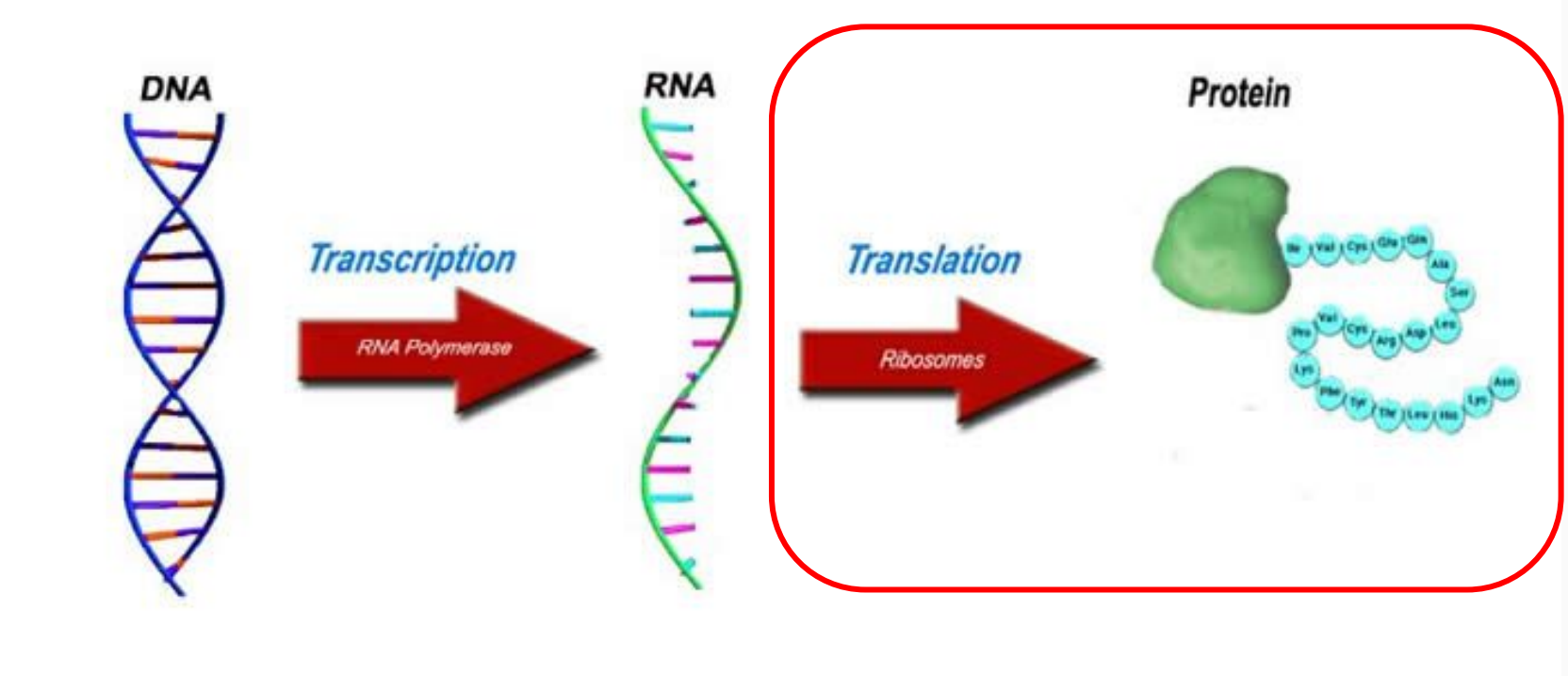
Proteomics

BF528

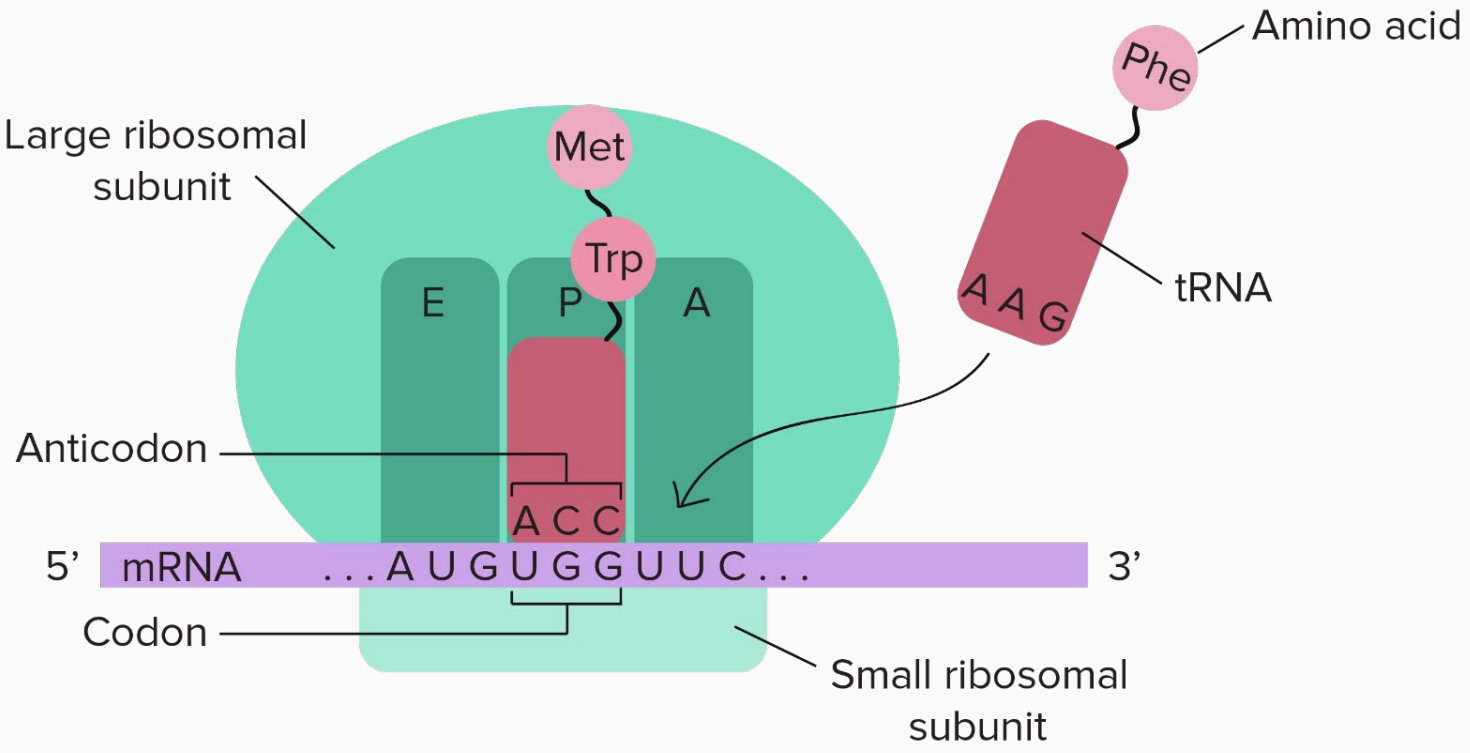
Life Is Made Of Proteins

- 40%-50% of dry mass of cells are proteins
- The “work horses” of the cell
- Participate in essentially all biological functions
- Amino acid precursors likely first biomolecules
- Cause of many human diseases

Central Dogma



Translating RNA to Amino Acids



RNA Sequence Implies Protein Sequence

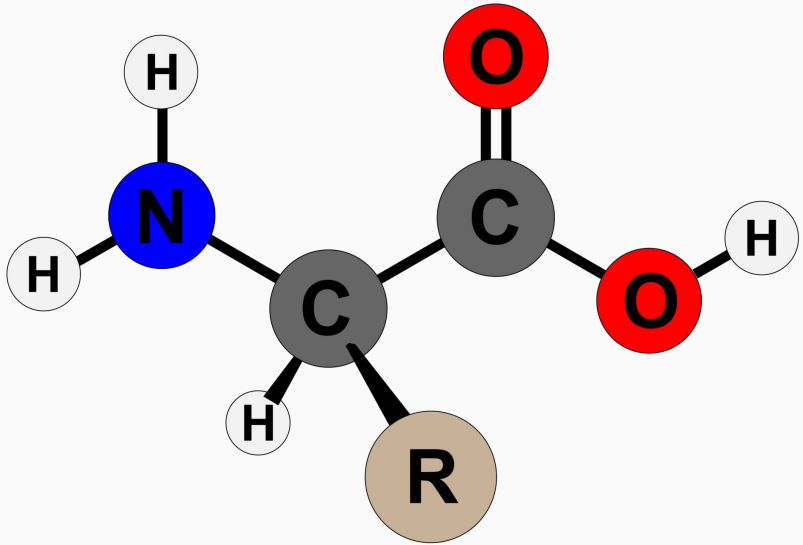
1st base	2nd base								3rd base
	T		C		A		G		
T	TTT	(Phe/F) Phenylalanine †	TCT	(Ser/S) Serine †	TAT	(Tyr/Y) Tyrosine †	TGT	(Cys/C) Cysteine †	T
	TTC		TCC		TAC		TGC		C
	TTA	(Leu/L) Leucine †	TCA		TAA	Stop (*Ochre*) *[note 2]	TGA	Stop (*Opal*) *[note 2]	A
	TTG →		TCG		TAG	Stop (*Amber*) *[note 2]	TGG	(Trp/W) Tryptophan †	G
C	CTT		CCT	(Pro/P) Proline †	CAT	(His/H) Histidine ‡	CGT	(Arg/R) Arginine ‡	T
	CTC		CCC		CAC		CGC		C
	CTA		CCA		CAA	(Gln/Q) Glutamine †	CGA		A
	CTG		CCG		CAG		CGG		G
A	ATT	(Ile/I) Isoleucine †	ACT	(Thr/T) Threonine †	AAT	(Asn/N) Asparagine †	AGT	(Ser/S) Serine †	T
	ATC		ACC		AAC		AGC		C
	ATA		ACA		AAA	(Lys/K) Lysine ‡	AGA	(Arg/R) Arginine ‡	A
	ATG →	(Met/M) Methionine †	ACG		AAG		AGG		G
G	GTT	(Val/V) Valine †	GCT	(Ala/A) Alanine †	GAT	(Asp/D) Aspartic acid ↓	GGT	(Gly/G) Glycine †	T
	GTC		GCC		GAC		GGC		C
	GTA		GCA		GAA	(Glu/E) Glutamic acid ↓	GGA		A
	GTG →		GCG		GAG		GGG		G

Amino Acids and their Properties

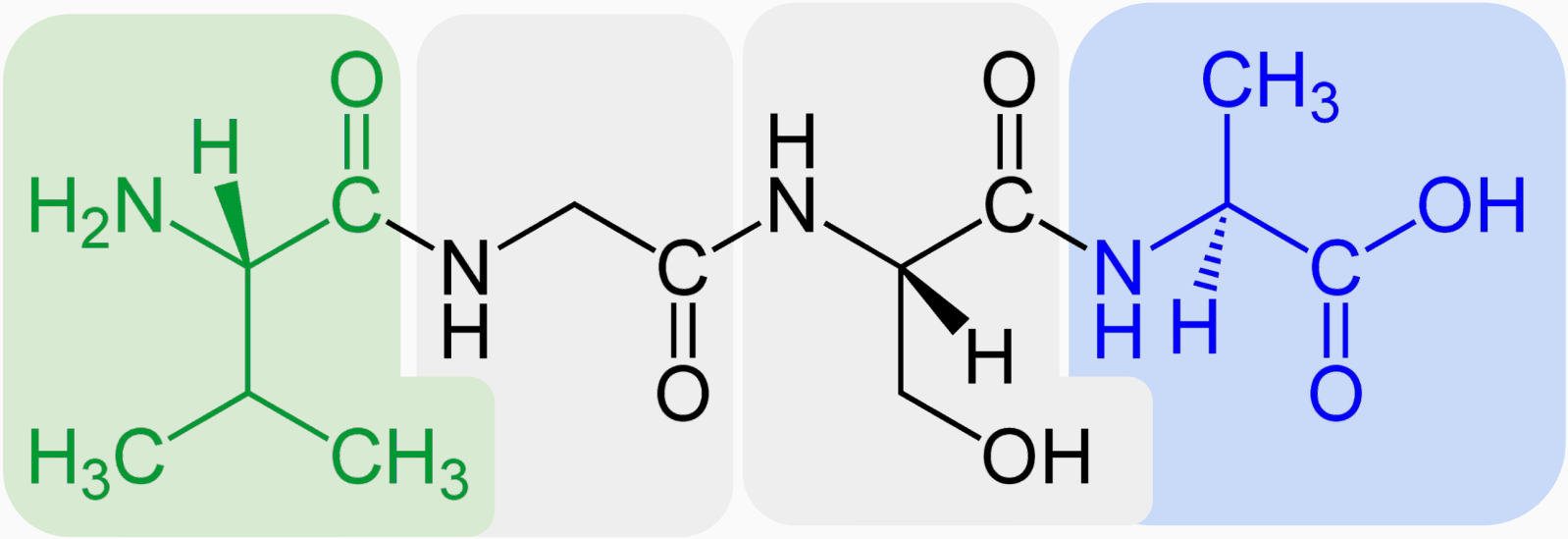
Name	Abbr.		Molecular Weight	Molecular Formula	Residue Formula	Residue Weight (-H_2O)	pKa[1]	pKb[2]	pKx[3]	pI[4]
Alanine	Ala	A	89.10	$C_3H_7NO_2$	C_3H_5NO	71.08	2.34	9.69	–	6.00
Arginine	Arg	R	174.20	$C_6H_{14}N_4O_2$	$C_6H_{12}N_4O$	156.19	2.17	9.04	12.48	10.76
Asparagine	Asn	N	132.12	$C_4H_8N_2O_3$	$C_4H_6N_2O_2$	114.11	2.02	8.80	–	5.41
Aspartic acid	Asp	D	133.11	$C_4H_7NO_4$	$C_4H_5NO_3$	115.09	1.88	9.60	3.65	2.77
Cysteine	Cys	C	121.16	$C_3H_7NO_2S$	C_3H_5NOS	103.15	1.96	10.28	8.18	5.07
Glutamic acid	Glu	E	147.13	$C_5H_9NO_4$	$C_5H_7NO_3$	129.12	2.19	9.67	4.25	3.22
Glutamine	Gln	Q	146.15	$C_5H_{10}N_2O_3$	$C_5H_8N_2O_2$	128.13	2.17	9.13	–	5.65
Glycine	Gly	G	75.07	$C_2H_5NO_2$	C_2H_3NO	57.05	2.34	9.60	–	5.97
Histidine	His	H	155.16	$C_6H_9N_3O_2$	$C_6H_7N_3O$	137.14	1.82	9.17	6.00	7.59
Hydroxyproline	Hyp	O	131.13	$C_5H_9NO_3$	$C_5H_7NO_2$	113.11	1.82	9.65	–	–
Isoleucine	Ile	I	131.18	$C_6H_{13}NO_2$	$C_6H_{11}NO$	113.16	2.36	9.60	–	6.02
Leucine	Leu	L	131.18	$C_6H_{13}NO_2$	$C_6H_{11}NO$	113.16	2.36	9.60	–	5.98
Lysine	Lys	K	146.19	$C_6H_{14}N_2O_2$	$C_6H_{12}N_2O$	128.18	2.18	8.95	10.53	9.74
Methionine	Met	M	149.21	$C_5H_{11}NO_2S$	C_5H_9NOS	131.20	2.28	9.21	–	5.74
Phenylalanine	Phe	F	165.19	$C_9H_{11}NO_2$	C_9H_9NO	147.18	1.83	9.13	–	5.48
Proline	Pro	P	115.13	$C_5H_9NO_2$	C_5H_7NO	97.12	1.99	10.60	–	6.30
Pyroglutamatic	Glp	U	139.11	$C_5H_7NO_3$	$C_5H_5NO_2$	121.09	–	–	–	5.68
Serine	Ser	S	105.09	$C_3H_7NO_3$	$C_3H_5NO_2$	87.08	2.21	9.15	–	5.68
Threonine	Thr	T	119.12	$C_4H_9NO_3$	$C_4H_7NO_2$	101.11	2.09	9.10	–	5.60
Tryptophan	Trp	W	204.23	$C_{11}H_{12}N_2O_2$	$C_{11}H_{10}N_2O$	186.22	2.83	9.39	–	5.89
Tyrosine	Tyr	Y	181.19	$C_9H_{11}NO_3$	$C_9H_9NO_2$	163.18	2.20	9.11	10.07	5.66
Valine	Val	V	117.15	$C_5H_{11}NO_2$	C_5H_9NO	99.13	2.32	9.62	–	5.96

Protein Primer

- Linear sequences of amino acids
- Directional:
  - N-terminus, C-terminus
- Polypeptide backbone
- Post translational modifications

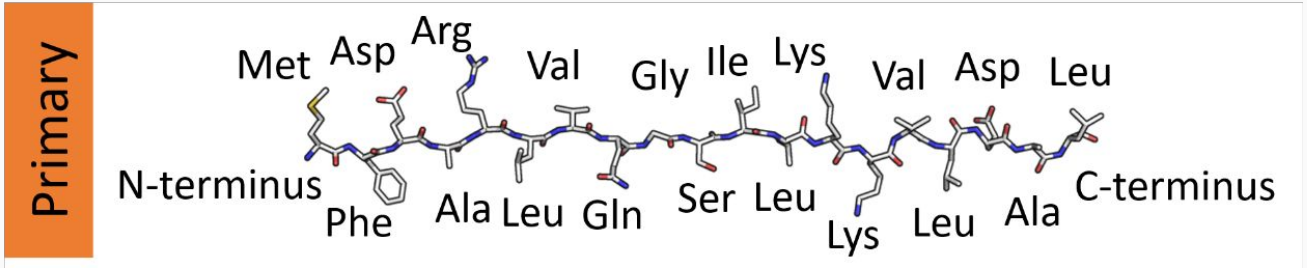


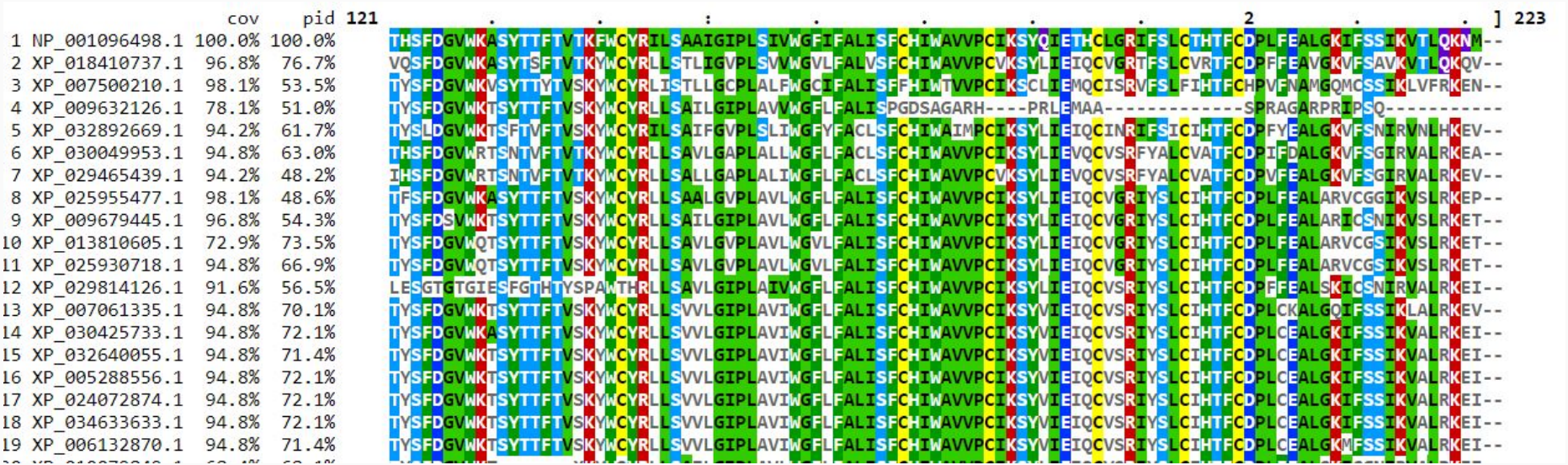
Peptides



Valine Glycine Serine Alanine

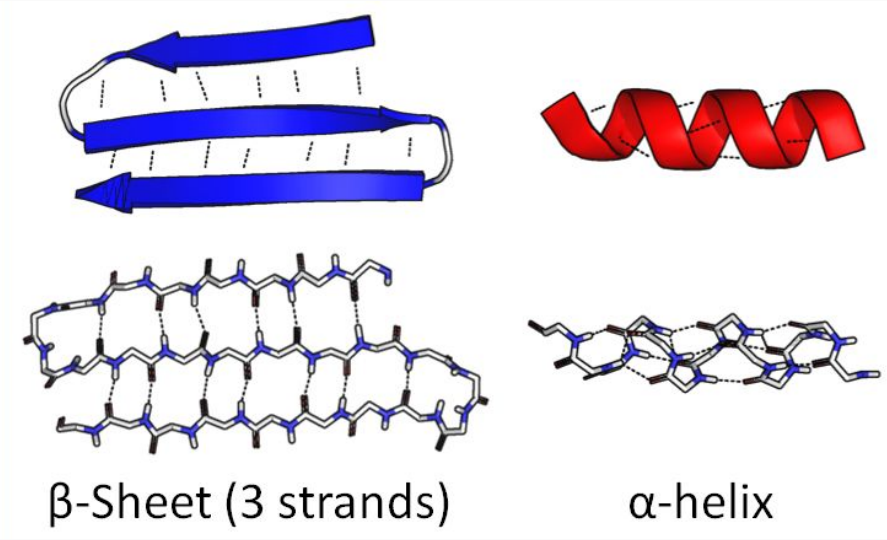
Proteins Have Sequence

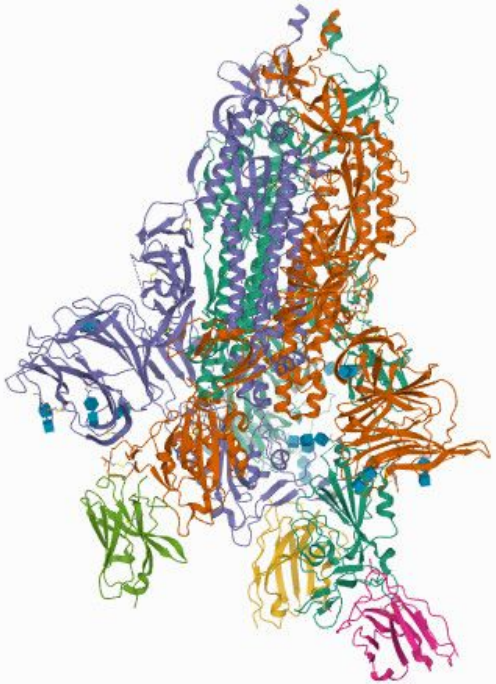




Proteins Have Structure

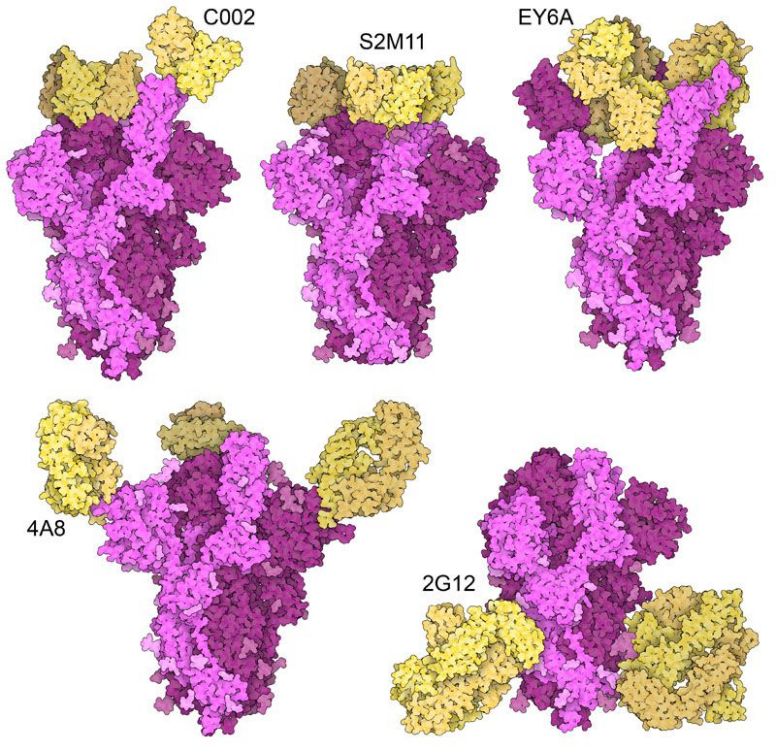
Secondary





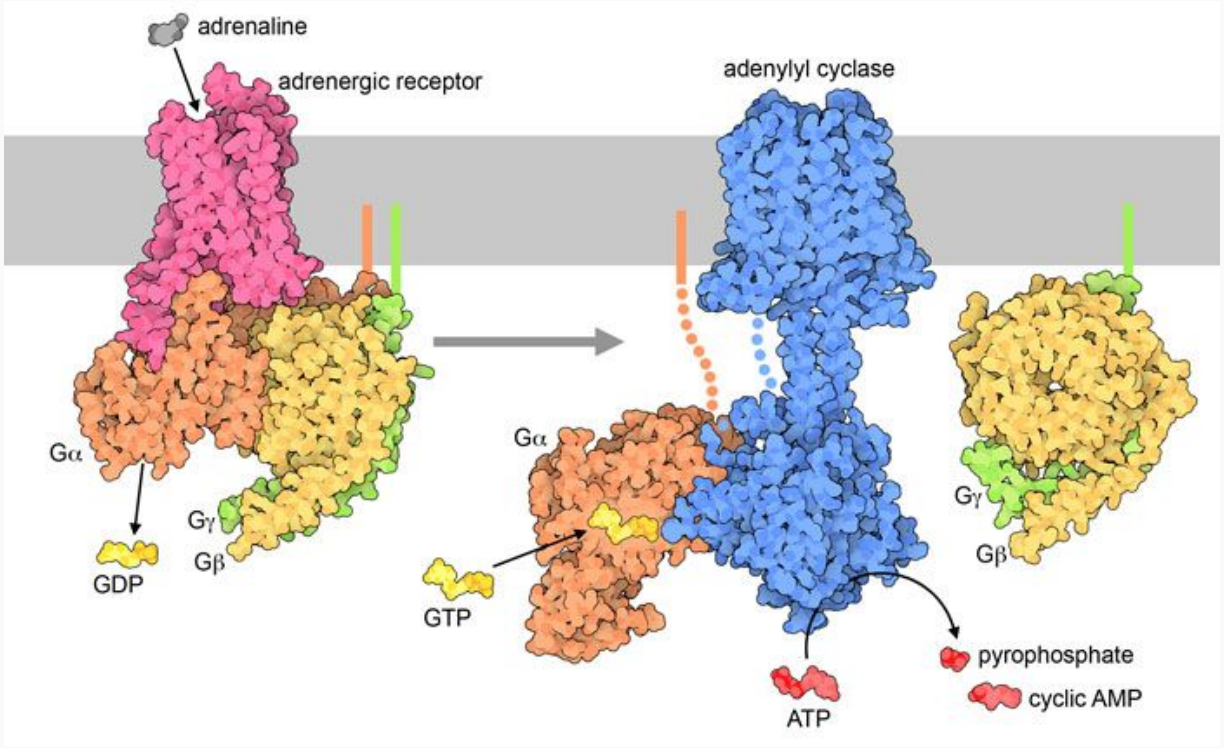
Cryo-EM structure of the SARS-CoV-2 spike protein bound to neutralizing nanobodies (Ty1)
https://www.rcsb.org/structure/6zxn

Proteins Have Shape



Structures of antibody Fab fragments (yellow) bound to SARS-CoV-2 spike protein (magenta). All structures include the variable domains of the antibodies, and some of the structures also include constant domains. http://pdb101.rcsb.org/motm/256

Proteins Have Function



Signaling with G-proteins. Hormones like adrenaline bind to a GPCR receptor (left), which binds to a heterotrimeric G-protein and releases GDP. Then the G-protein separates into two pieces, and the G-alpha subunit binds to GTP and activates adenylyl cyclase (right).

Measuring Proteins Indirectly

- Infer protein sequence from DNA sequence
  - Gene model prediction
  - Homology
- Infer protein sequence from RNA sequence
- Infer protein abundance from expression
- Immunohistochemistry
- Others...

SNPs and Variants

dbSNP Short Genetic Variations

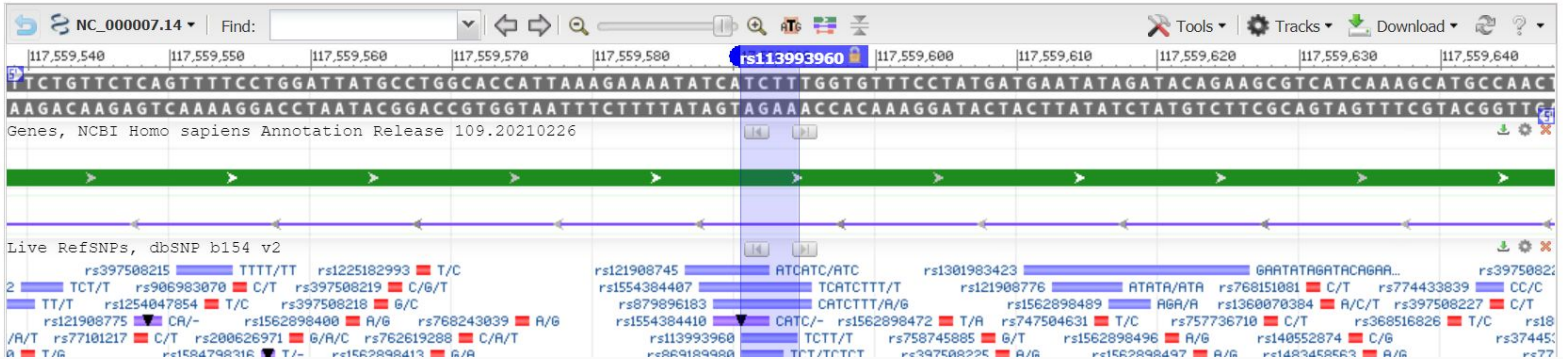
Search for terms | Search

Examples: rs268, BRCA1 and more | Advanced search

rs113993960 Cystic fibrosis transmembrane conductance regulator (CTFR)

Current Build 154
Released April 21, 2020

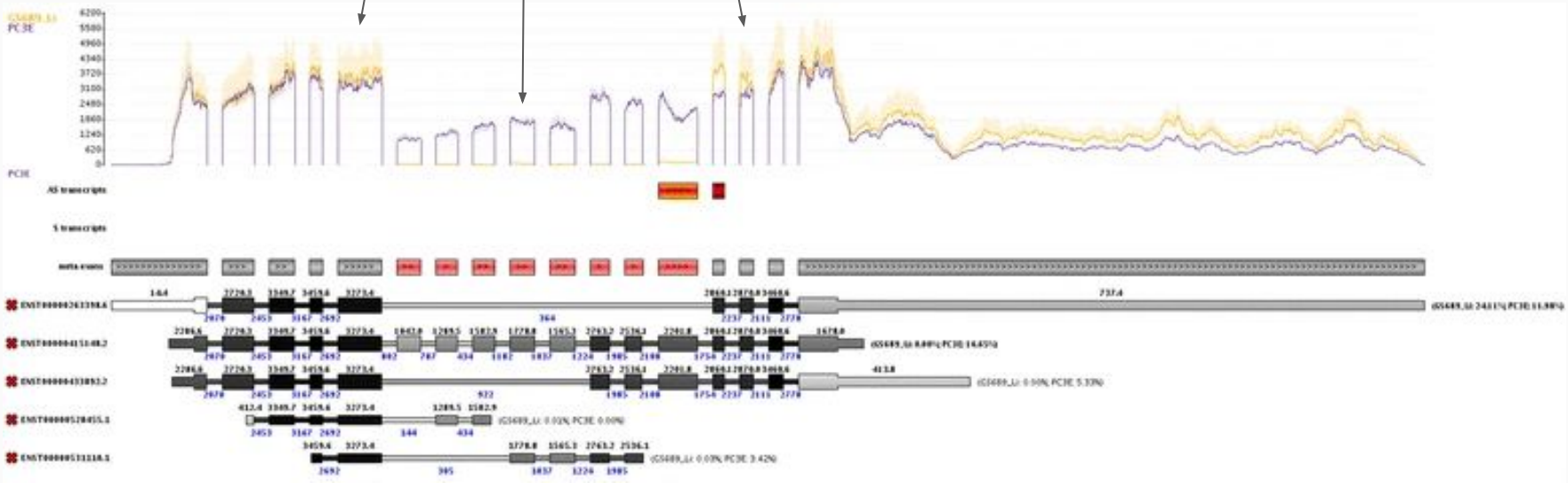
Organism	*Homo sapiens*	**Clinical Significance**	Reported in ClinVar
Position	chr7:117559591-117559594 (GRCh38.p12)	**Gene : Consequence**	CFTR : Inframe Deletion CFTR-AS1 : Intron Variant
Alleles	delCTT	**Publications**	52 citations
Variation Type	Indel Insertion and Deletion		LitVar 100



Loss of phenylalanine (F) residue at amino acid position 508

RNA Isoforms

Alternative splicing alters amino acid sequence

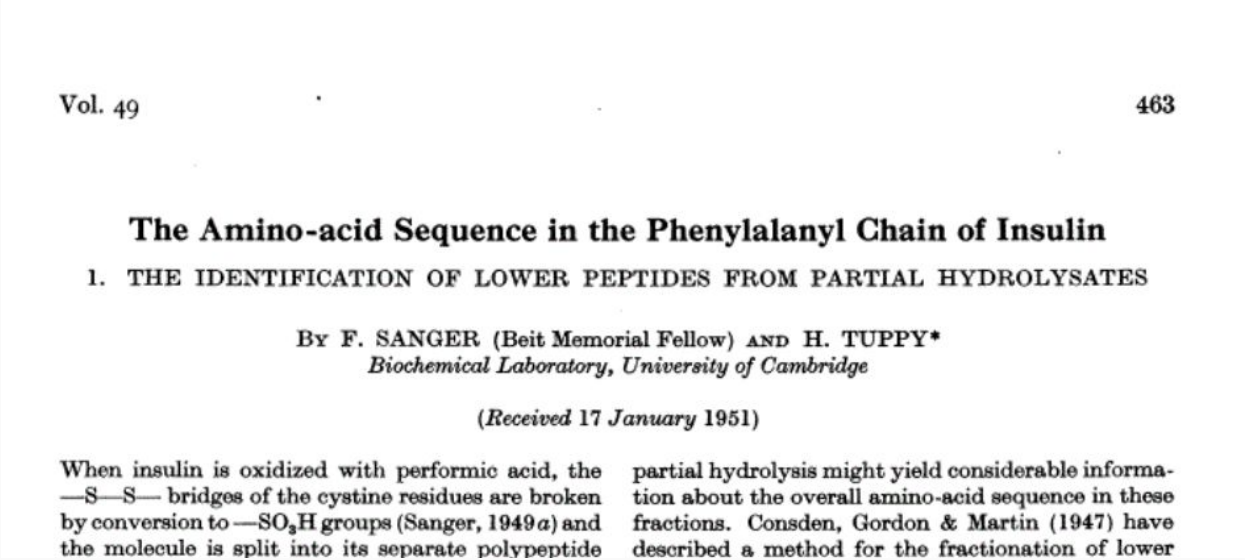


Proteomics

“Proteomics is the large-scale study of proteins. Proteins are vital parts of living organisms, with many functions. The **proteome** is the entire set of proteins that is produced or modified by an organism or system.” - Wikipedia

Proteomics History

First biological sequences ever determined were amino acids!



Vol. 49 463

The Amino-acid Sequence in the Phenylalanyl Chain of Insulin

1. THE IDENTIFICATION OF LOWER PEPTIDES FROM PARTIAL HYDROLYSATES

By F. SANGER (Beit Memorial Fellow) and H. TUPPY*
Biochemical Laboratory, University of Cambridge

(*Received* 17 *January* 1951)

When insulin is oxidized with performic acid, the —S—S— bridges of the cystine residues are broken by conversion to —SO_3H groups (Sanger, 1949*a*) and the molecule is split into its separate polypeptide

partial hydrolysis might yield considerable information about the overall amino-acid sequence in these fractions. Consden, Gordon & Martin (1947) have described a method for the fractionation of lower

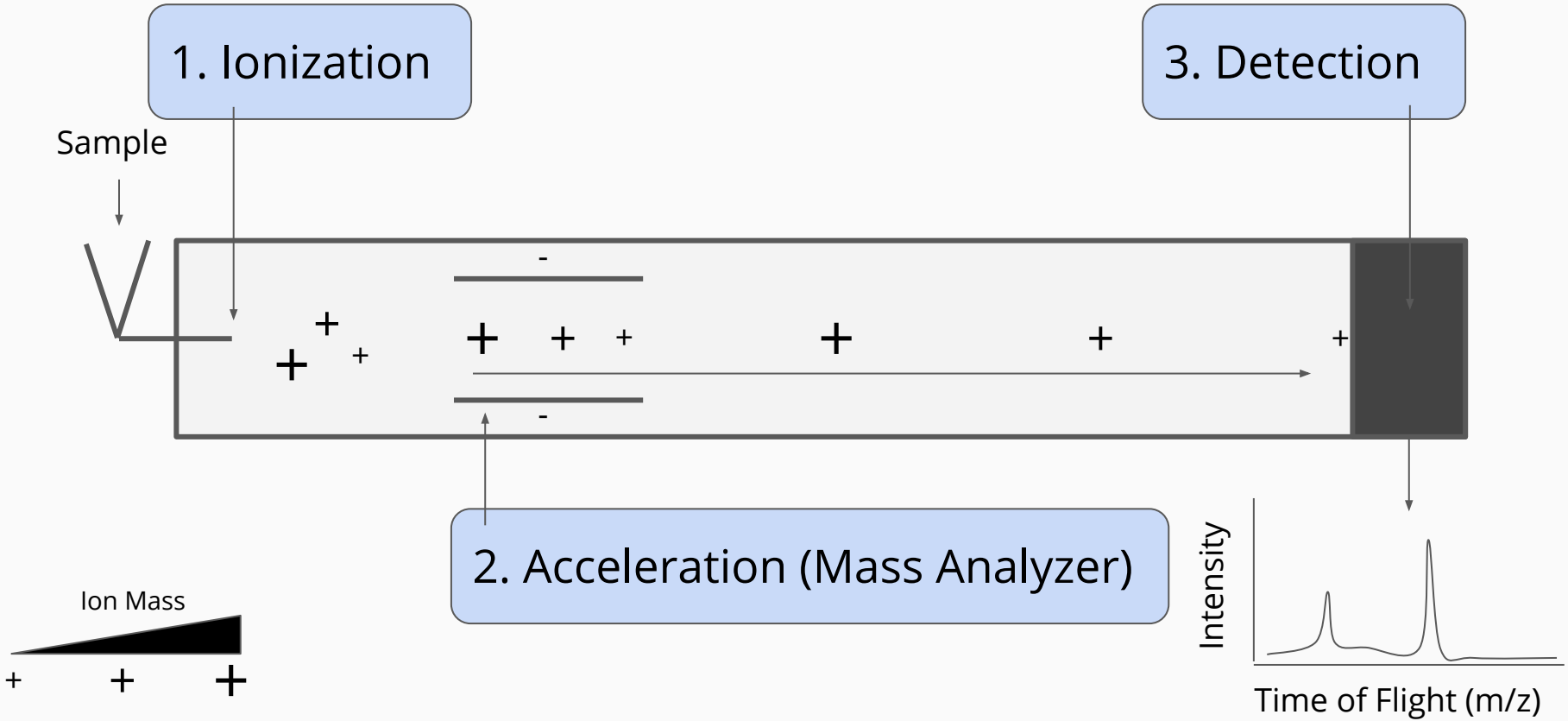
Measuring Proteins Directly

- Although proteins are sequences, no current technology can directly sequence amino acids
- Modular nature of amino acids enables decomposition, measurement, and “puzzling together” of complete proteins

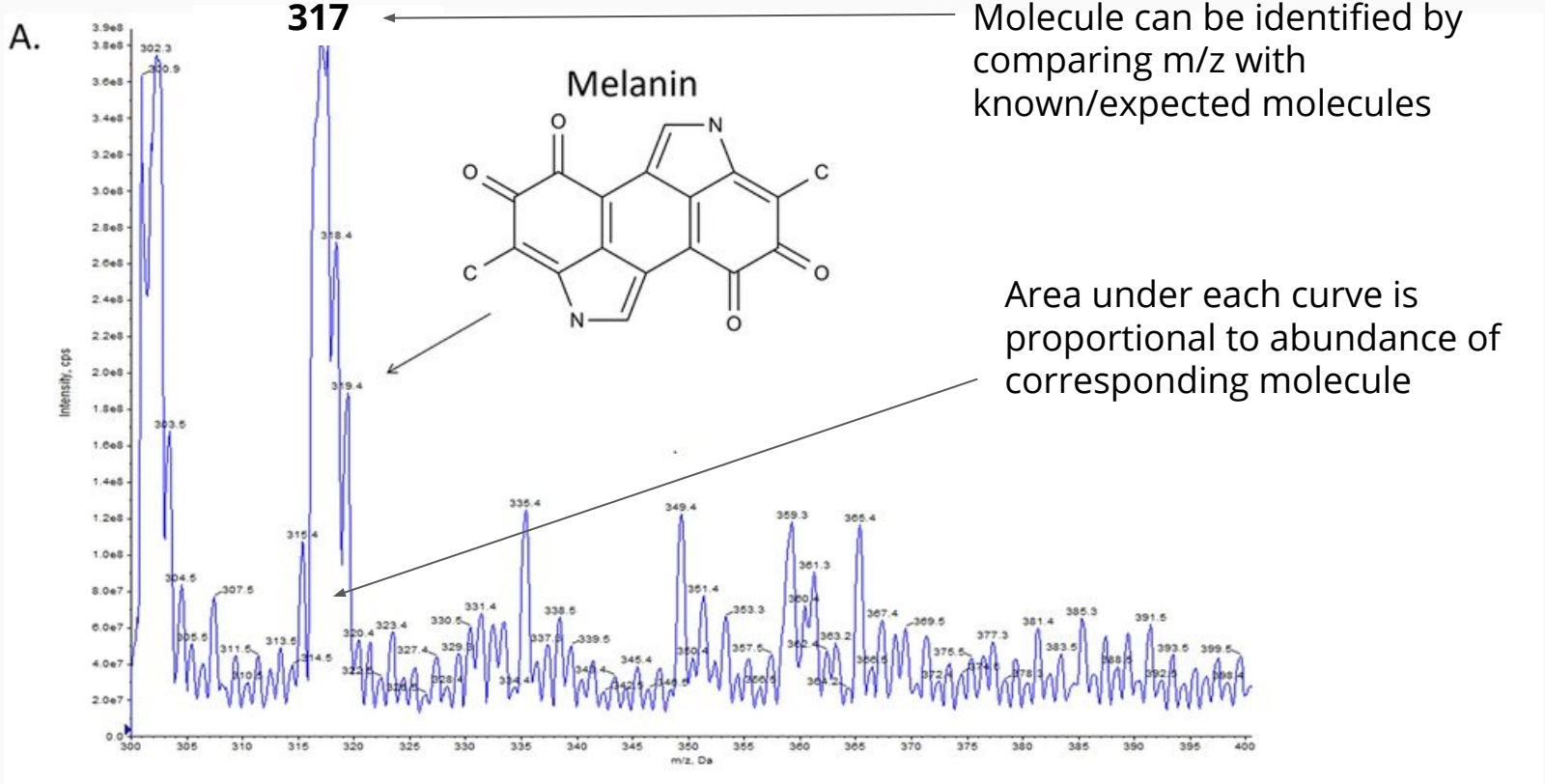
Mass Spectrometry (MS)

- Measures mass and charge of molecules
- Measures any ions
  - not just proteins/peptides!
- **Only** measures ions
- Can measure both **identity** and **abundance** of ions

Conceptual MS Architecture



Mass/Charge Spectra



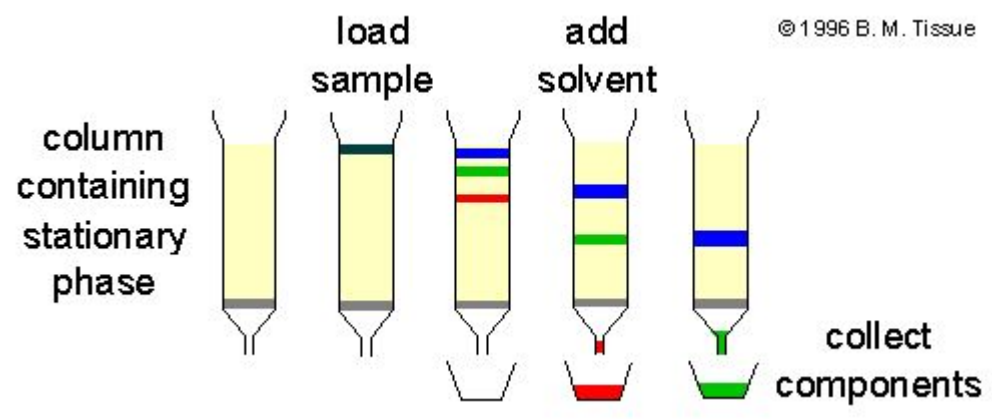
Mass Spec Proteomics Workflow

1. Sample collection/processing
2. Fractionation (Liquid Chromatography)
3. Digestion
4. Ionization
5. Mass Analyzer
6. Ion Detection
7. Spectra Analysis

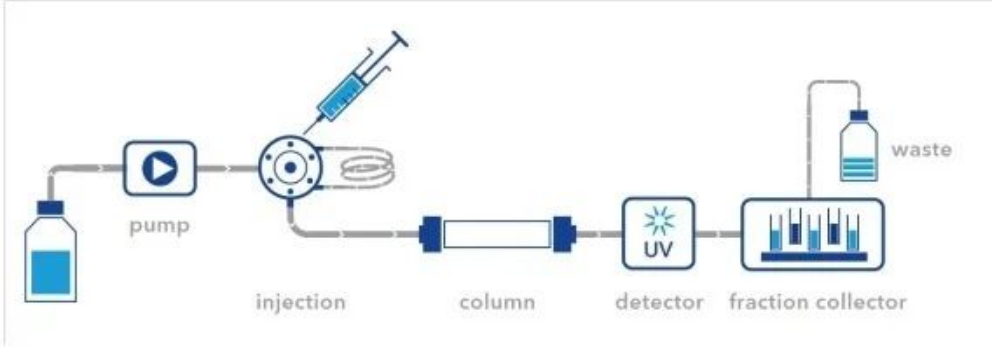
Chromatography

- Method for separating constituent molecules from a mixture
- Separated using physical properties of molecules, e.g. polarity
- Sample dissolved in mobile phase, passed through stationary phase
- Molecules separated by amount of interference passing through column

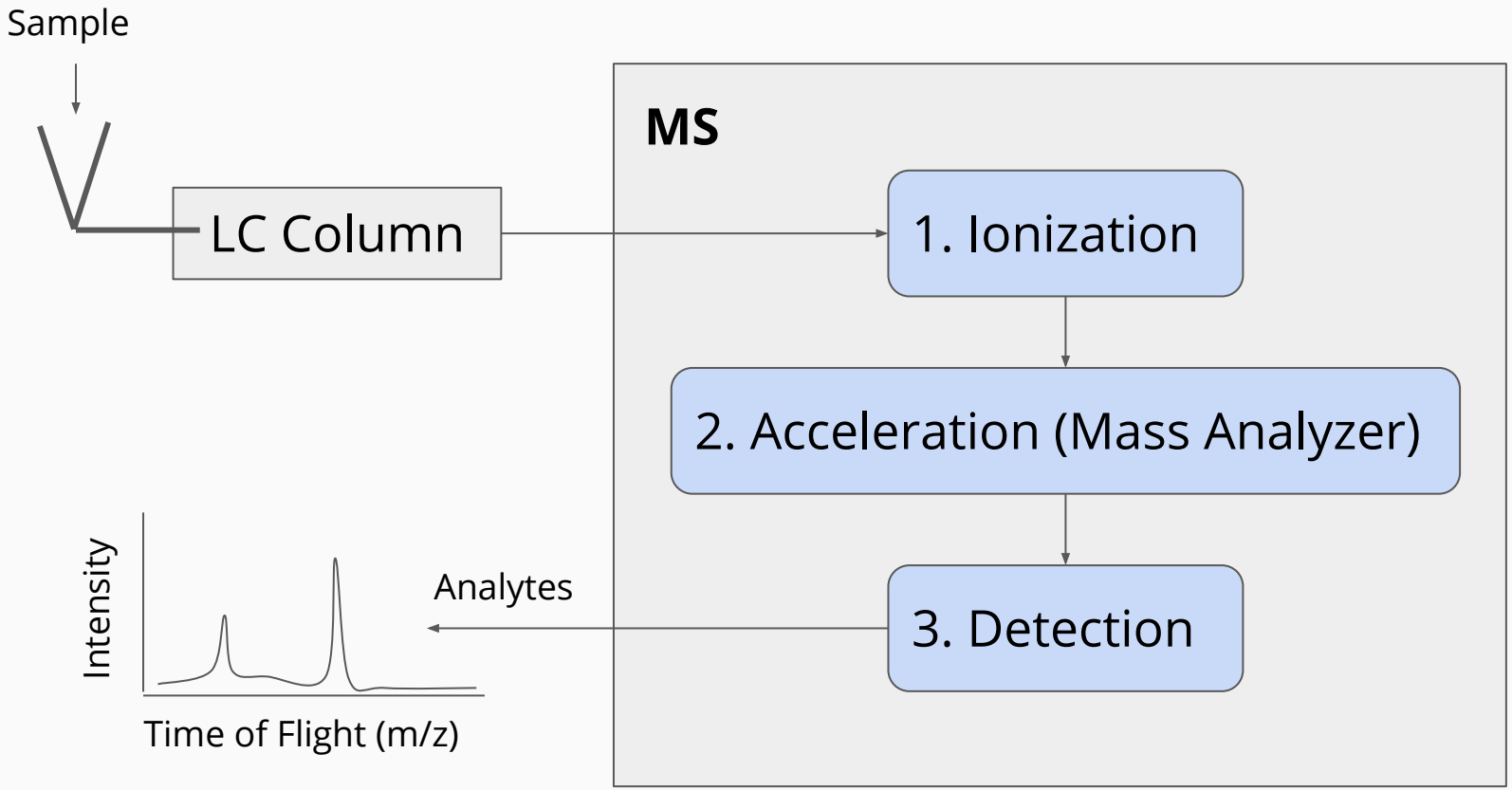
Column Chromatography



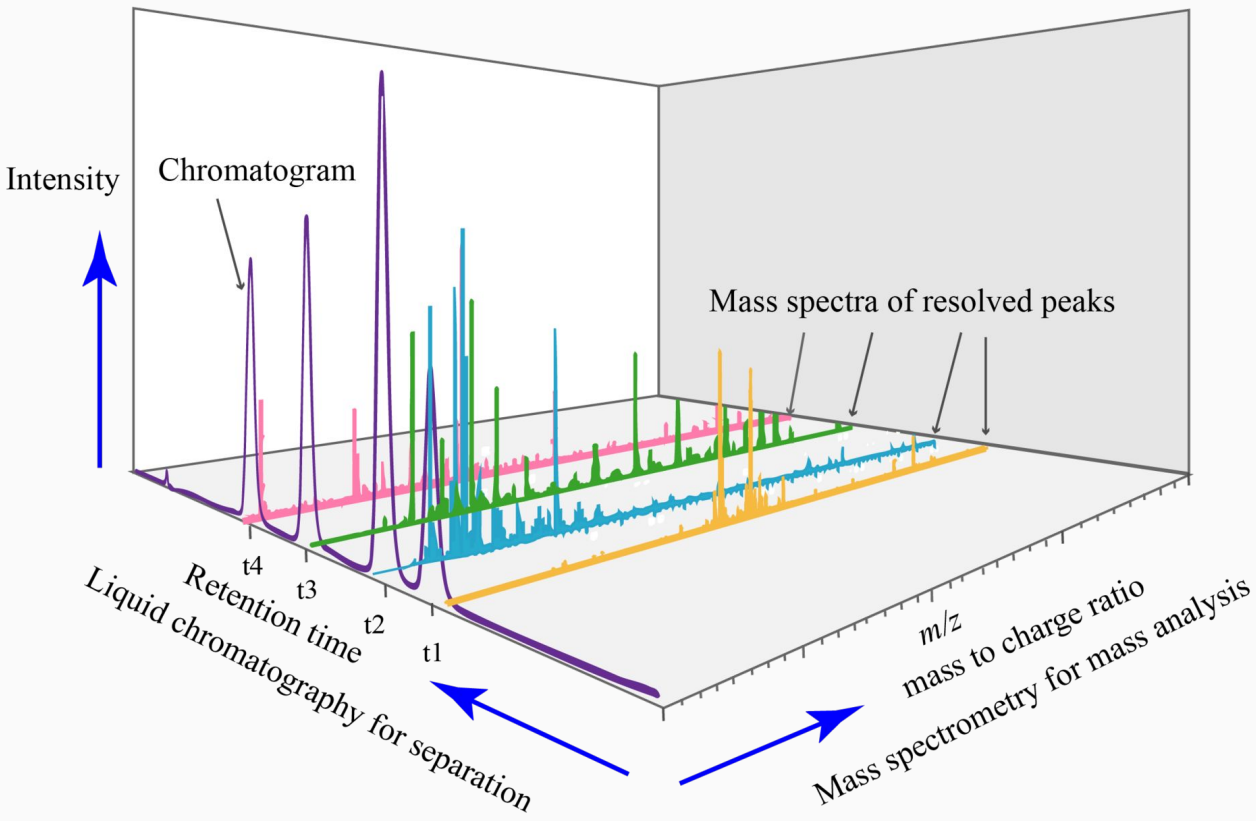
Liquid Chromatography



Liquid Chromatography Mass Spec (LCMS)

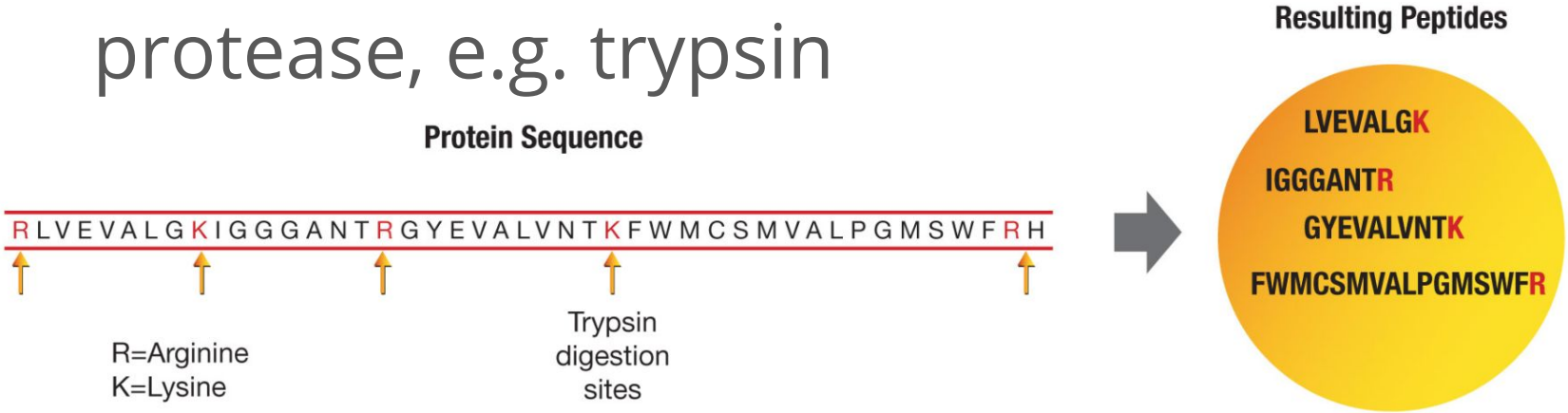


Each LC Fraction Has Its Own Spectra

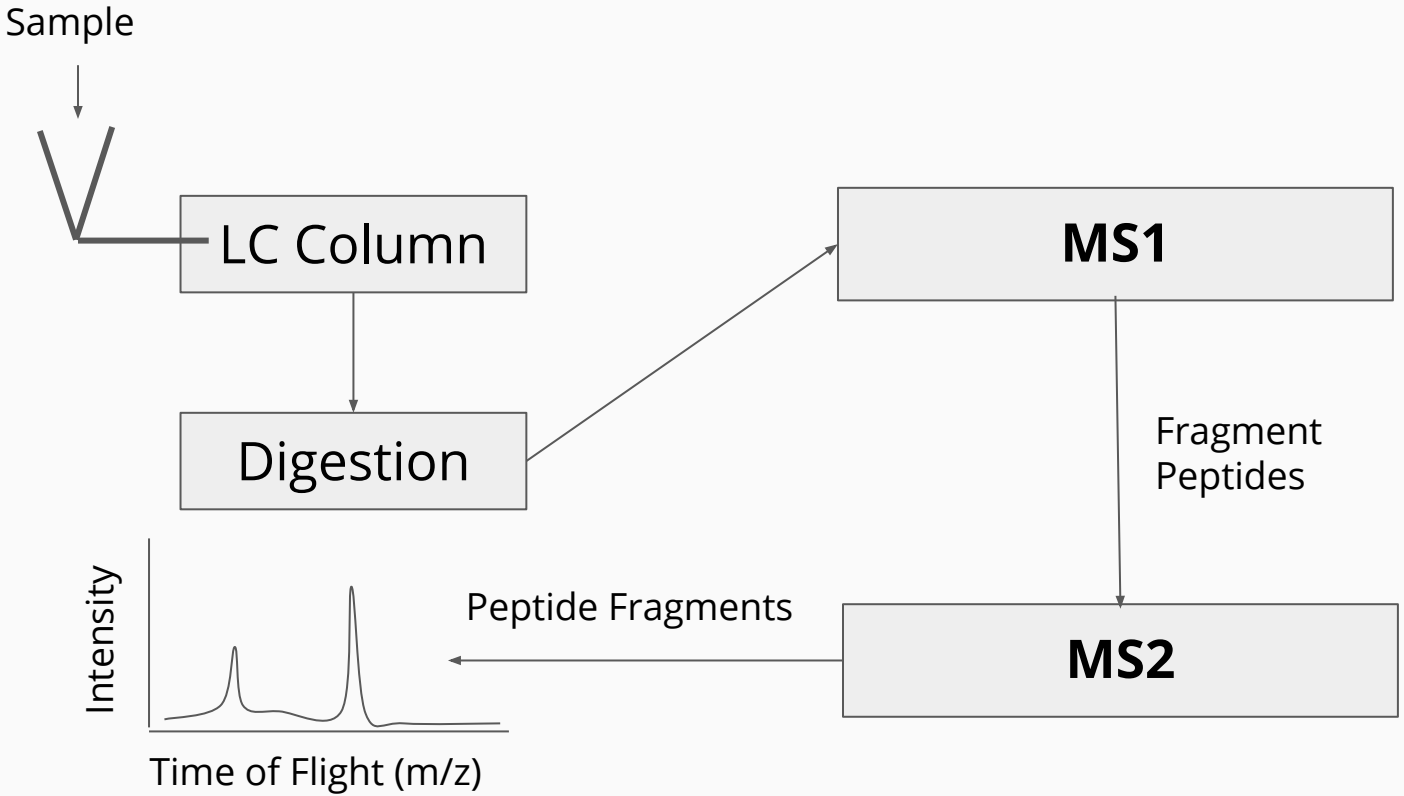


Protein Digestion

- Proteins are macromolecules → big!
- Measuring entire mass of a protein may not be enough to identify it
- Digest into predictable peptides with protease, e.g. trypsin

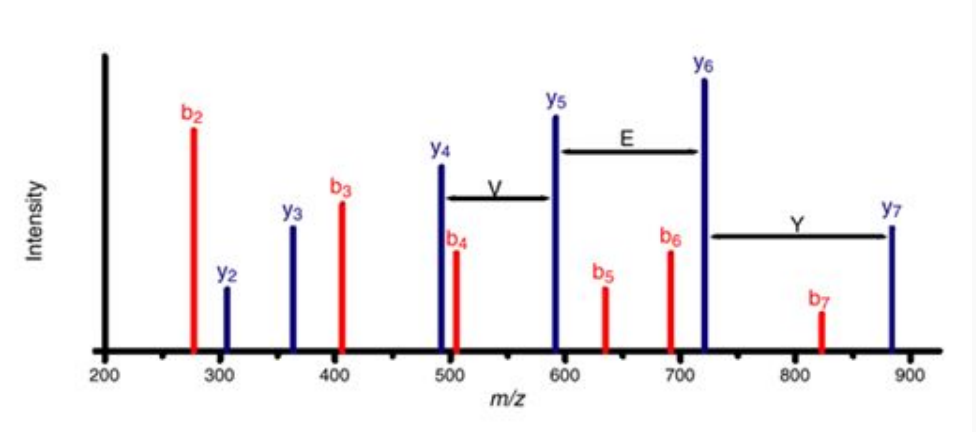


LC/MS for Proteomics



Peptide Identification

1. Peptide fragmented into constituent components
2. Each component has a m/z
3. Sequence of peaks (aka *ladder*) uniquely identifies originating peptide
4. Spectral pattern searched in a database of known spectra to identify peptide



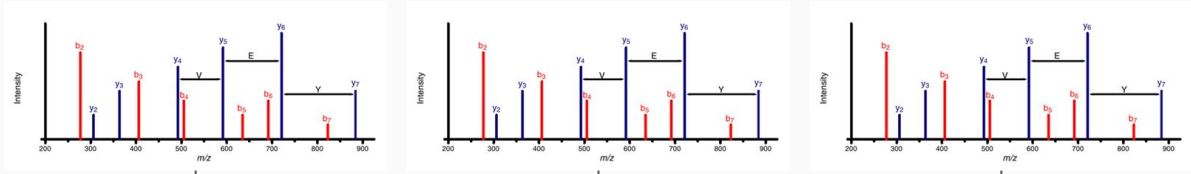
Spectra Databases

Many peptide spectral databases available

- UniProtKB
- Human Proteome Map
- PeptideAtlas
- Post-Translational Modification databases

Protein Identification

Identify compatible peptides using spectral databases



Peptide 1 Peptide 2 Peptide 3

Identify compatible proteins based on identified peptides + digestion method using genomics+protein databases

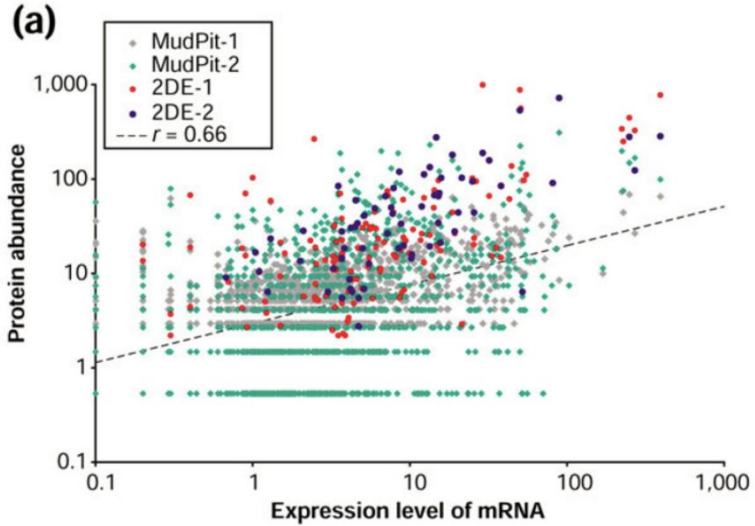
Possible proteins

Quantification

- Above analysis produces a mapping of peaks → peptides → proteins
- Height/AUC of peaks proportional to peptide intensity
- Peptide intensity can be mapped to a protein abundance matrix
- Analyze/interpret much like e.g. RNA-Seq

Caveats & Considerations

RNA vs protein abundance are not well correlated



Obtained from Genome Biol. 2003;4(9):117

Caveats & Considerations

- Proteins can also be post-translationally modified → even more complex!
- Spectral databases are incomplete
- Some protein families are very complex
  - Glycoproteins
- Not all important macromolecules are proteins
  - Lipids, DNA, metabolites, etc