS TPDAVMGNPKK

      70      80      90     100     110     120
VKAHGKKVLG AFSDGLAHLD NLKGTFFATLS ELHCDKLHVD PENFRLLGNV LVCVLAHHFGG

     130     140
KEFTPPVQAA YQKVVAGVAN ALAHKYH
```

Theoretical pI/Mw: 6.74 / 15998.41

NetPhos to predict phosphorylation sites: Example of an ExPASy program for proteomics analysis

147 Sequence

MVHLTPEEKSAVTALWGKVVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHL
NLKGTFTATLSLHCDKLVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAAYQKVVAGVANALAHKYH

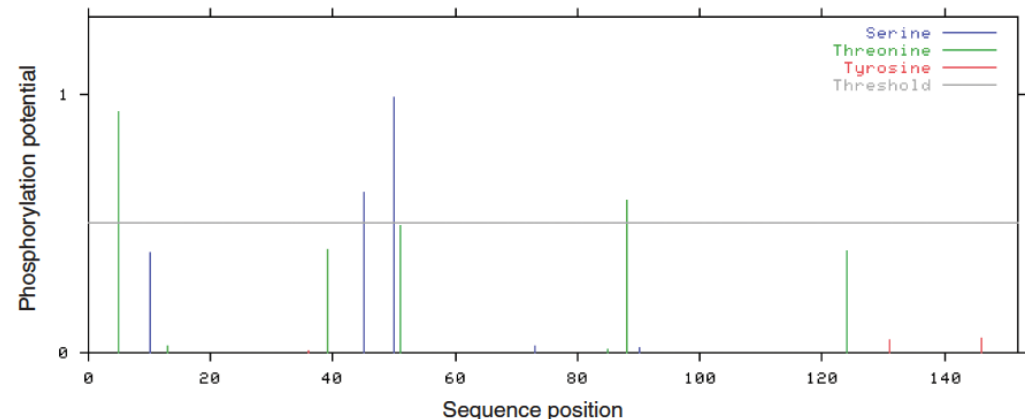
.....T.....S.....S.....
.....T.....

Phosphorylation sites predicted: Ser: 2 Thr: 2 Tyr: 0

Serine predictions

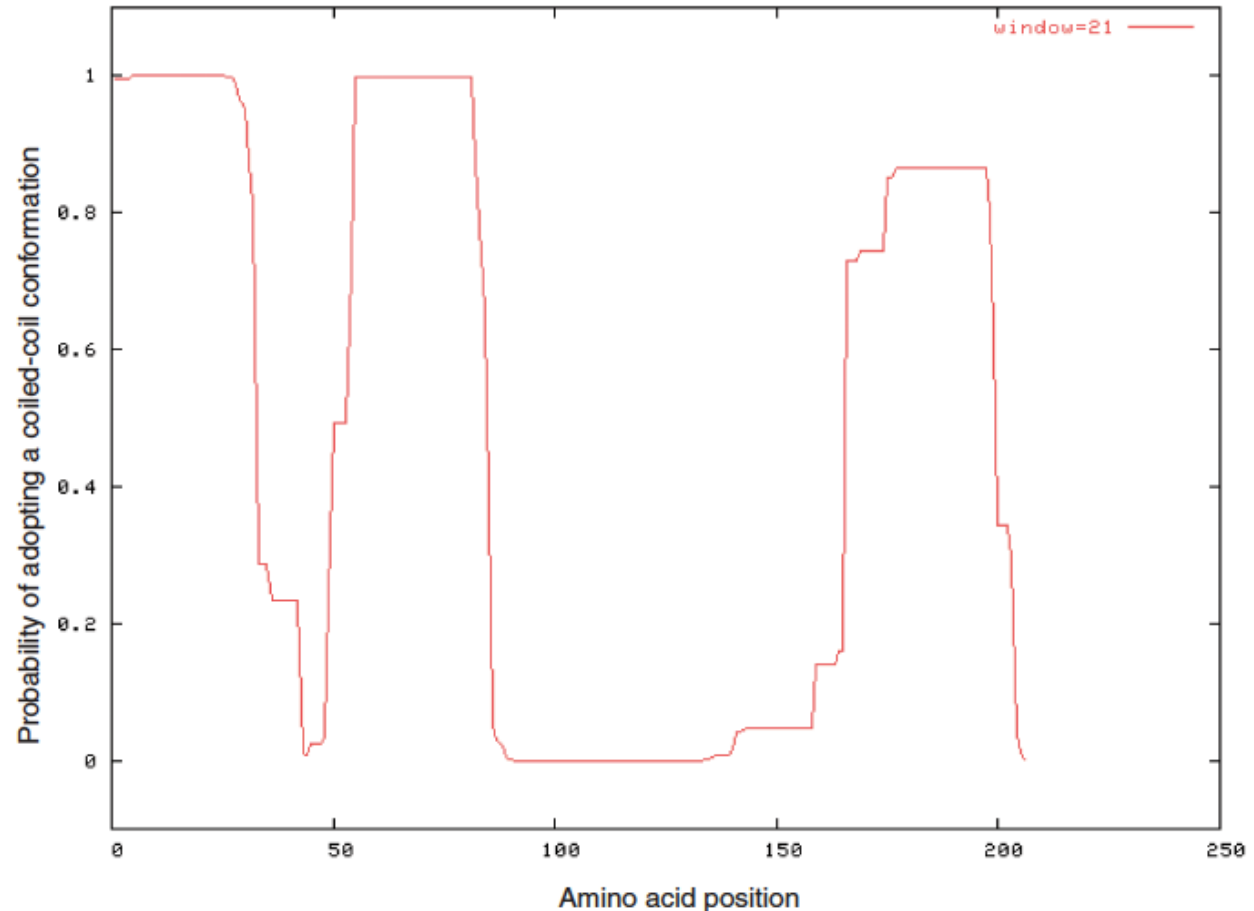
Name	Pos	Context	Score	Pred
		V		
Sequence	10	PEEKSAVTA	0.389	.
Sequence	45	RFFESFGDL	0.621	*S*
Sequence	50	FGDLSTPDA	0.987	*S*
Sequence	73	LGAFSDGLA	0.026	.
Sequence	90	FATLSELHC	0.020	.

^

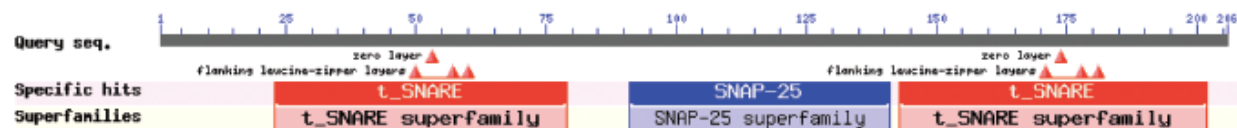


COILS program assesses the likelihood that a protein sequence forms a coiled-coil structure (implicated in protein-protein interactions)

(a) COILS output for SNAP-25



(b) Domains from Conserved Domain Database (NCBI)



Introduction to Perspectives 3 and 4: Gene Ontology (GO) Consortium

The Gene Ontology Consortium

An ontology is a description of concepts. The GO Consortium compiles a dynamic, controlled vocabulary of terms related to gene products.

There are three organizing principles:

- Molecular function

- Biological process

- Cellular compartment

You can visit GO at <http://www.geneontology.org>.

There is no centralized GO database. Instead, curators of organism-specific databases assign GO terms to gene products for each organism.

The Gene Ontology Consortium: Evidence Codes

IC Inferred by curator

IDA Inferred from direct assay

IEA Inferred from electronic annotation

IEP Inferred from expression pattern

IGI Inferred from genetic interaction

IMP Inferred from mutant phenotype

IPI Inferred from physical interaction

ISS Inferred from sequence or structural similarity

NAS Non-traceable author statement

ND No biological data

TAS Traceable author statement

GO terms are assigned to NCBI Gene entries

GeneOntology

Provided by [GOA](#)

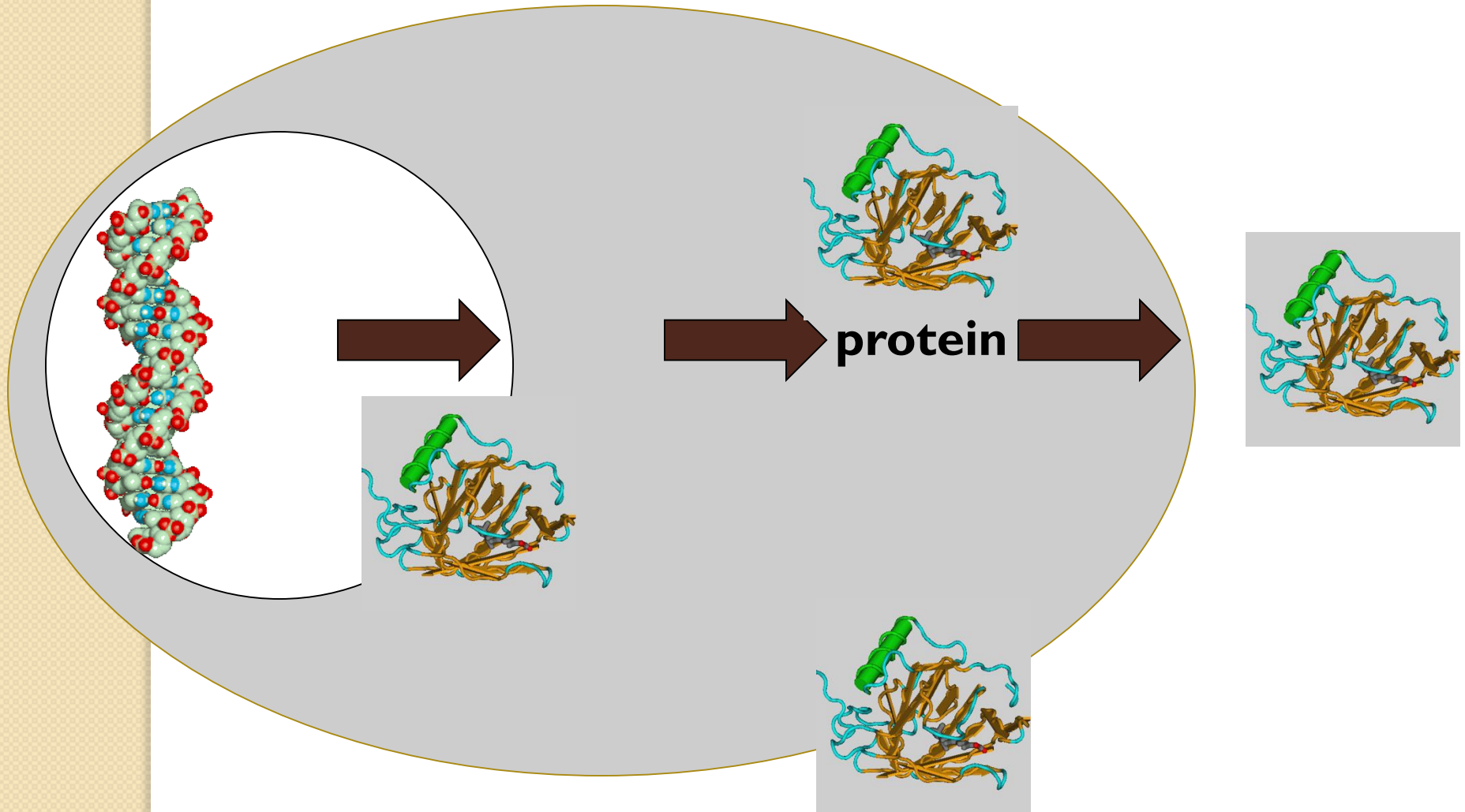
Function	Evidence
heme binding	IEA
hemoglobin binding	IDA PubMed
iron ion binding	IEA
metal ion binding	IEA
molecular function	ND
oxygen binding	IDA PubMed
oxygen binding	IEA
oxygen transporter activity	IEA
oxygen transporter activity	NAS PubMed
selenium binding	IDA PubMed

Process	Evidence
biological process	ND
nitric oxide transport	NAS PubMed
oxygen transport	IEA
oxygen transport	NAS PubMed
oxygen transport	TAS PubMed
positive regulation of nitric oxide biosynthetic process	NAS PubMed
transport	IEA

Component	Evidence
hemoglobin complex	IEA
hemoglobin complex	NAS PubMed
hemoglobin complex	TAS PubMed

Perspective 3: Protein localization

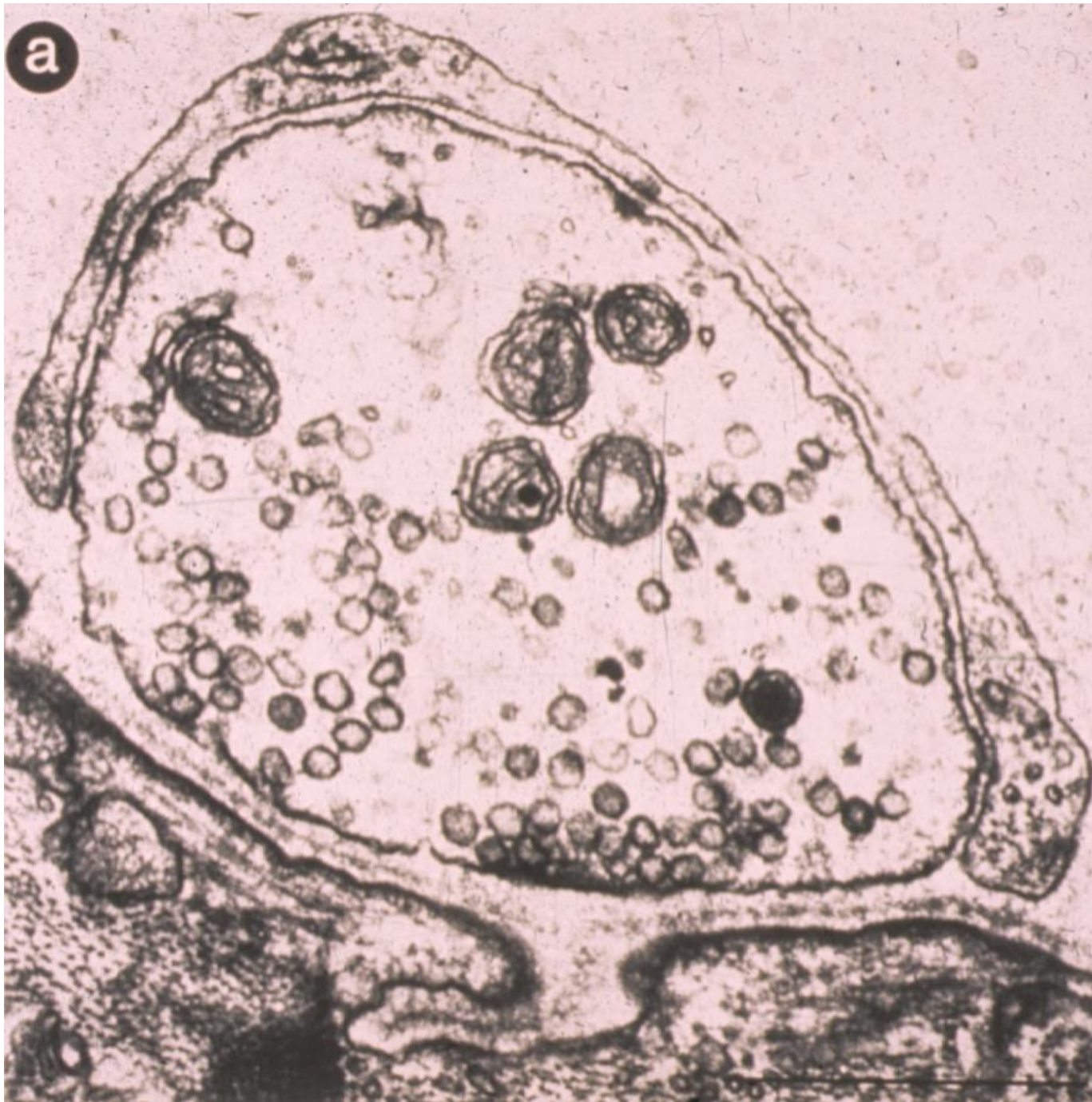
Protein localization



Protein localization

Proteins may be localized to intracellular compartments, cytosol, the plasma membrane, or they may be secreted. Many proteins shuttle between multiple compartments.

A variety of algorithms predict localization, but this is essentially a cell biological question.



Results of Subprograms

PSG: a new signal peptide prediction method

N-region: length 2; pos.chg 1; neg.chg 0
H-region: length 14; peak value 10.03
PSG score: 5.63

GvH: von Heijne's method for signal seq. recognition

GvH score (threshold: -2.1): 3.93
possible cleavage site: between 16 and 17

>>> Seems to have a cleavable signal peptide (1 to 16)

Tmpred: predict membrane topology of proteins

Usage: Paste your sequence in one of the supported [formats](#) into the sequence field below and press the "Run TMpred" button.

Make sure that the format button (next to the sequence field) shows the correct format

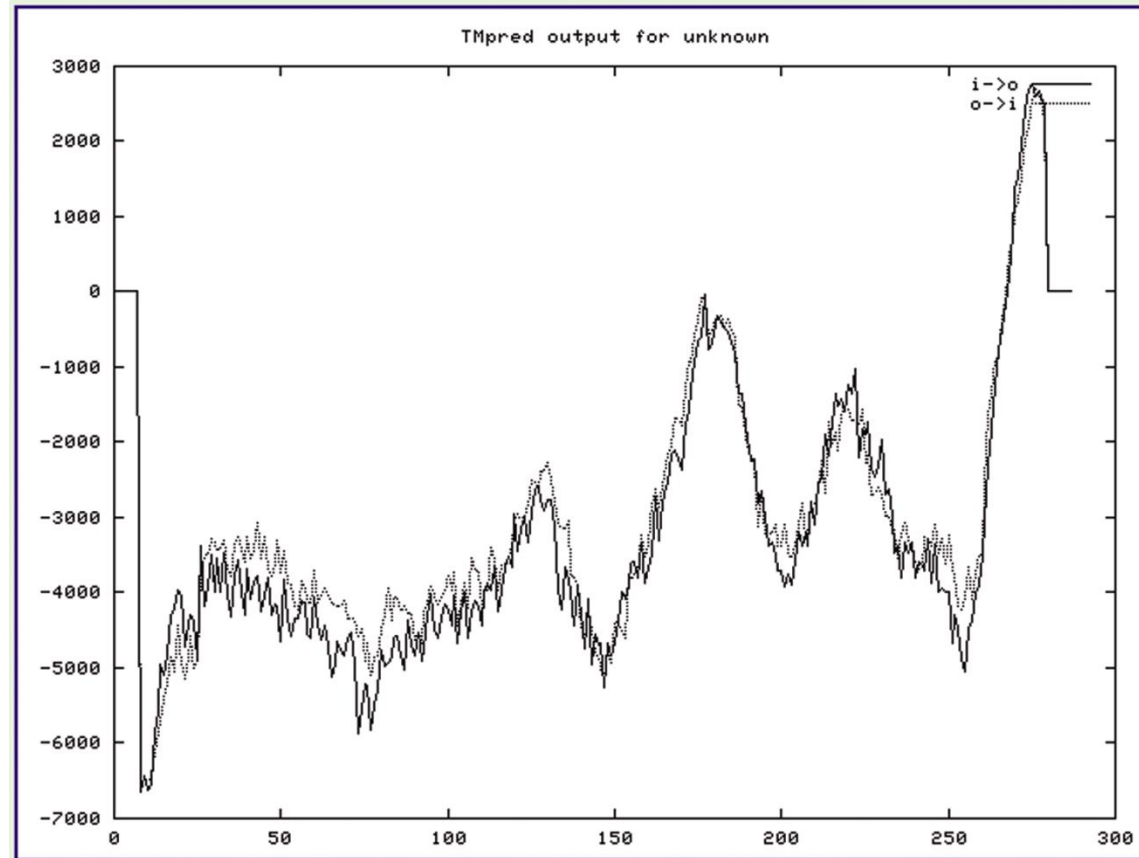
Choose the minimal and maximal length of the hydrophobic part of the transmembrane helix

Output format	<input type="text" value="html"/> minimum <input type="text" value="17"/> maximum <input type="text" value="33"/>
Query title (optional)	<input type="text"/>
Input sequence format	<input type="text" value="Plain Text"/>
Query sequence: or ID or AC or GI (see above for valid formats)	<div>MKDRTQELRTAKDSDDDDVAVTVDRDRFMDEFFEQVEEIRGFIDKIAENVVEEVKRKHSAILASPNPDEKTKEELEELMS DIKKTANKVRSKLKSIEQSIEQEEGLNRSSADLRIRKQHSTLSRKFEVMSEYNATQSDYRERCKGRIQRQLEITGRTT TSEELEDMLESGNPAIFASGIIMDSSISKQALSEIETRHSEIIKLENSIRELHDMFMDMAMLVESQGEMIDRIEYNVEHA VDYVERAVSDTKKAVKYQSKARRKKIMIIICCVILGIVIASTVGGIFA</div>
<div>Run TMpred</div> <div>Clear Input</div>	

TMpred: predict membrane topology of proteins

2 possible models considered, only significant TM-segments used

```
-----> slightly preferred model: N-terminus inside  
1 strong transmembrane helices, total score : 2757  
# from   to length score orientation  
1  266  284 (19)    2757 i-o  
  
-----> alternative model  
1 strong transmembrane helices, total score : 2690  
# from   to length score orientation  
1  266  288 (23)    2690 o-i
```



Perspective 4: Protein function

Protein function

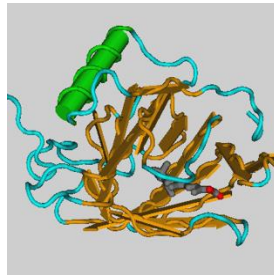
Function refers to the role of a protein in the cell. We can consider protein function from a variety of perspectives.

I. Biochemical function (molecular function)

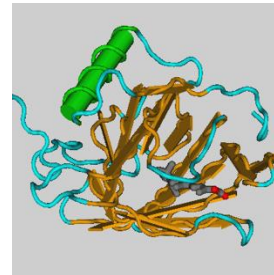


RBP binds retinol,
could be a carrier

2. Functional assignment based on homology



RBP
could be
a carrier
too



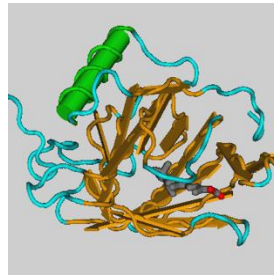
Other
carrier
proteins

3. Function based on structure



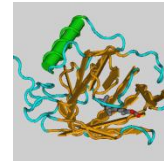
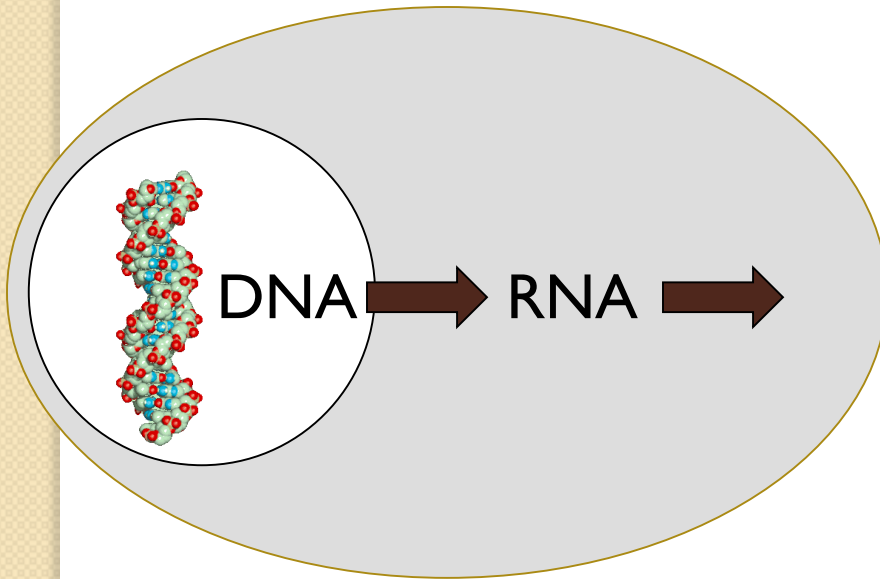
RBP forms a calyx

4. Function based on ligand binding specificity



RBP binds vitamin A

5. Function based on cellular process



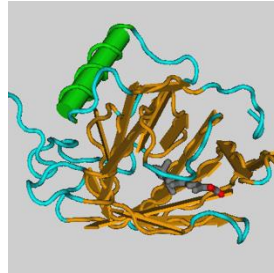
**RBP is abundant,
soluble, secreted**

6. Function based on biological process



RBP is essential for vision

7. Function based on “proteomics” or high throughput “functional genomics”



High throughput analyses show...

RBP levels elevated in renal failure

RBP levels decreased in liver disease

Functional assignment of enzymes: the EC (Enzyme Commission) system

Oxidoreductases	1,003
Transferases	1,076
Hydrolases	1,125
Lyases	356
Isomerases	156
Ligases	126

Functional assignment of proteins: Clusters of Orthologous Groups (COGs)

Information storage and processing

Cellular processes

Metabolism

Poorly characterized

Perspective

Our understanding of the properties of proteins has advanced dramatically, from the level of biochemical function to the role of proteins in cellular processes. Advances in instrumentation have propelled mass spectrometry into a leading role for many proteomics applications.

Pitfalls

Many of the experimental and computational strategies used to study proteins have limitations.

- Two-dimensional protein gels are most useful for studying relatively abundant proteins, but thousands of proteins expressed at low levels are harder to characterize.
- Experimental approaches are extremely challenging in practice, as shown by the ABRF critical assessments.
- Many computational approaches suffer from high false positive error rates, reflecting the difficulty of obtaining adequate training sets.