Chapter 12
Protein analysis and proteomics

Bioinformatics and Functional Genomics
3rd edition (2015)
Jonathan Pevsner, Ph.D.

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.

B&FG 3e
Page 539

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.

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

B&FG 3e
Page 542

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

Work groups

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

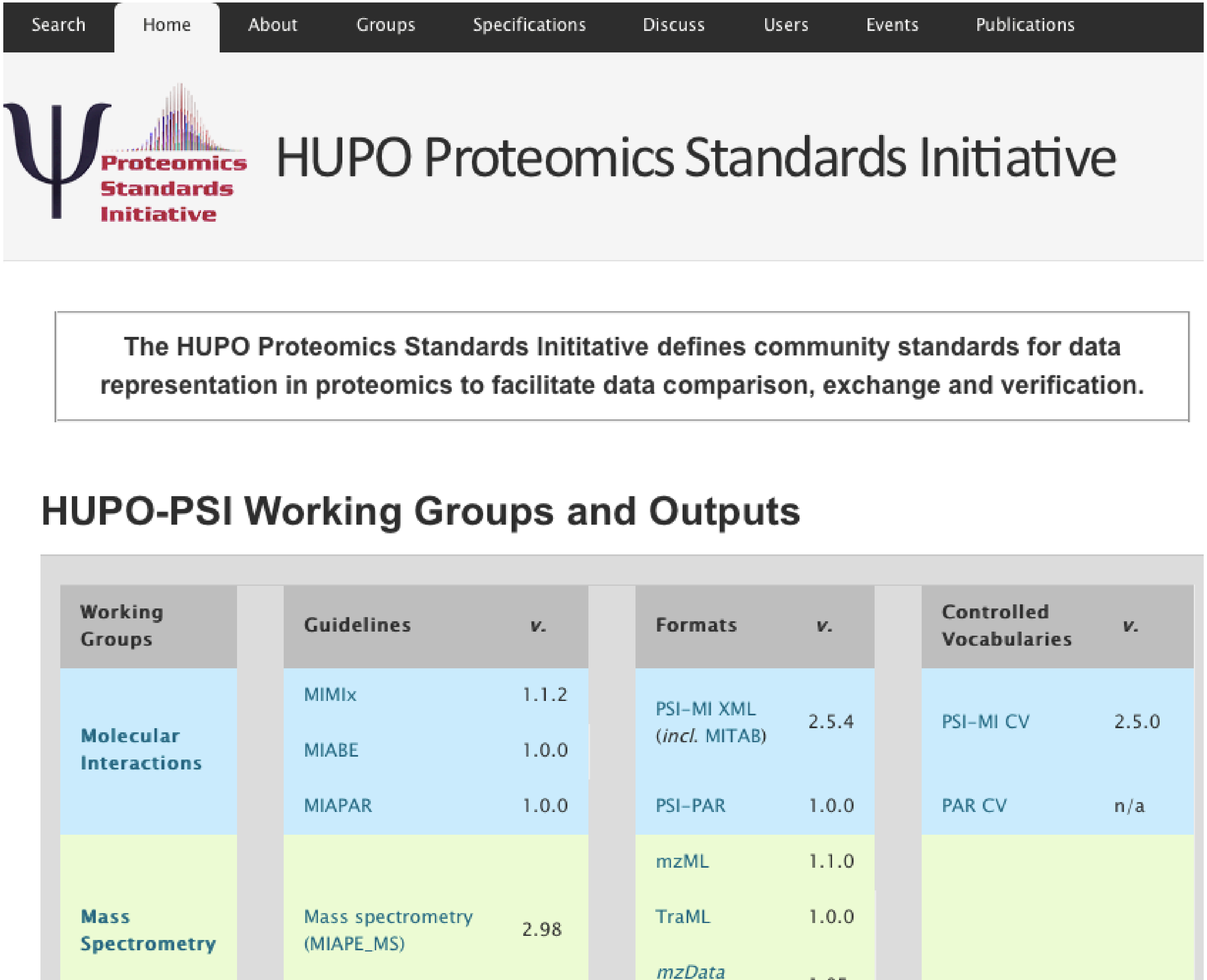
Themes

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

B&FG 3e
Page 542

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

http://www.psidev.info/



HUPO Proteomics Standards Initiative

The HUPO Proteomics Standards Inititative 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	Mass spectrometry (MIAPE_MS)	2.98	mzML	1.1.0		
			TraML	1.0.0		
			mzData	1.05		

B&FG 3e
Page 542

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

B&FG 3e
Page 543

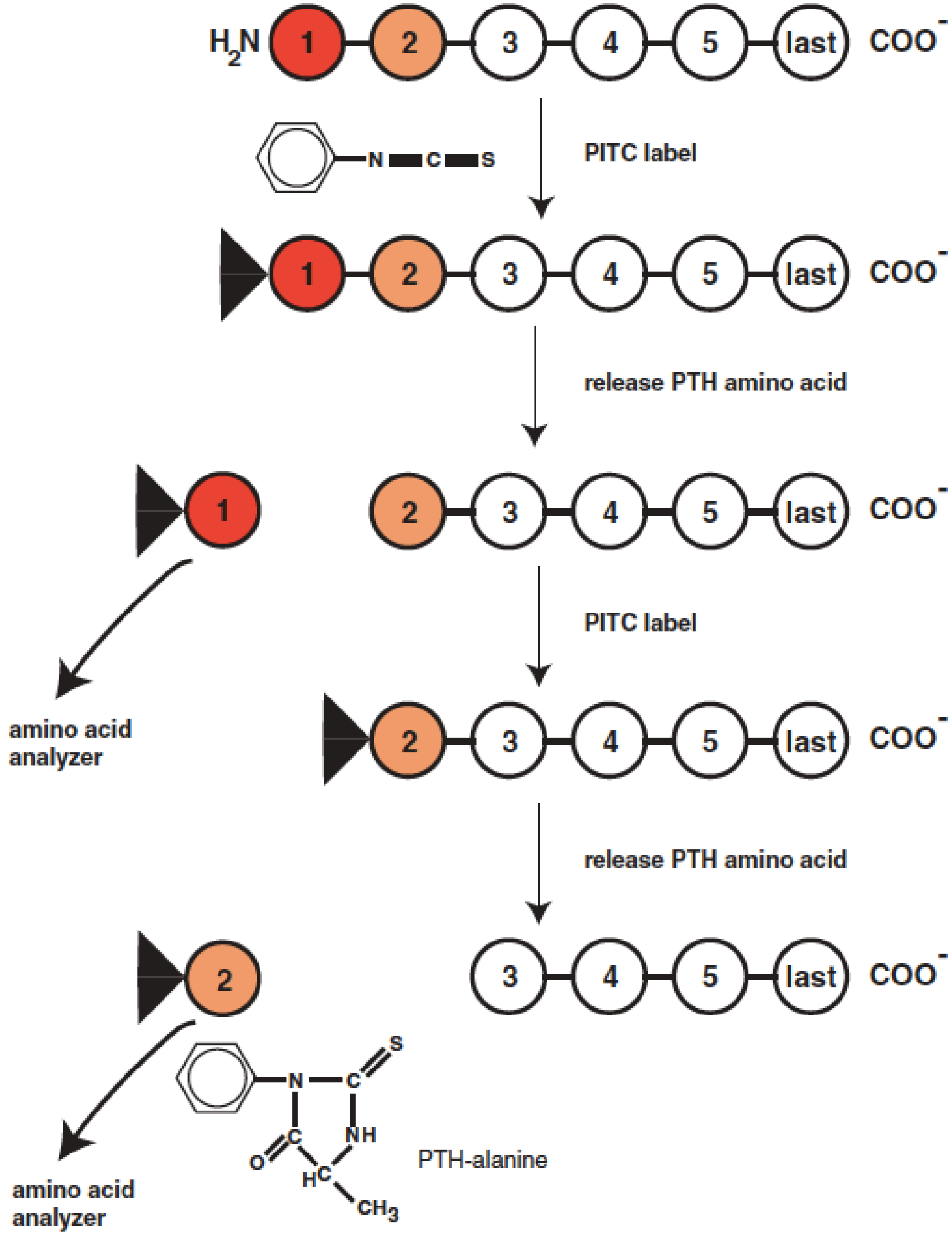
Protein sequencing by Edman degradation

Beginning in the 1949 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.

B&FG 3e
Fig. 12.1
Page 543

Protein sequencing by Edman degr



B&FG 3e
Fig. 12.1
Page 544

Polyacrylamide gel electrophoresis (PAGE)

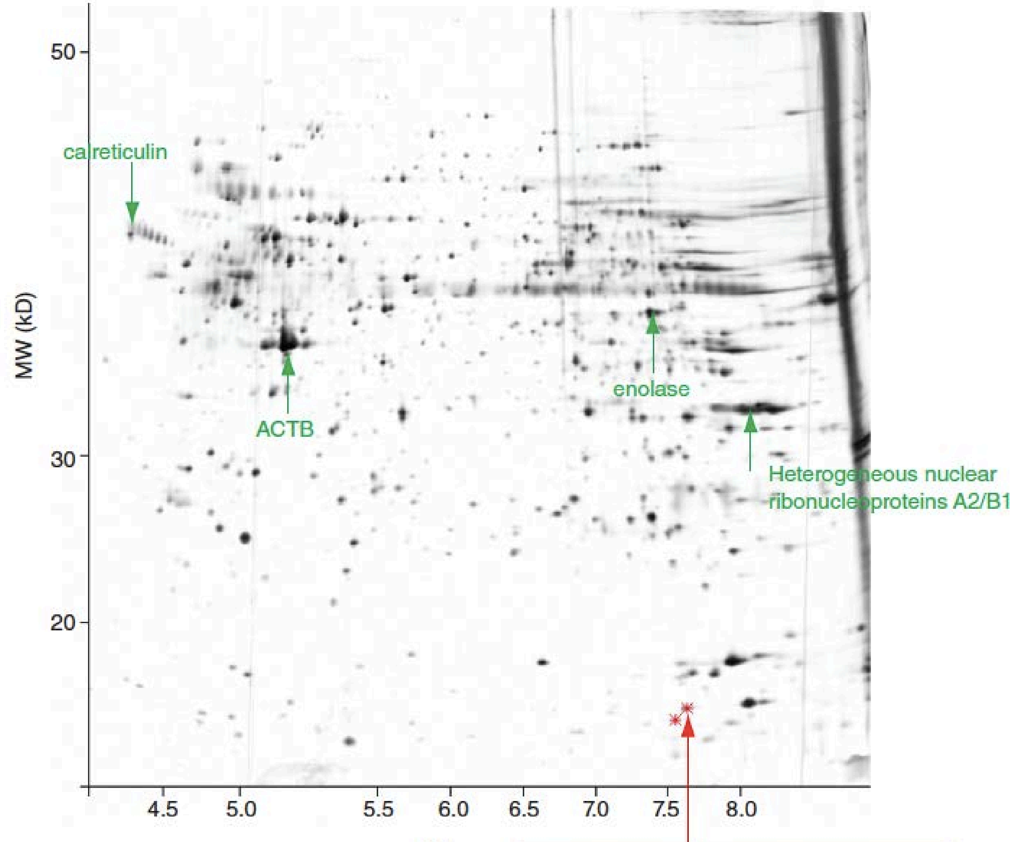
Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) is useful to separate proteins based on molecular mass.

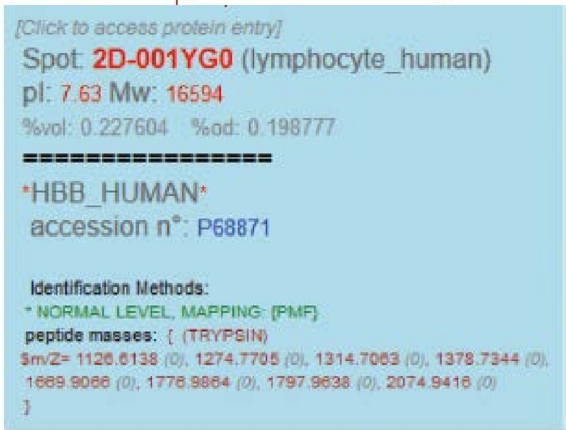
Two dimensional SDS-PAGE includes a second separation of proteins in the basis of charge: a protein migrates in an electric field to its isoelectric point, the pH at which the net charge is neutral.

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).

B&FG 3e
Page 544

Polyacrylamide gel electrophoresis (PAGE)





[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)

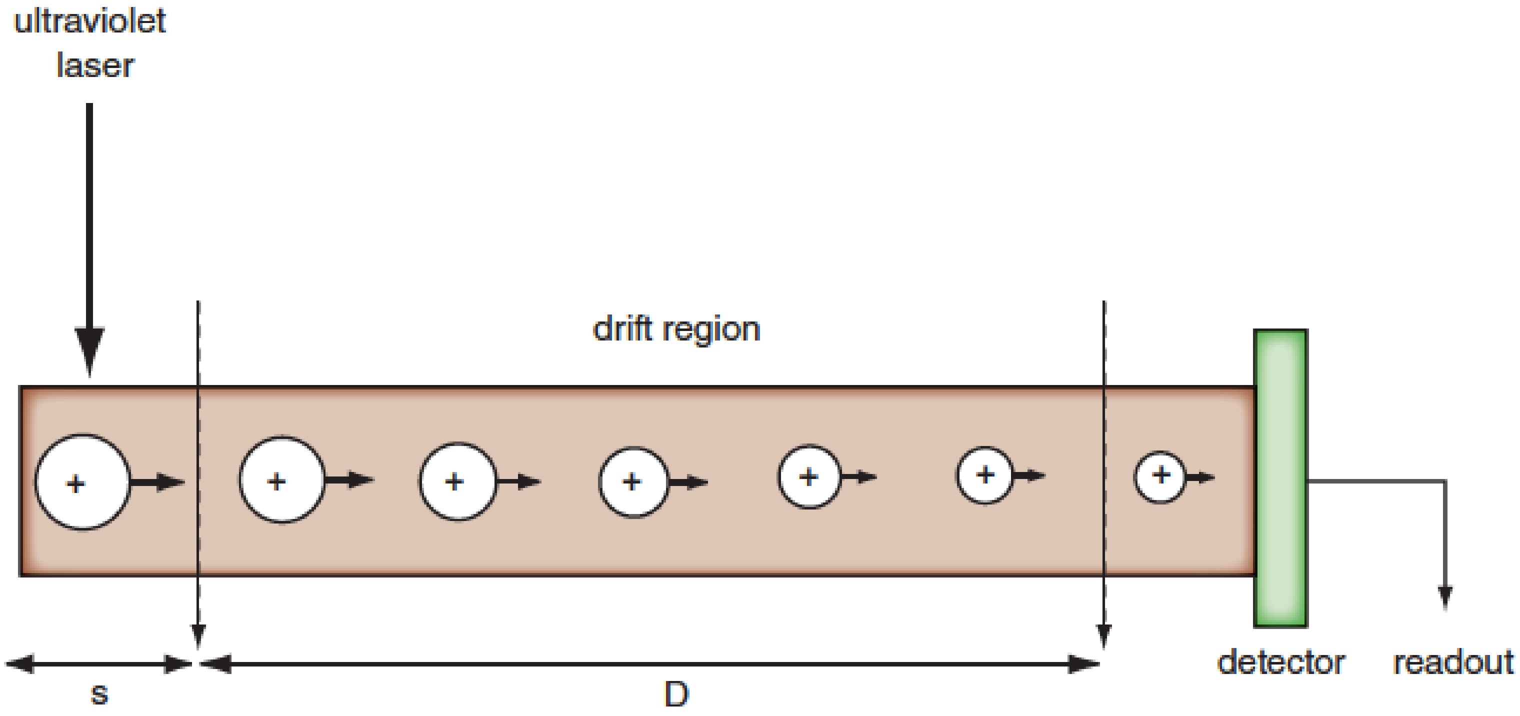
$m/Z= 1126.6138 (0), 1274.7705 (0), 1314.7063 (0), 1378.7344 (0), 1669.9066 (0), 1776.9864 (0), 1797.9638 (0), 2074.9416 (0)

}

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

B&FG 3e
Fig. 12.2
Page 545

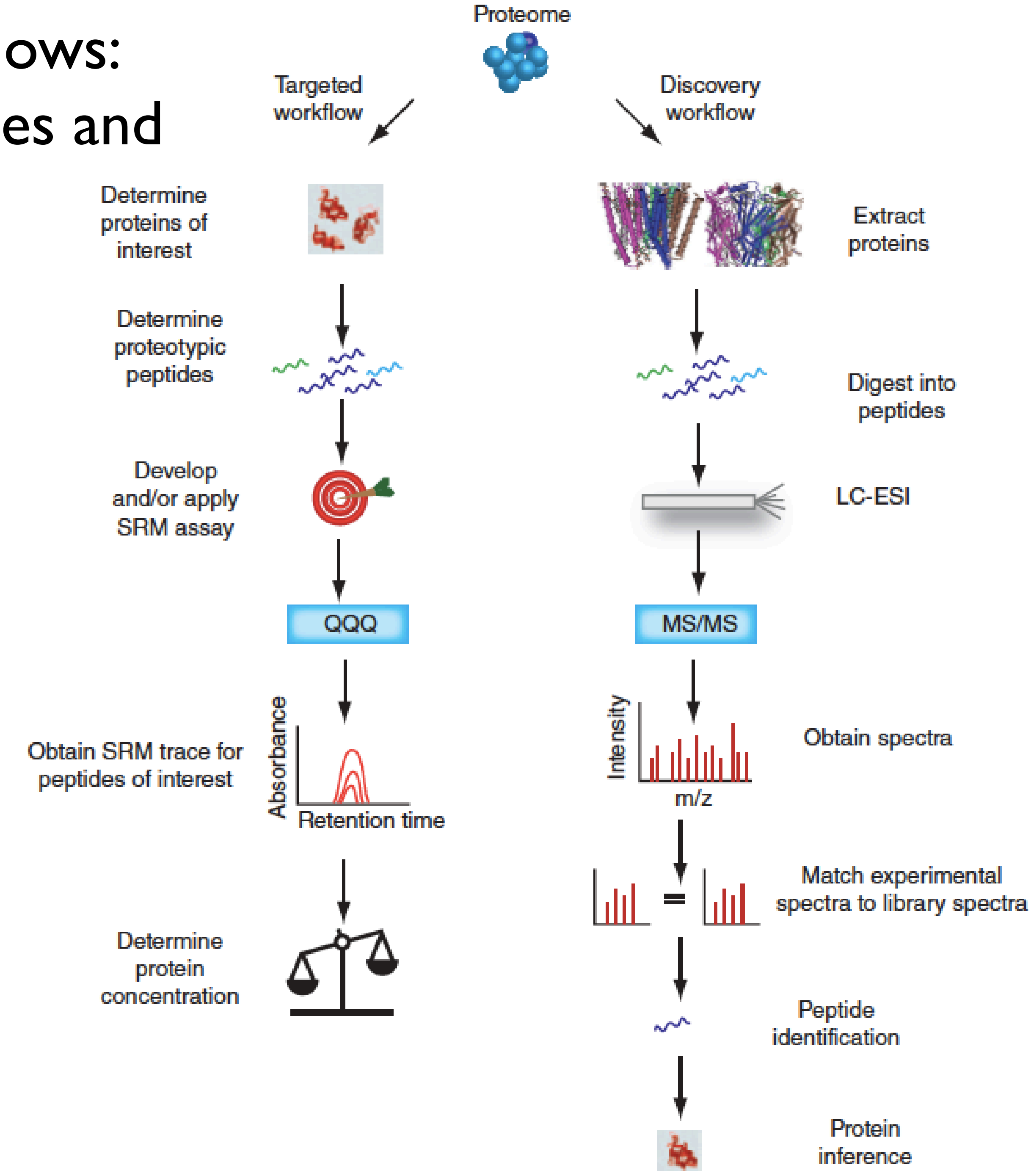
Matrix-assisted laser desorption/ionization time-of-flight spectroscopy (MALDI-TOF)



Mass spectrometry (MS) enables sensitive identification of proteins

B&FG 3e
Fig. 12.4
Page 548

Two MS workflows: targeted analyses and discovery


Proteome
Targeted workflow
Discovery workflow
Determine proteins of interest
Extract proteins
Determine proteotypic peptides
Digest into peptides
Develop and/or apply SRM assay
LC-ESI
QQQ
MS/MS
Obtain SRM trace for peptides of interest
Absorbance
Retention time
Intensity
m/z
Obtain spectra
Match experimental spectra to library spectra
Peptide identification
Determine protein concentration
Protein inference

B&FG 3e
Fig. 12.5
Page 549

Outline

Introduction

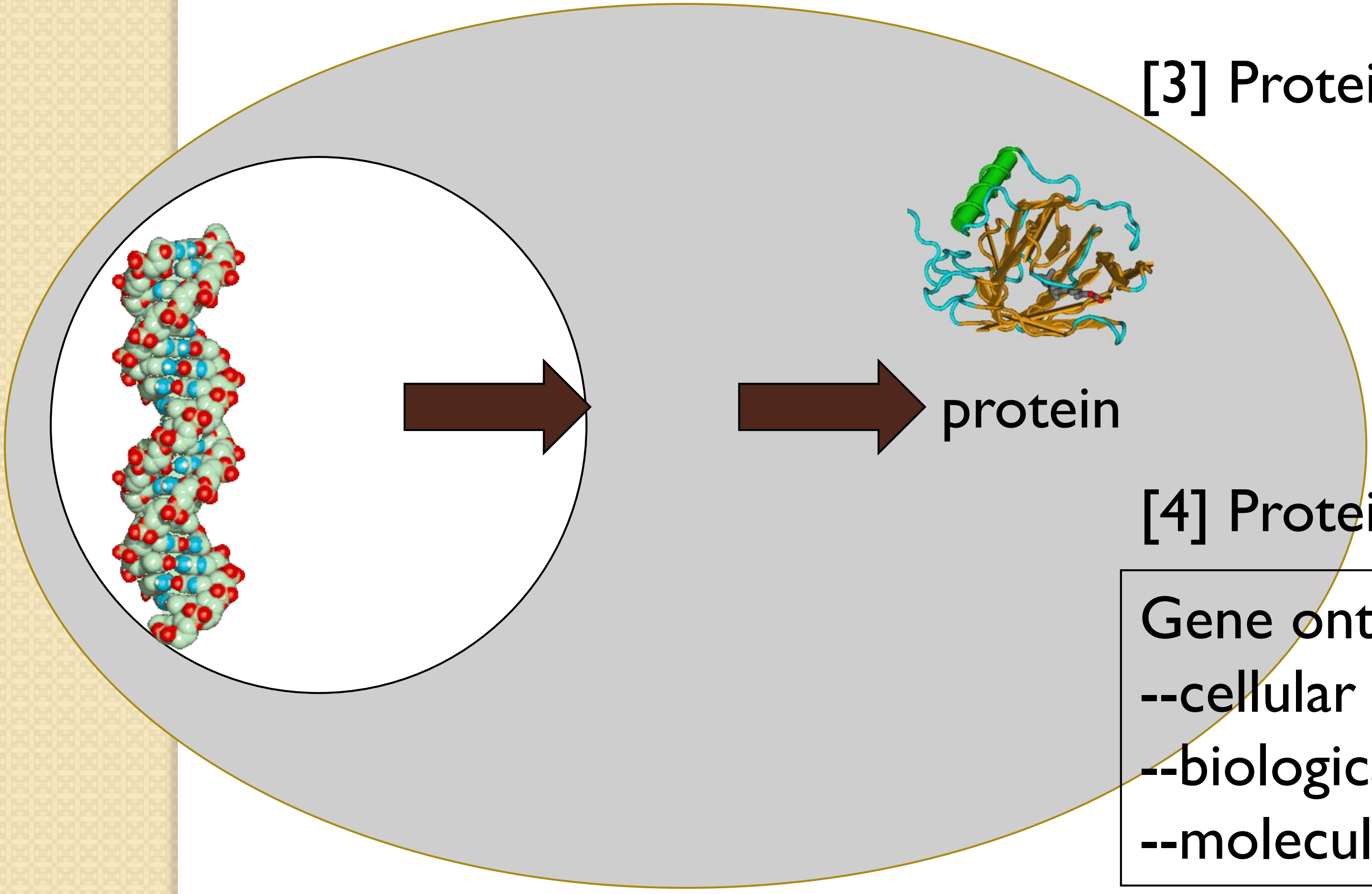
Techniques for identifying proteins

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

B&FG 3e
Fig. 12.7
Page 552

Perspective 1:
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

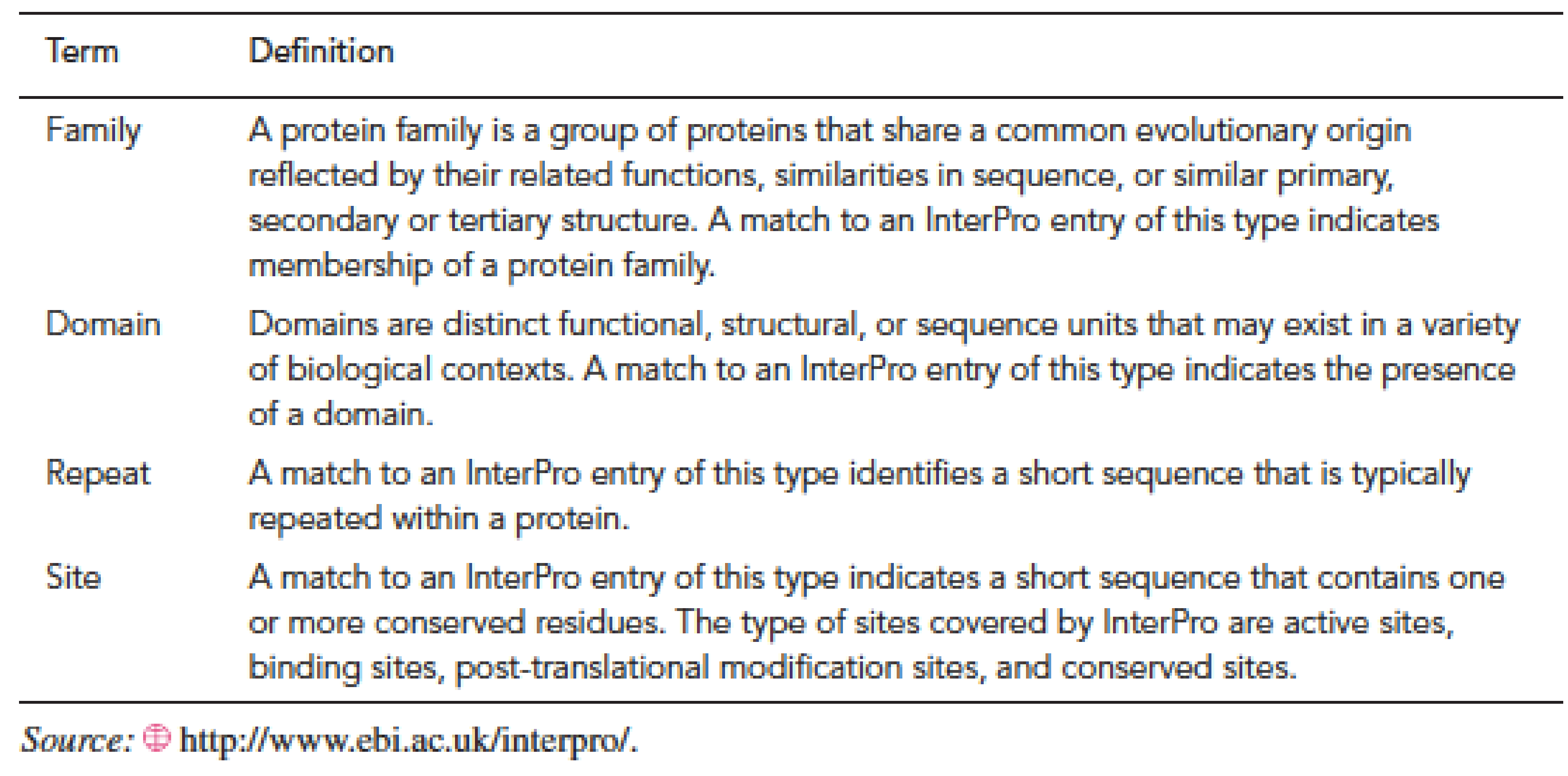
Domain:

- 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

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

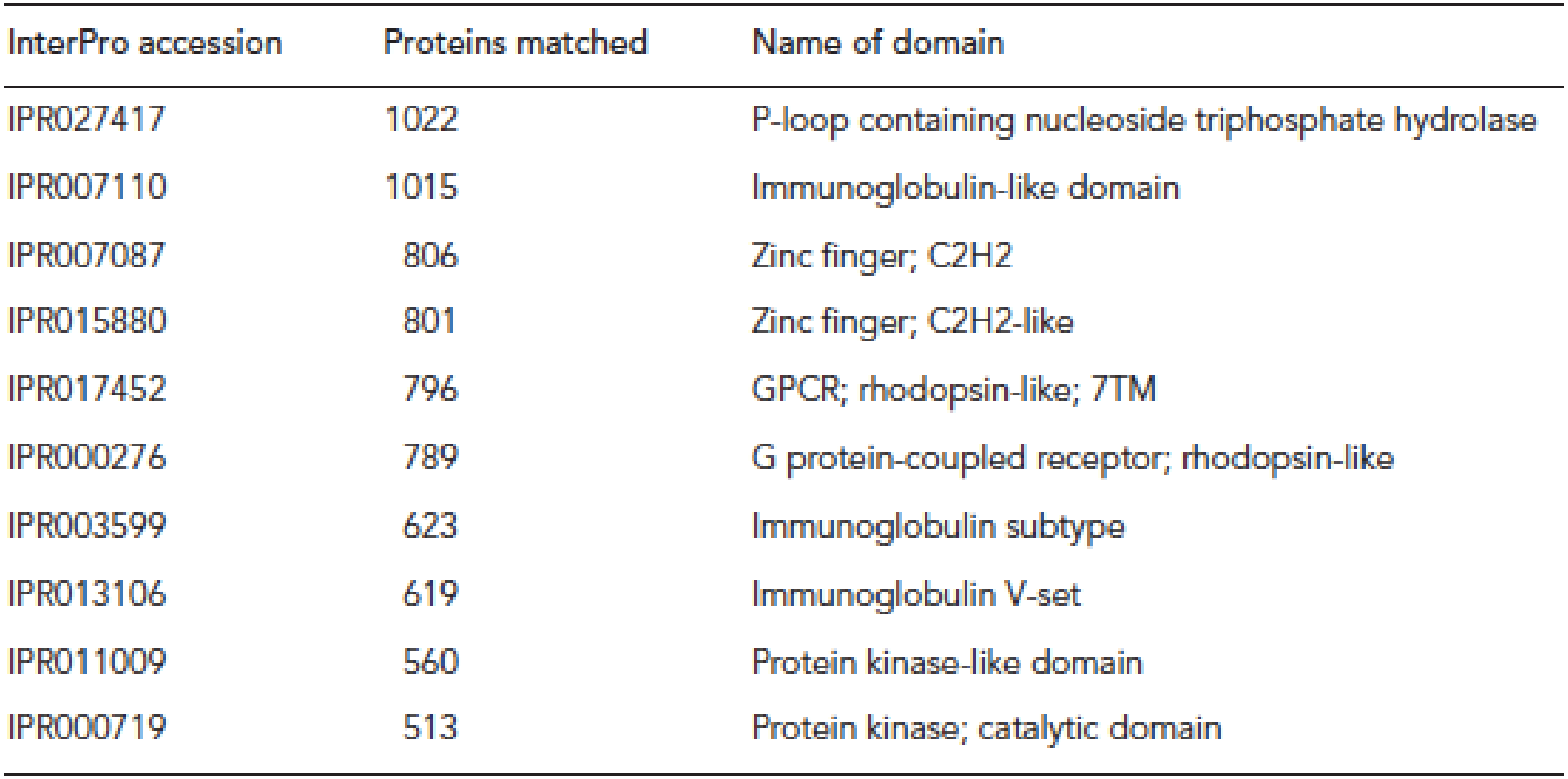


Term	Definition
Family	A protein family is a group of proteins that share a common evolutionary origin reflected by their related functions, similarities in sequence, or similar primary, secondary or tertiary structure. A match to an InterPro entry of this type indicates membership of a protein family.
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/.

B&FG 3e
Tab. 12.1
Page 553

10 most common domains (human)

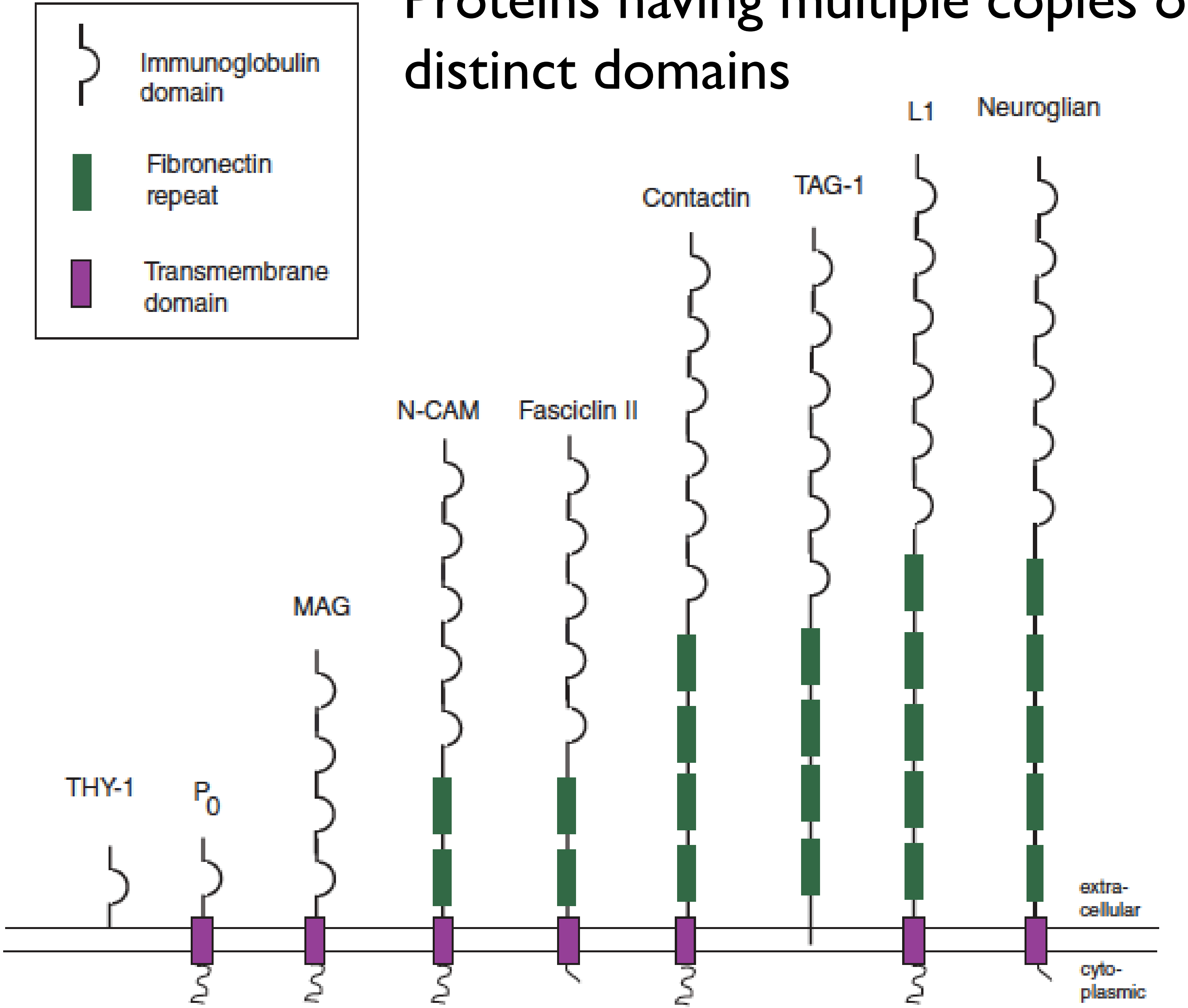


InterPro accession	Proteins matched	Name of domain
IPR027417	1022	P-loop containing nucleoside triphosphate hydrolase
IPR007110	1015	Immunoglobulin-like domain
IPR007087	806	Zinc finger; C2H2
IPR015880	801	Zinc finger; C2H2-like
IPR017452	796	GPCR; rhodopsin-like; 7TM
IPR000276	789	G protein-coupled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

B&FG 3e
Tab. 12.3
Page 554

Source: InterPro (2015)

Proteins having multiple copies of distinct domains



B&FG 3e
Fig. 12.8
Page 555

Definition of a domain

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

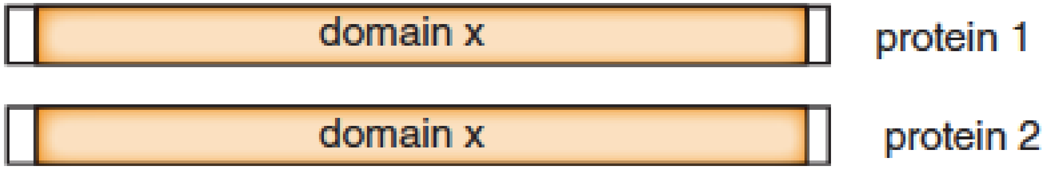
A domain is an independent structural unit, found alone or in conjunction with other domains or repeats. Domains are evolutionarily related.

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

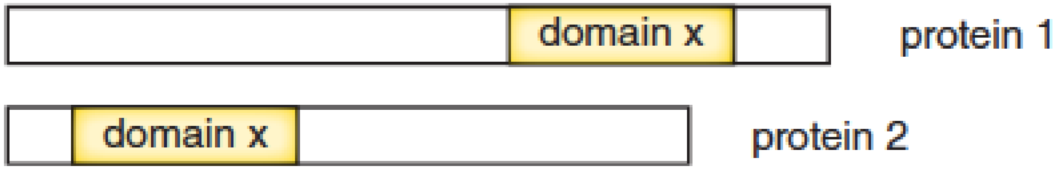
A domain is a conserved structural entity with distinctive secondary structure content and a hydrophobic core. Homologous domains with common functions usually show sequence similarities.

Varieties of protein domains

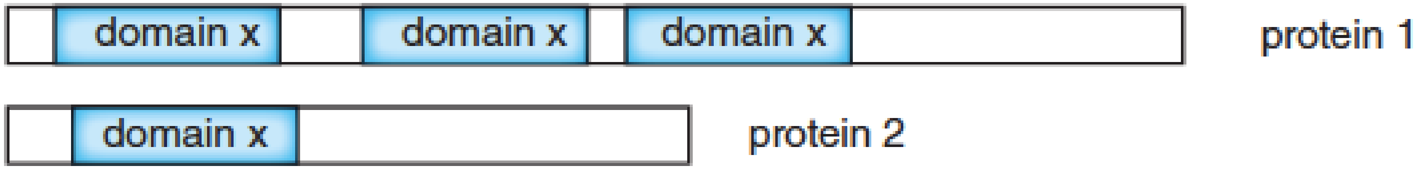
Extending along the length of a protein



Occupying a subset of a protein sequence

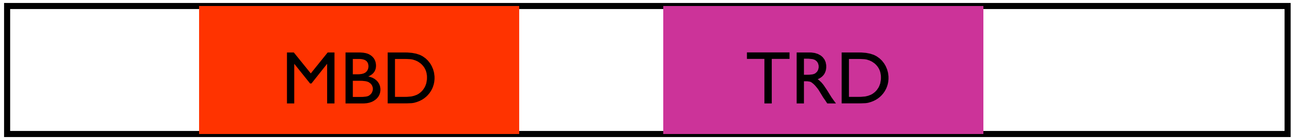


Occurring one or more times



B&FG 3e
Fig. 12.9
Page 556

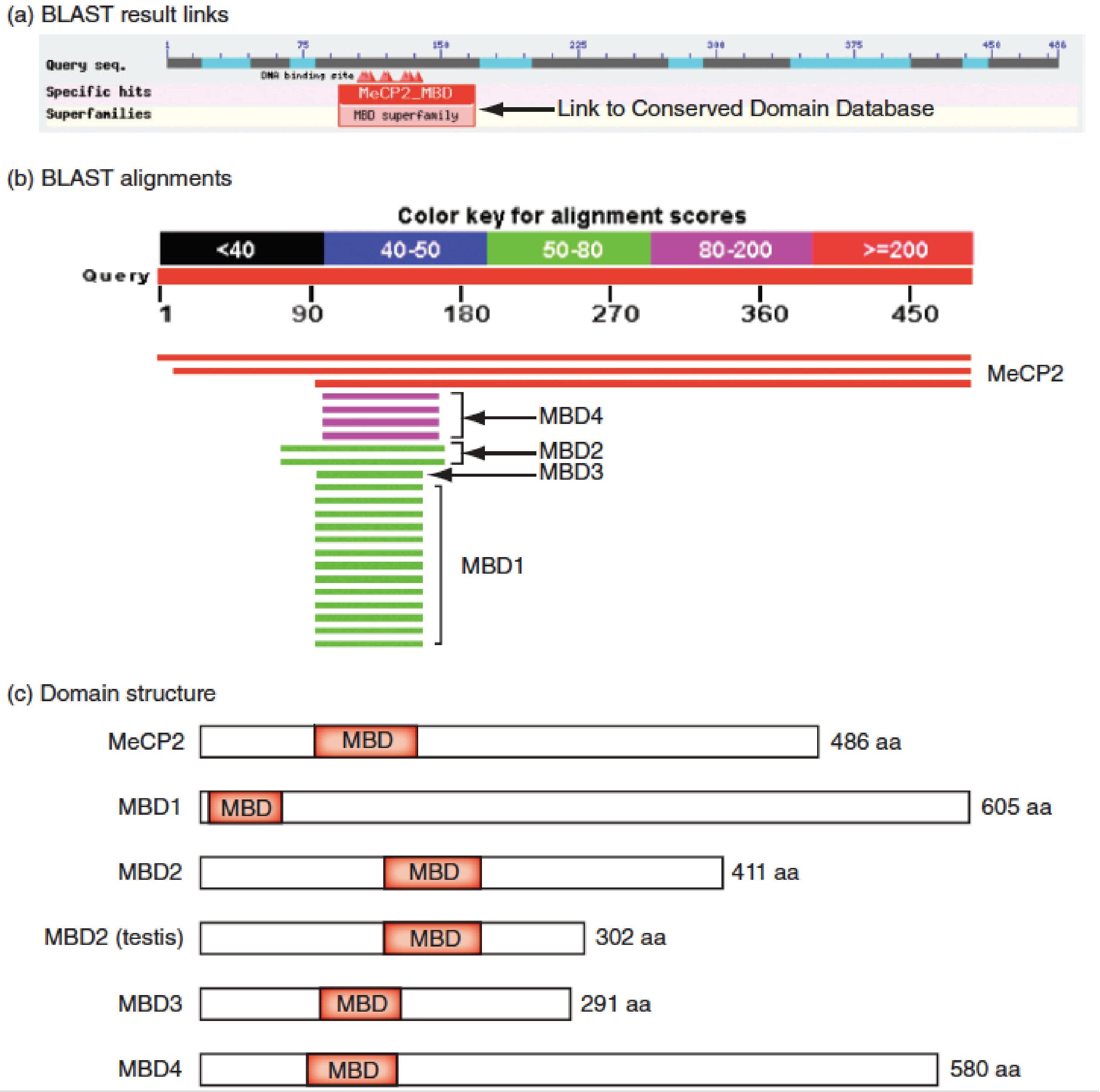
Example of a protein with domains: Methyl CpG binding protein 2 (MeCP2)



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.

Result of an MeCP2 BLASTP search: A methyl-binding domain shared by several proteins



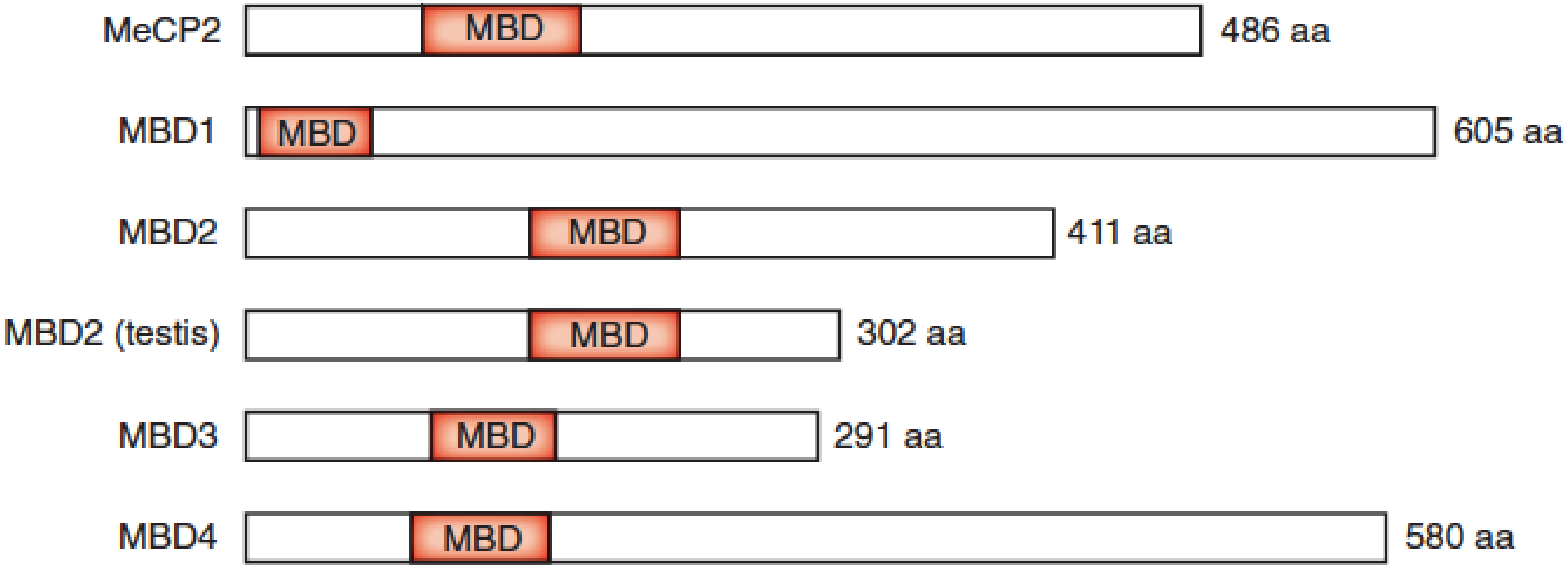
(a) BLAST result links

(b) BLAST alignments

(c) Domain structure

B&FG 3e
Fig. 12.10
Page 557

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.

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

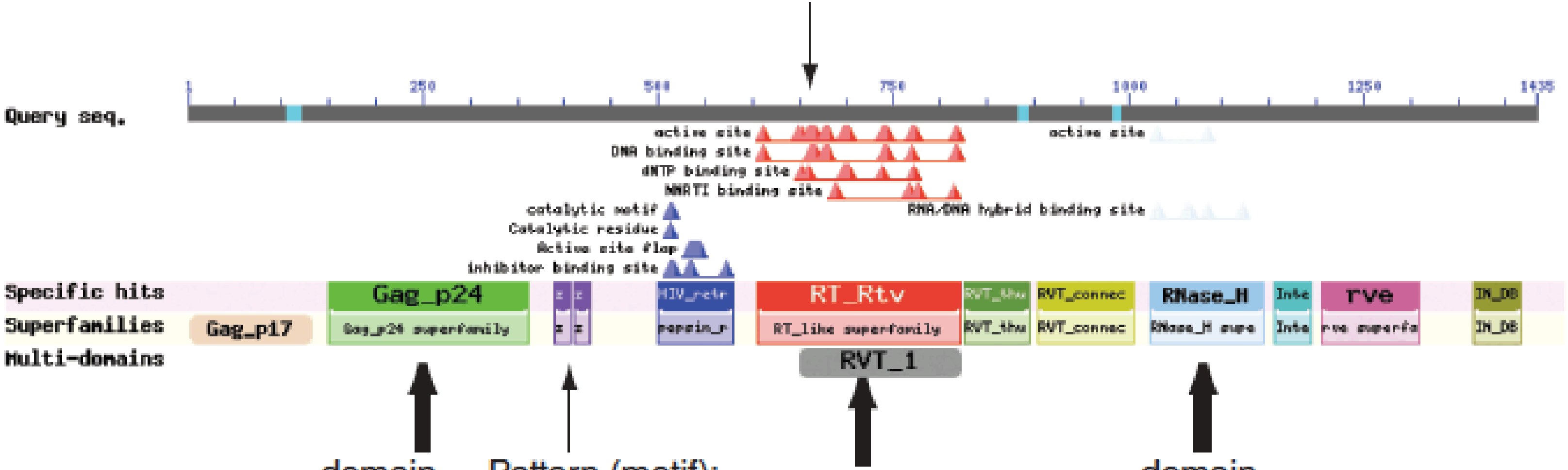
We will explore HIV-1 pol and other proteins at the Expert Protein Analysis System (ExPASy) server.

Visit www.expasy.org/

Searches for a multidomain protein: HIV gag-pol

Gag-pol at Conserved Domain Database

Pattern (motif): several amino acids within
the reverse transcriptase domain
(active site, DNA binding site, dNTP binding site)



domain
(Pfam 00607,
gag gene protein p24)

Pattern (motif):
zinc knuckle
(CX2CX4HX4C
zinc binding motif)

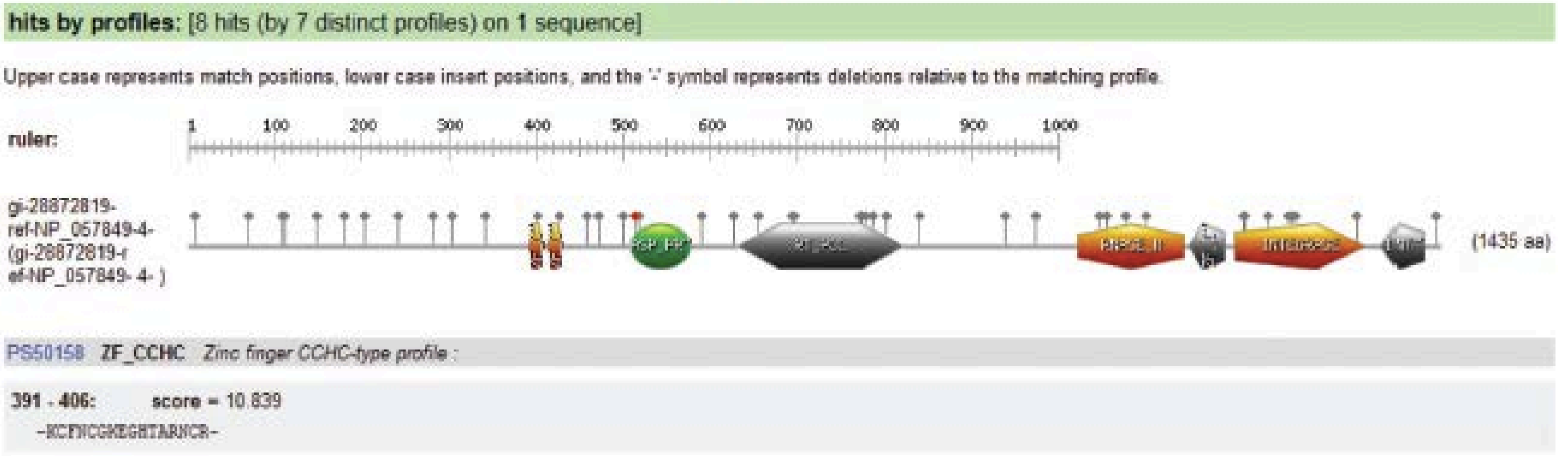
domain
(Pfam 00078,
reverse
transcriptase)

domain
(Pfam 00075,
RNase H)

B&FG 3e
Fig. 12.11
Page 558

Searches for a multidomain protein: HIV gag-pol

PROSITEscan for Gag-pol (zinc finger CCHC-type profile)



hits by profiles: [8 hits (by 7 distinct profiles) on 1 sequence]

Upper case represents match positions, lower case insert positions, and the '-' symbol represents deletions relative to the matching profile.

ruler:

gi-28872819-
ref-NP_057849-4-
(gi-28872819-r
ef-NP_057849- 4-)

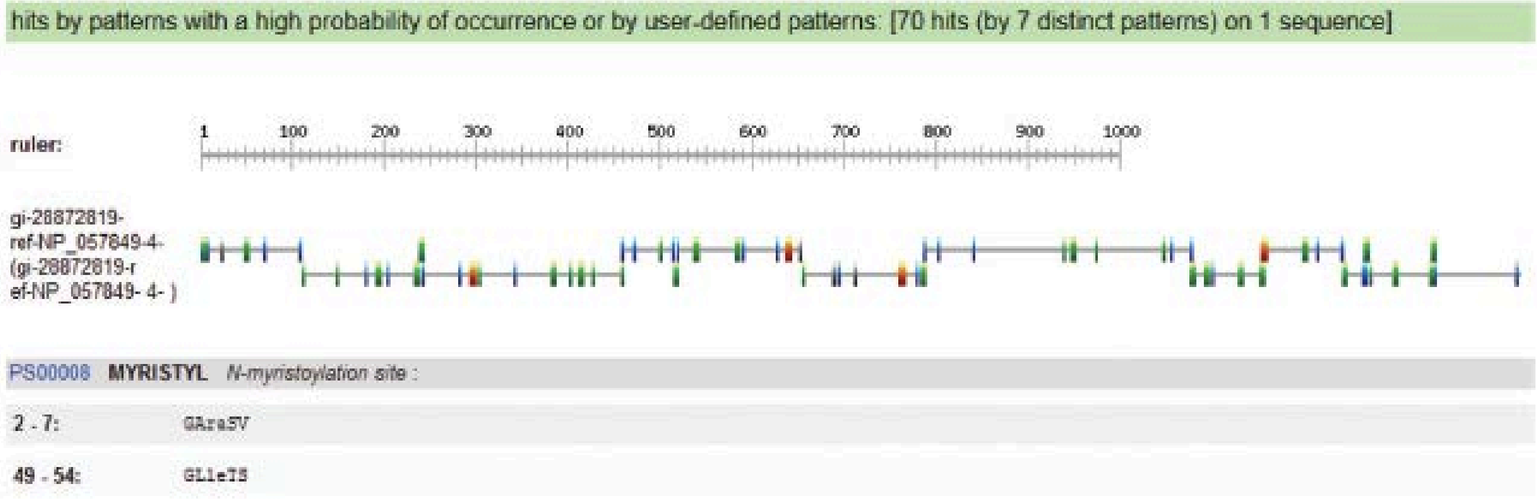
(1435 aa)

PS50158 **ZF_CCHC** *Zinc finger CCHC-type profile :*

391 - 406: **score** = 10.839

```
-KCFNCGKEGHTARNCR-
```

PROSITEscan for Gag-pol (N-myristoylation sites)



hits by patterns with a high probability of occurrence or by user-defined patterns: [70 hits (by 7 distinct patterns) on 1 sequence]

ruler:

gi-28872819-
ref-NP_057849-4-
(gi-28872819-r
ef-NP_057849- 4-)

PS00008 **MYRISTYL** *N-myristoylation site :*

2 - 7: GAraSV

49 - 54: GLleTS

B&FG 3e
Fig. 12.11
Page 558

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

Three protein databases recently merged to form UniProt:

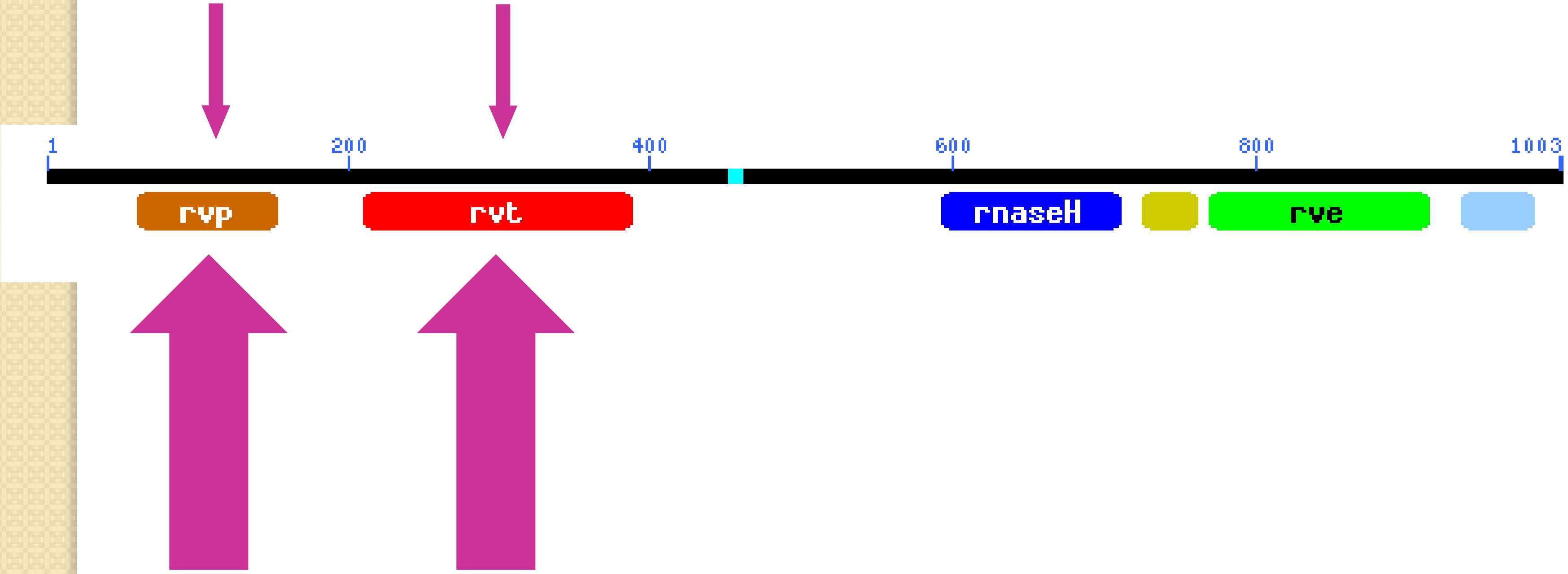
- SwissProt
- TrEMBL (translated European Molecular Biology Lab)
- Protein Information Resource (PIR)

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

Proteins can have both domains and motifs (patterns)

Motif (several residues)

Motif (several residues)



Domain (aspartyl protease)

Domain (reverse transcriptase)

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 1: Protein domains and motifs

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

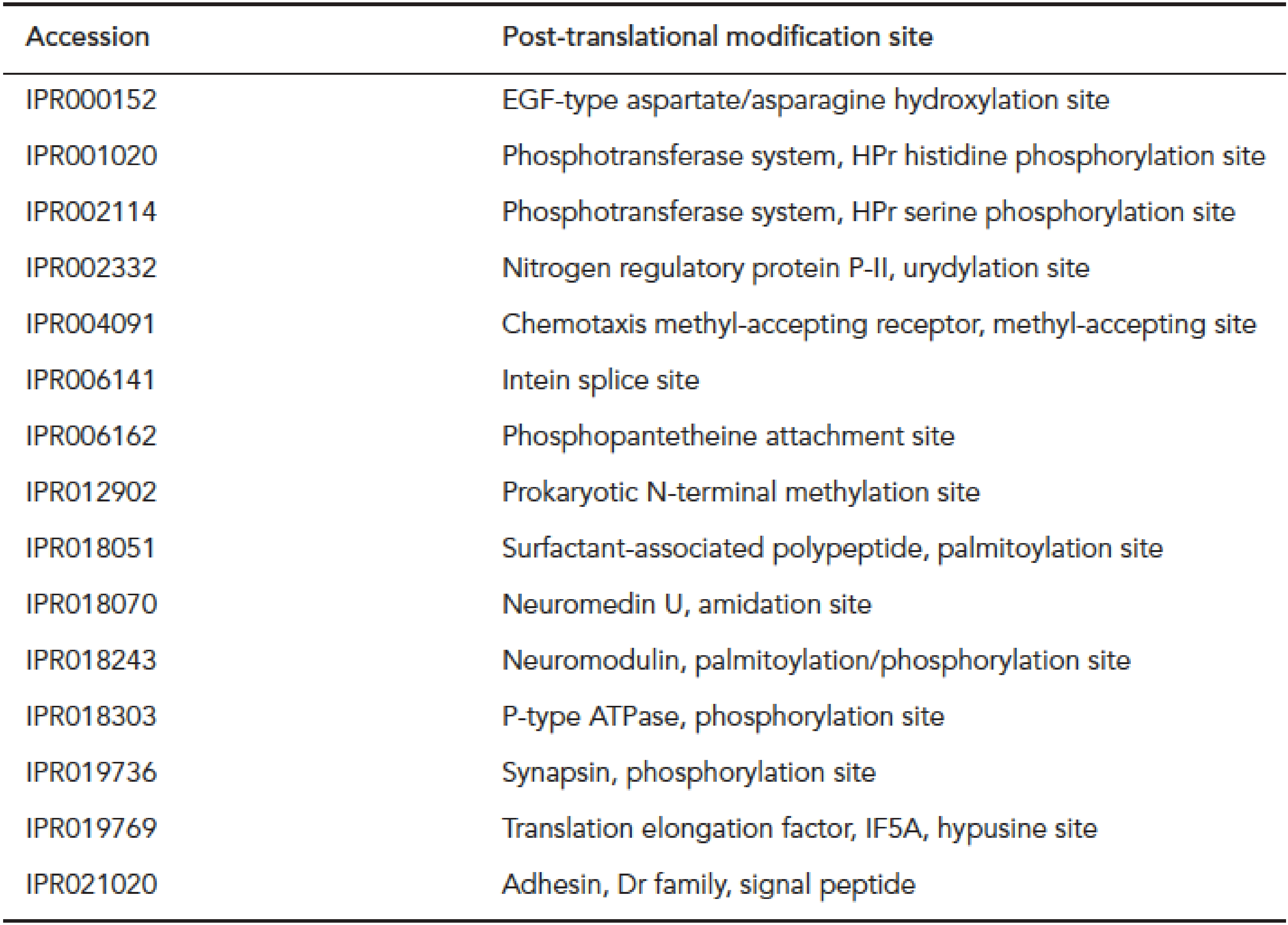
You can learn about domains in databases such as InterPro and Pfam.

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

Perspective 2:
Physical properties of proteins

B&FG 3e
Page 559

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, urydylation 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

B&FG 3e
Tab. 12.5
Page 559

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.

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

B&FG 3e
Page 566

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.

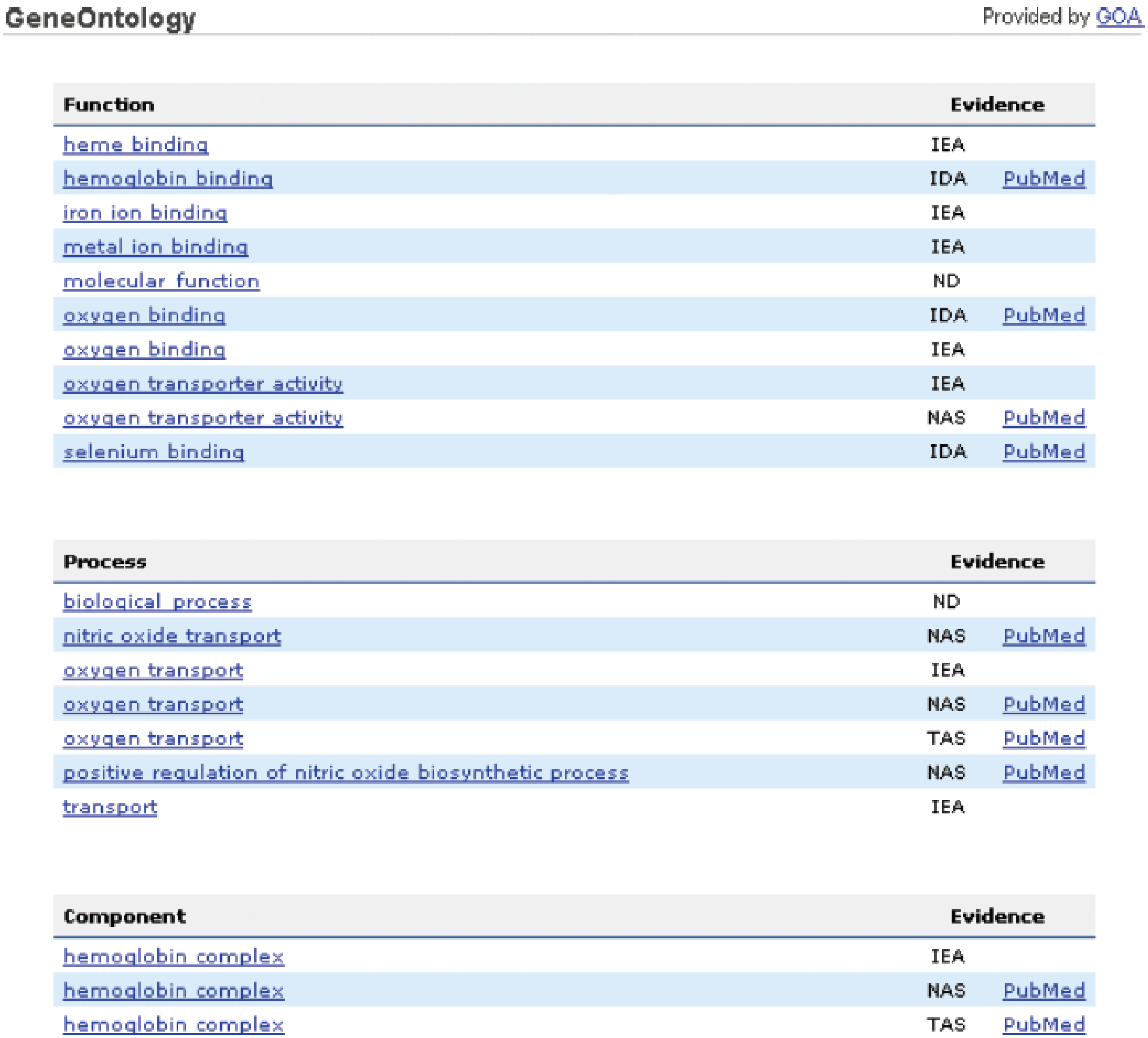
B&FG 3e
Page 566

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

B&FG 3e
Tab. 12.8
Page 568

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	

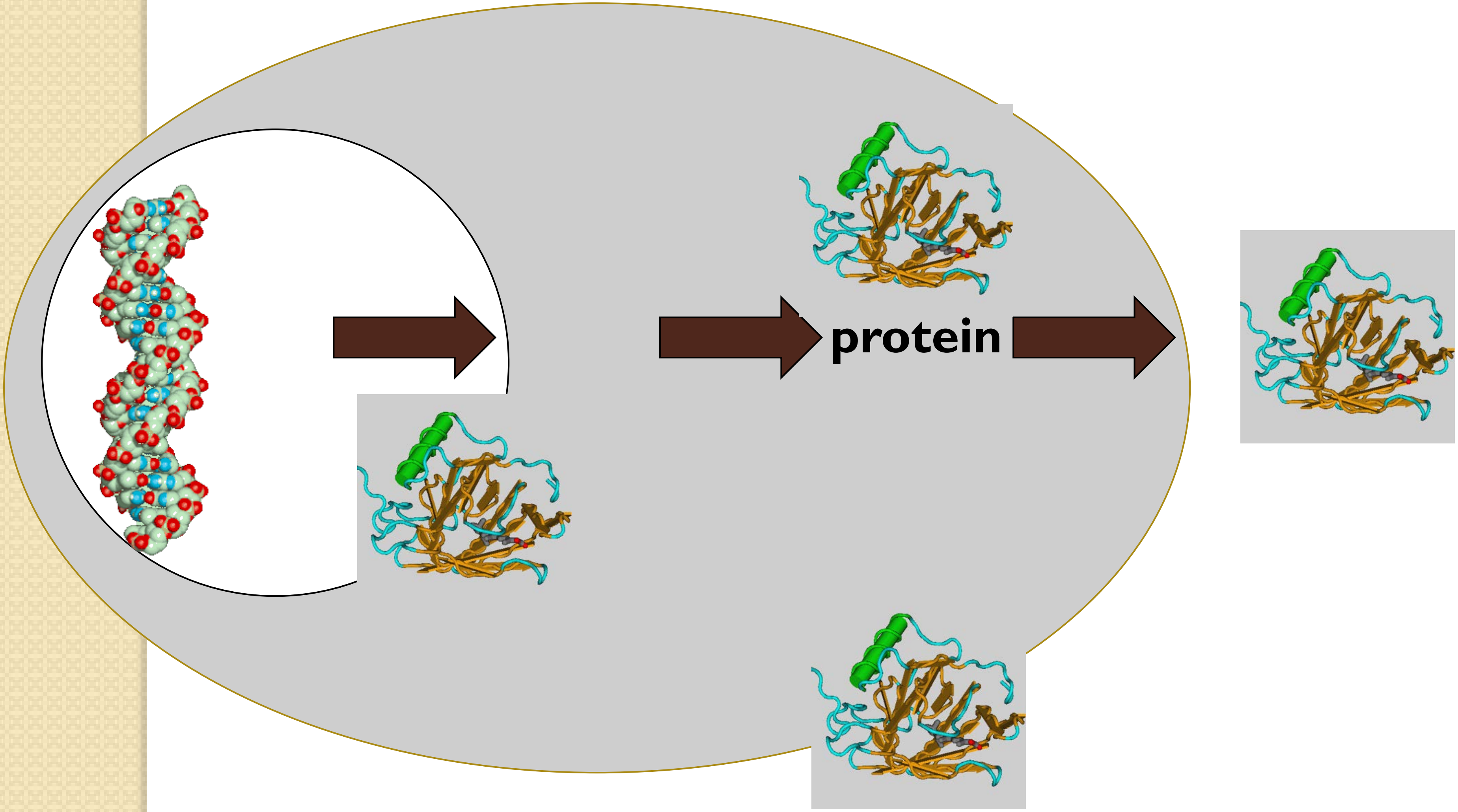
Component	Evidence	
hemoglobin complex	IEA	
hemoglobin complex	NAS	PubMed
hemoglobin complex	TAS	PubMed

B&FG 3e
Fig. 12.16
Page 569

Perspective 3:
Protein localization

B&FG 3e
Page 570

Protein localization

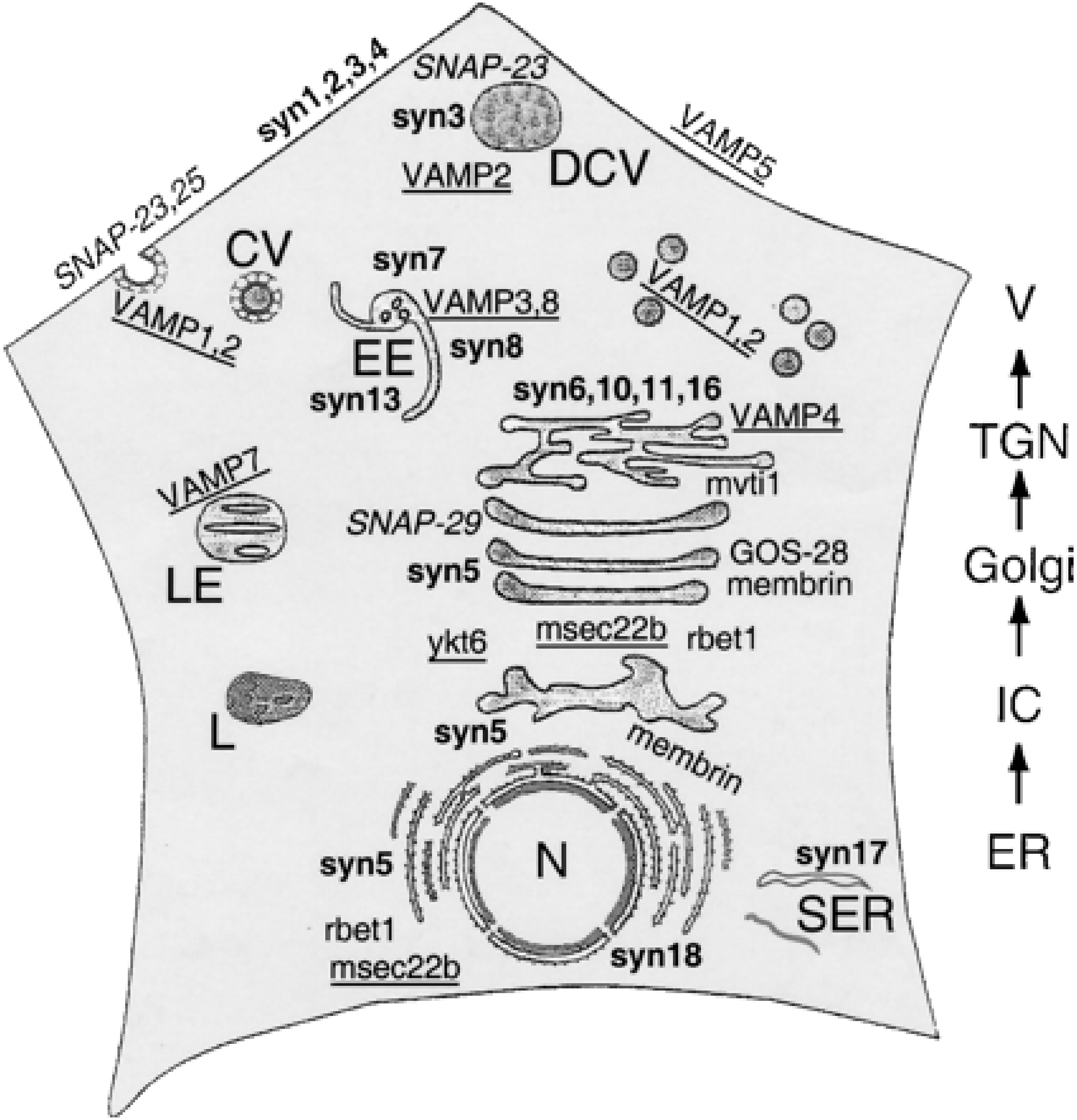


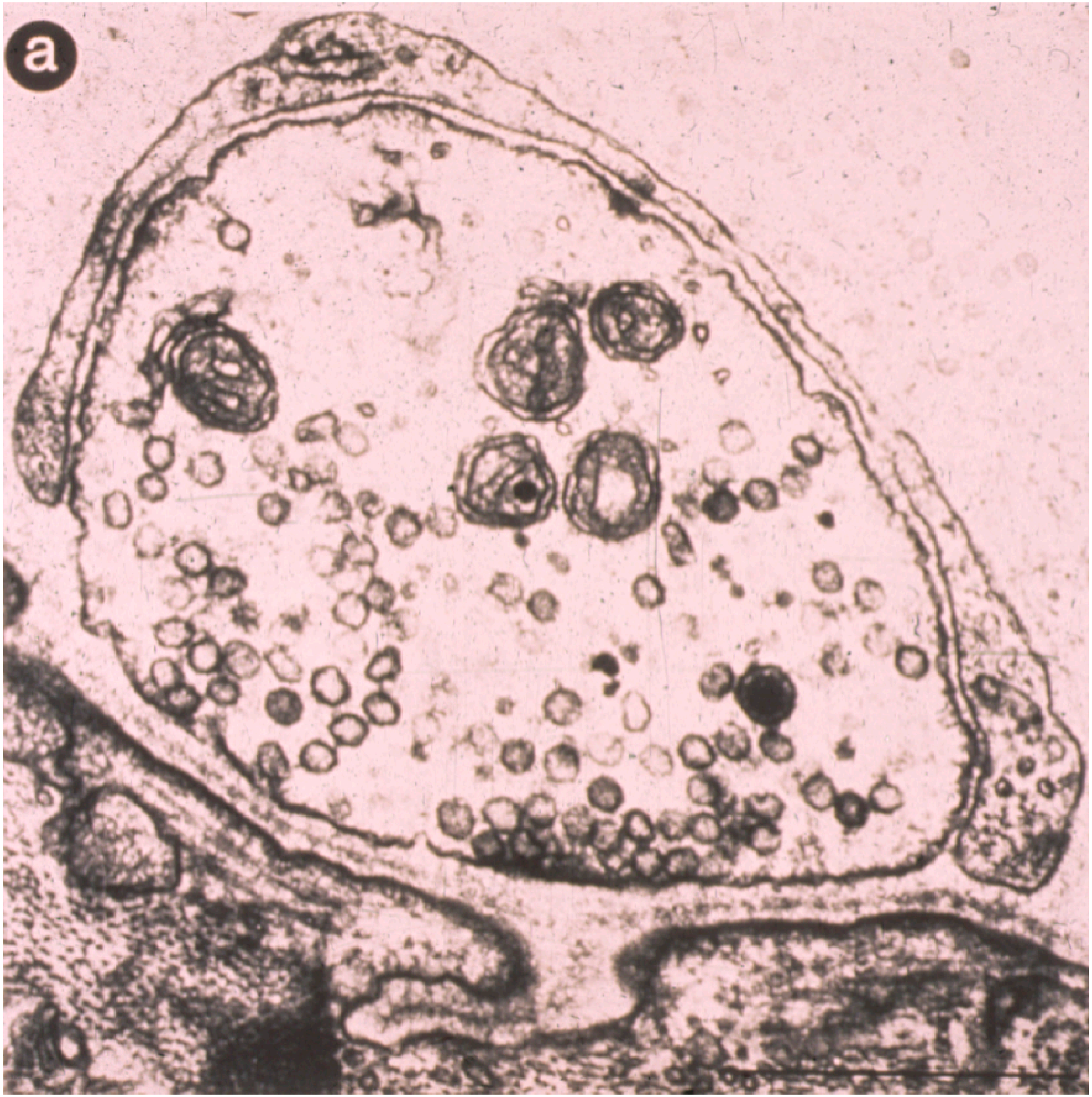
Protein localization

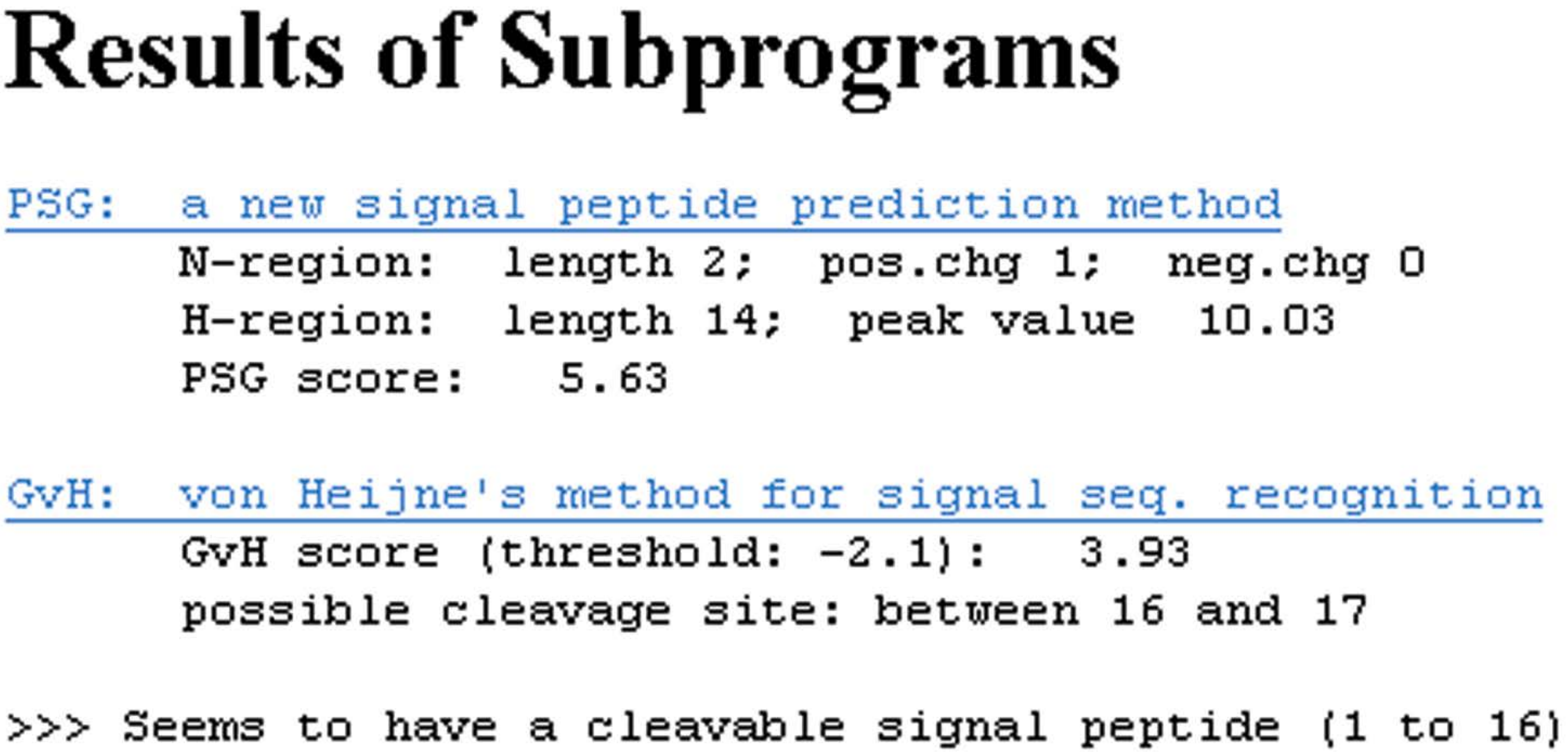
Proteins may be localized to intracellular compartments, cytosol, the plasma membrane, or they may be secreted. Many proteins shuttle between multiple compartments.

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

B&FG 3e
Page 570


syn1,2,3,4
SNAP-23
syn3
VAMP2
DCV
VAMP5
SNAP-23,25
VAMP1,2
CV
syn7
VAMP3,8
EE
syn8
syn13
VAMP1,2
syn6,10,11,16
VAMP4
mvti1
VAMP7
LE
SNAP-29
syn5
GOS-28
membrin
ykt6
msec22b
rbet1
L
syn5
membrin
N
syn5
rbet1
msec22b
syn18
syn17
SER
V
TGN
Golgi
IC
ER


a



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)
```

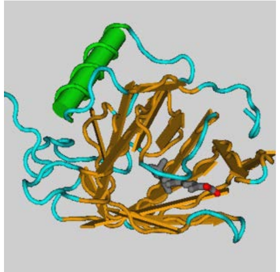
Perspective 4:
Protein function

B&FG 3e
Page 570

Protein function

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

1. Biochemical function (molecular function)



RBP binds retinol,
could be a carrier

B&FG 3e
Fig. 12.18
Page 572

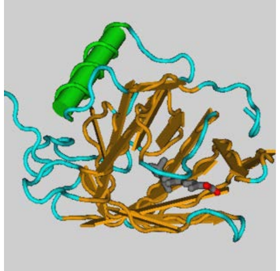
2. Functional assignment based on homology

RBP
could be
a carrier
too

Other
carrier
proteins

B&FG 3e
Fig. 12.18
Page 572

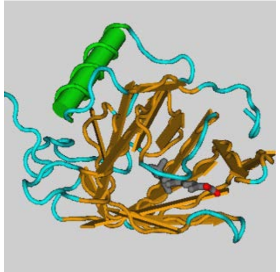
3. Function
based on structure



RBP forms a calyx

B&FG 3e
Fig. 12.18
Page 572

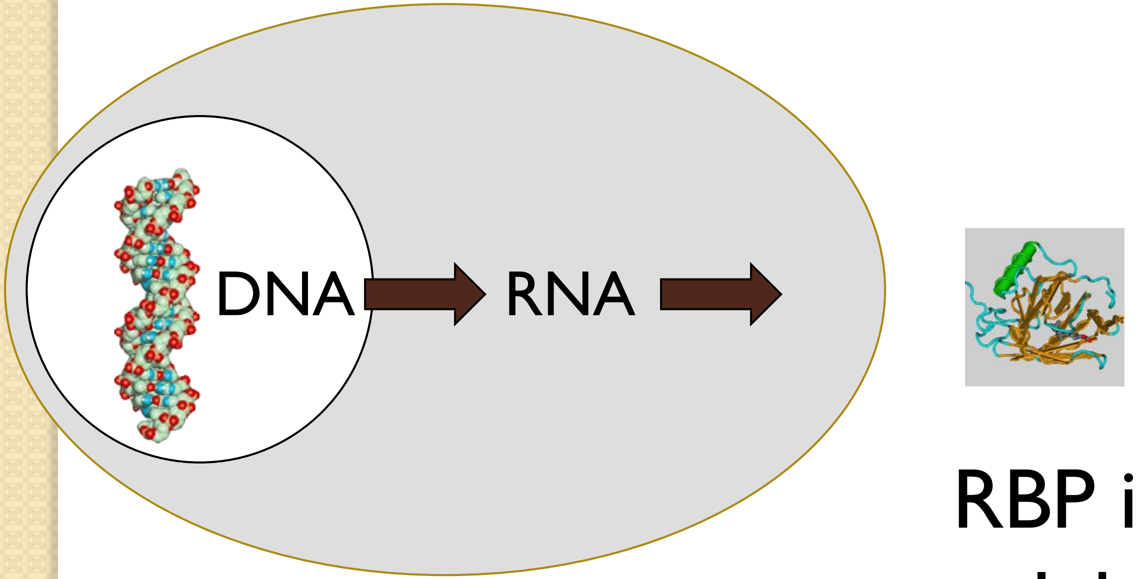
4. Function based on ligand binding specificity



RBP binds vitamin A

B&FG 3e
Fig. 12.18
Page 572

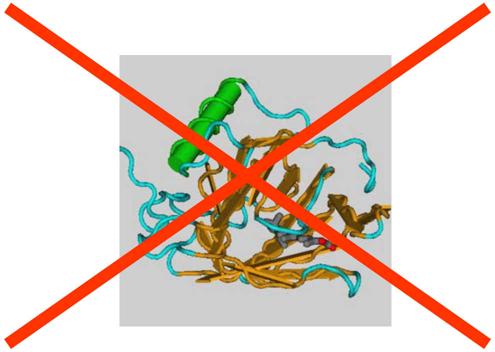
5. Function based on cellular process



RBP is abundant, soluble, secreted

B&FG 3e
Fig. 12.18
Page 572

6. Function based on biological process



RBP is essential for vision

B&FG 3e
Fig. 12.18
Page 572

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

B&FG 3e
Fig. 12.18
Page 572

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.

B&FG 3e
Page 574

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.

B&FG 3e
Page 574