Synopsis: Molecules of interest to biochemists may be classified as **small molecules** and **macromolecules**. Small molecules are similar to those encountered in conventional organic chemistry, and are important in metabolism, which we deal with in the second half of the course. Macromolecules are huge by comparison - molar masses from 10^4 to over 10^9 g.mol^{-1}. What makes it possible to comprehend structures of this magnitude is their **modular construction** from much simpler smaller molecular units. The basis of macromolecule assembly is the **reversible formation** of certain kinds of bonds, e.g. **ester or amide bonds** to link up smaller subunits into long chains.
Proteins are chains of linked **amino acids**. Each amino acid has a unique side chain. Since the α-amino/α-carboxylate core is constant, the **side chain** R determines the specific properties of a particular amino acid and the role it plays in a protein.

REVIEW: CHEM*1040 notes regarding electronegativity and Lewis structures.

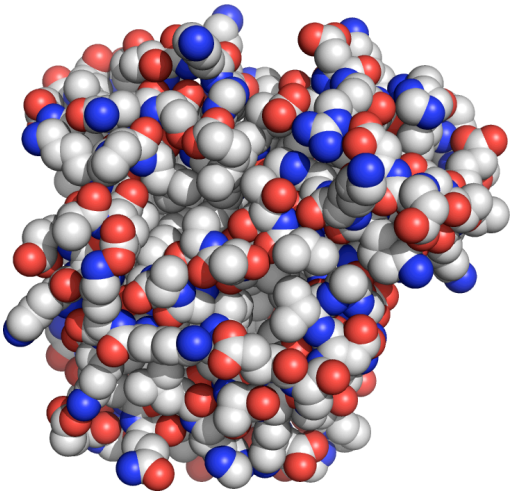
Classes of molecules found in biochemistry:

Small molecules

- **Sugars, amino acids, nucleotides, fatty acids, simple carboxylic acid derivatives**
- Interconversions of small molecules may be used to store or release energy, which is the basis of **metabolism**
- Particular kinds of small molecules may serve as building blocks for **macromolecules**

Macromolecules

- **Proteins,** made as chains of amino acids
- **Nucleic acids,** made as chains of nucleotides
- **Polysaccharides,** made as chains of simple sugars



Proteins form complex structures capable of many functions, including structural components of cells, catalysis of reactions and communication processes. For this reason, the first half of the semester will focus on proteins and their role.

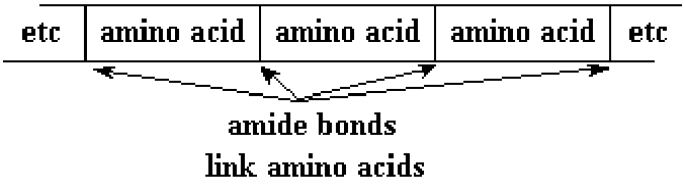
For example**, myoglobin** (right) is a protein that stores O_2 in muscle tissue.

Typical protein molecules have molecular masses between 10 000 and 100 000 g.mol^{-1}, so they contain literally thousands of atoms. Myoglobin has a molecular mass of 16 500 g.mol $^{-1}$.

Because proteins and other macromolecules are so large, biochemists use a unit called the **kiloDalton (kDa).** One Dalton is simply 1 g.mol^{-1}, so 11 000 g.mol^{-1} becomes 11 kDa. Typical proteins are therefore between 10 and 100 kDa, while myoglobin is 16.5 kDa. The largest known single protein molecule is titin at 10 000 kDa.

The building block principle of macromolecule structure

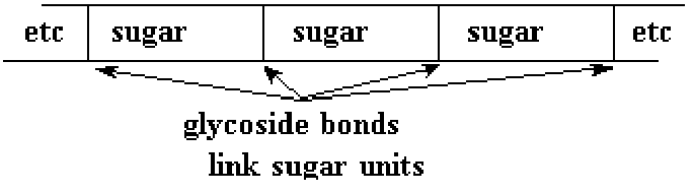
Proteins are chains of linked **amino acids**:
Each protein has a **unique sequence of different amino acids**, and a well-defined **size and structure**. The arrangement of amino acids in the chain determines the properties and function of the protein. A protein of 100 amino acids has a mass of about 11 000 g.mol^{-1}, about 110 g.mol^{-1} per amino acid. Proteins are between 10 and 10 000 kDa (10^4 to 10^7 g.mol^{-1}).



Two other kinds of macromolecule will be dealt with later in the semester:

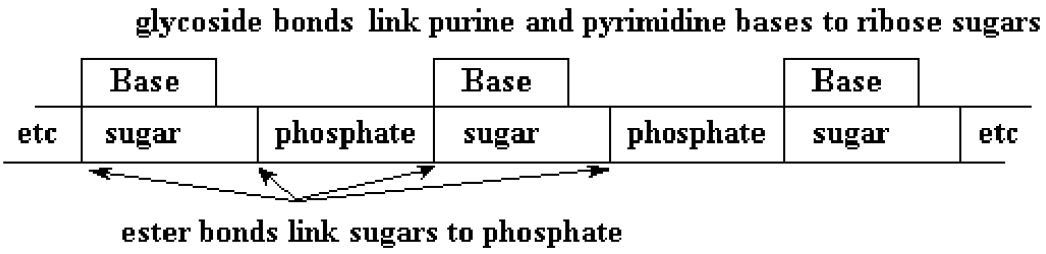
1. A **polysaccharide** is a chain of sugars:

   Most polysaccharides, e.g. starch, are **simple repetitive structures** of one or two sugars, with no definite size. Some polysaccharides are used for storage of sugars; others act in simple structural roles.



2. **Nucleic acids** DNA and RNA are a bit more complex:

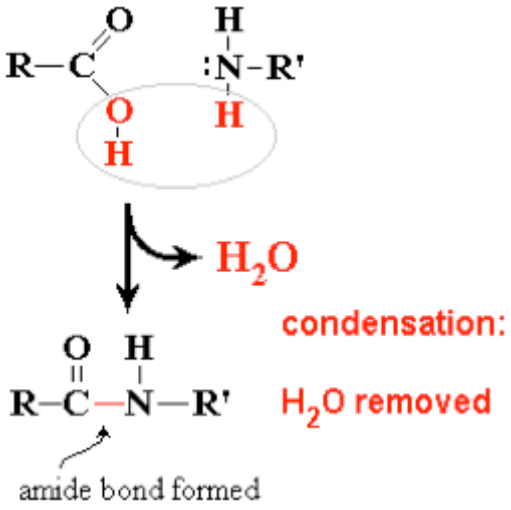
   The backbone is simple and repetitive; but different bases are attached giving nucleic acids unique and characteristic sequences. The repeating unit, base + sugar + phosphate is called a **nucleotide**.



Bonding between subunits

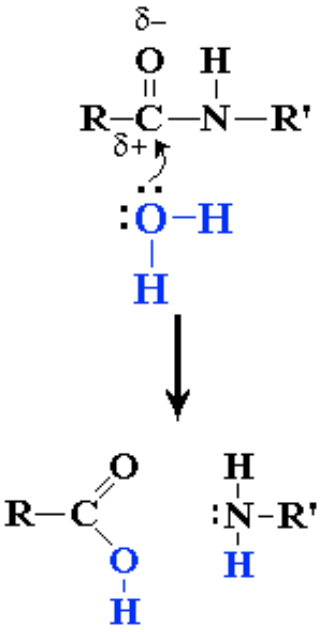
The types of bond that link subunits in macromolecules are formed by a process called **condensation**, since the process involves **elimination of the elements of H_2O**.

e.g **amino acids** contain both carboxylic acid and amino groups, and these allow the formation of an **amide bond** by condensation:



The converse of condensation is the **attack of H_2O on the amide bond**, which restores the original carboxylic acid and amino group and thus **unlinks** the two units. This is called **hydrolysis.**

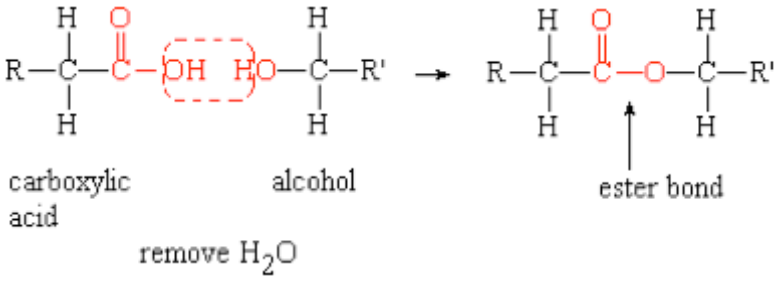
The **carbonyl group C=O** of the amide is the **point of weakness** that allows H_2O to attack**.**



Bonds formed by condensation and broken by hydrolysis:

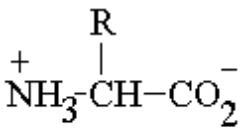
carboxylic acid + amino group ⇔ amide

carboxylic acid + alcohol ⇔ ester

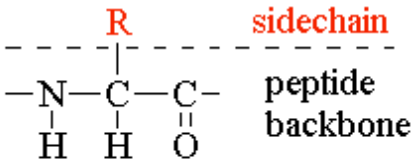


Proteins are chains of amino acids

Twenty different amino acids are found in proteins. They have in common the following structure, distinguished by different side chain groups, shown as R.



as single amino acid



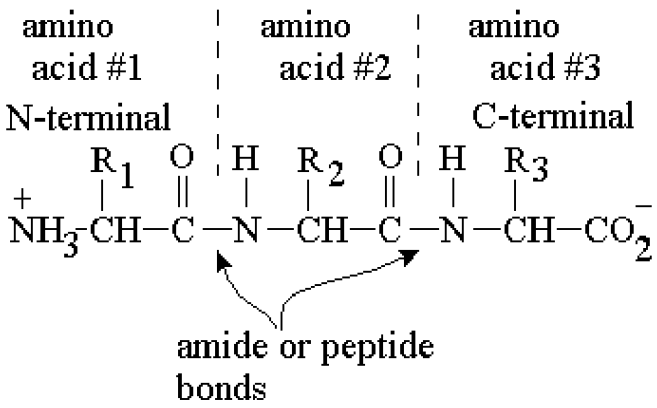
as part of a polypeptide chain

Each amino acid has an **amino group** and a **carboxylate group**, and can thus form either end of an amide bond. (What's different about the amino acid **proline**? - look it up and find out!)

After removal of the H_2O in condensation, the portion of the amino acid found in the chain is called an amino acid **residue.**

Large numbers of amino acids can be strung together in a chain:
Each amino acid in the chain can have a different side chain **R; these give the protein molecule as a whole its unique properties.**

On average, each amino acid contributes about **110 Da** to the molar mass of a protein. 153 amino acids make up the 16.5 kDa myoglobin molecule.



Some terminology to learn:

One end of the chain has an amino acid with an uncombined amino group, usually protonated; this is called the **N-terminal amino acid**. The other end has an amino acid with an uncombined carboxylate group, and is called the **C-terminal amino acid**.

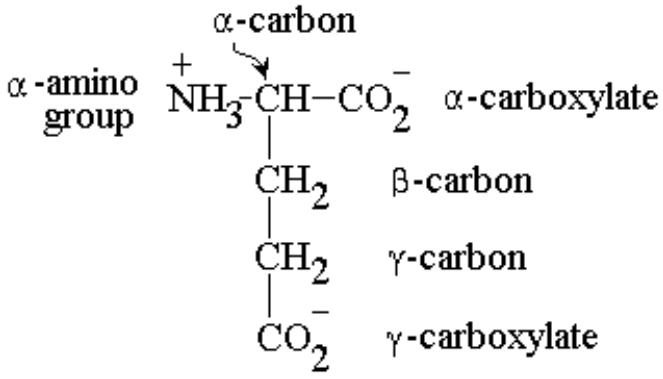
Amino acid chains linked in this way are called peptide **chains; the amide bonds linking them are called peptide bonds.**

A **polypeptide** is a large peptide chain, usually the complete amino acid chain of a protein.
Typical proteins are several hundred amino acids long.
***Poly* = Greek for "many"**

An **oligopeptide** is a smaller peptide chain, often applied to a fragment of a larger protein
***Oligo* = Greek for "a few"**

Structural layout of amino acids

Amino acids in a peptide chain have identical backbone but **unique side chains R.** Specific properties of a protein are determined by the particular functional groups present in the side chains. The amino acid **glutamate** (**Glu** for short) is an example.



In biochemical nomenclature, Greek letters identify the carbon atoms of the structural core of an amino acid, as in the example at left:

Functional groups are labeled according to the core atom to which they are attached, e.g. **α-amino, α-carboxylate**, **γ-carboxylate**. The α-amino and α-carboxylate become linked up in the peptide bonds making up the backbone.

The side chain properties that have most influence on the behaviour of the protein include:

- **polarity**
- **hydrogen bonding ability**
- **charge**

There are twenty different amino acids commonly found in protein chains, described by a full name, a three-letter abbreviation, or a single letter symbol. Full structures of amino acids and other properties can be found in Lehninger p.73-75 and on the amino acids Euchre deck (available online).

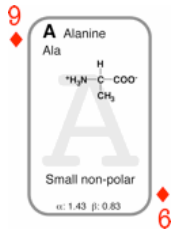
You should be prepared to *reproduce* the structures of the amino acid constituents of proteins and the complete covalent structures of proteins, and to know the single- and three-letter abbreviations for the amino acids.

Classification of amino acids by properties

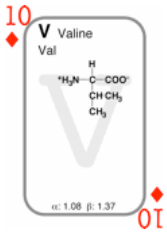
The "**inverted pyramid**" below is a convenient memory aid which groups amino acids according to **common properties and structures**. There is also a Euchre deck of the amino acids and nucleotides in DNA. If you play euchre, try these cards out!

6 amino acids with very non-polar side chains:

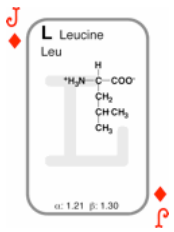
Alanine
Ala, A



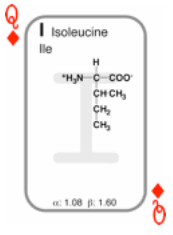
Valine
Val, V



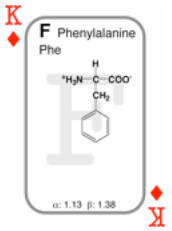
Leucine
Leu, L



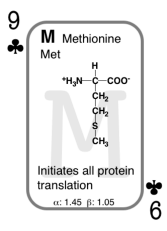
Isoleucine
Ile, I



Phenylalanine
Phe, F

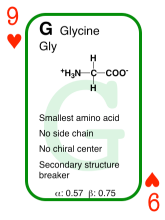


Methionine
Met, M

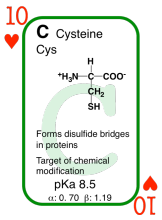


5 amino acids with medium to moderately non-polar side chains:

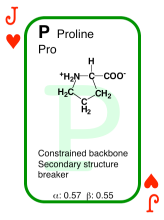
Glycine
Gly, G



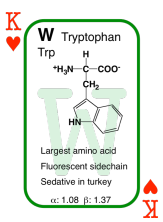
Cysteine
Cys, C



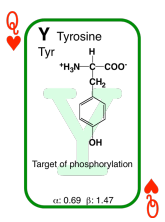
Proline
Pro, P



Tryptophan
Trp, W

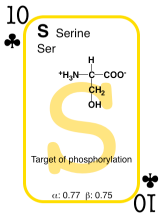


Tyrosine
Tyr, Y

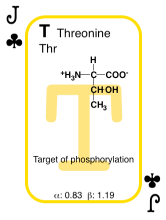


4 amino acids with polar uncharged side chains:

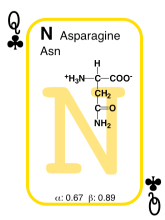
Serine
Ser, S



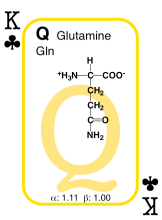
Threonine
Thr, T



Asparagine
Asn, N

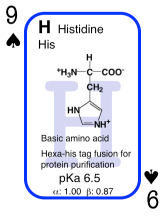


Glutamine
Gln, Q

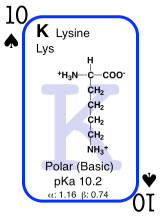


3 amino acids with positively charged side chains:

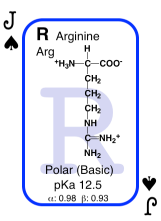
Histidine
His, H



Lysine
Lys, K

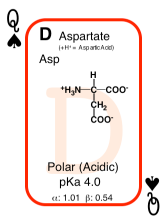


Arginine
Arg, R

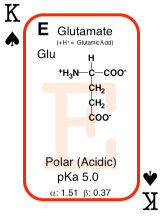


2 amino acids with negatively charged side chains:

Aspartate
Asp, D



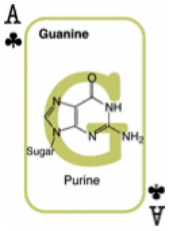
Glutamate
Glu, E



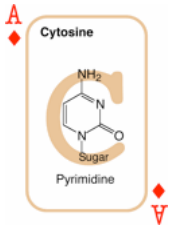
And, just for the record, here are the 4 aces in the euchre deck:

4 nucleotides in DNA

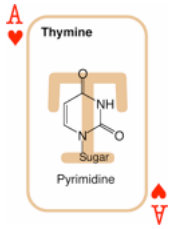
Guanine
G



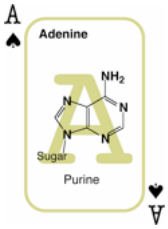
Cytosine
C



Thymine
T



Adenine
A



(You'll get into the nucleotides in the 2nd half of the course)

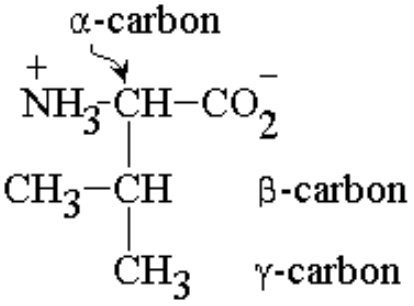
Synopsis: Amino acid side chains may be classed according to polarity, hydrogen bonding ability and ionic properties. Biochemical reactions occur in aqueous solution at close to neutral pH. Many biochemical substances such as amino acids include weak acid groups such as carboxylates, SH or phenolic OH, or weak bases such as amines or some ring N compounds. The behaviour of such groups is highly dependent on whether they are protonated or deprotonated.

REVIEW: CHEM*1040 notes regarding weak acids and bases.

Amino acids with very non-polar side chains: Ala, Val, Leu, Ile, Phe, Met

These side chains are dominated by **hydrocarbon**, i.e. consists only of **C-C** and **C-H** bonds, e.g. **Valine**

Hydrocarbon is **non-polar** and **hydrophobic,** or water avoiding.

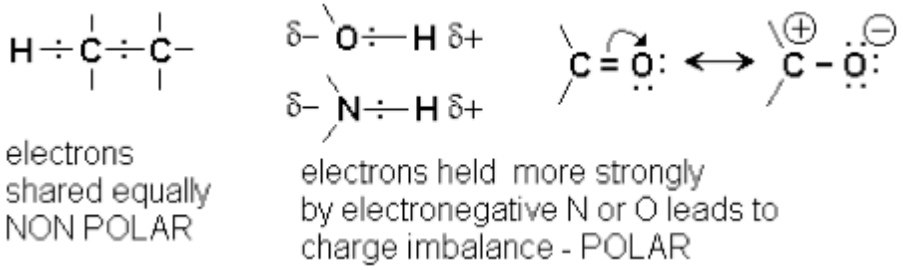


Polar and non-polar properties

Polarity is a property that occurs when atoms in a molecule have very different electronegativity. **Electronegativity** refers to the tendency of a nucleus to hold electrons (it does **NOT** mean possession of negative charge).

very electronegative O > N > S > C ≈ H moderately electronegative

Since C and H have similar electronegativity, bonding electrons are **evenly distributed** in hydrocarbon regions. In contrast, O and N are more electronegative, and in bonds such as O-H or C=O, electrons shift towards the electronegative O atom, creating a **dipole.** We say therefore that **C=O and O-H are polar,** while **C-C and C-H are non-polar.**



Non-polar groups interact well with each other and poorly with polar groups. Hydrocarbon regions of a molecule are also generally chemically **unreactive**.

Hydrocarbon side chains are described as hydrophobic, i.e. they avoid H_2O, which is very polar. Hydrocarbon side chains tend to cluster together, so as to minimize the area of direct contact between hydrocarbon and H_2O, a property known as the **hydrophobic effect**. The hydrophobicity of a side chain is simply related to the number of CH, CH_2 or CH_3 groups. Note: methionine has one S atom in its hydrocarbon chain; however S is much less electronegative than O, so methionine fits in the very non-polar class.

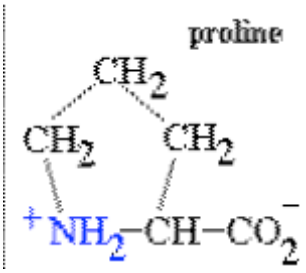
Amino acids with moderately non polar side chains: Gly, Cys, Pro, Trp, Tyr

Glycine has a side chain that is simply **H** linked directly to the α-carbon**: $+NH_3CH_2CO_2^-$**
Although it's a non-polar bond, it's not large enough to make the whole molecule very non-polar.
Note: The α-carbon of all other natural amino acids is **chiral,** with **L- configuration.**

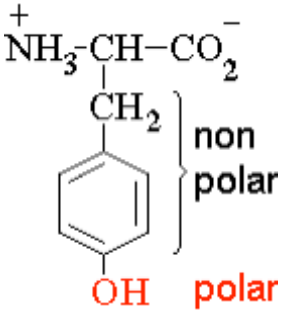
Glycine is the only **non-chiral** structure, since it has **two identical H substituents on the α-carbon.**

Cysteine has the side chain **$-CH_2-SH$,** which is not very polar, because S is much less electronegative than O.

Proline is unique because the side chain links back to the -N, forming a 5-member ring.



Tyrosine and **tryptophan** are the most hydrophobic amino acids, based on their total surface area of CH atoms, however the hydrophobicity is partly offset by the presence of a polar OH group in Tyr shown at right, or slightly polar NH group in Trp, so overall, these two amino acids are only somewhat non-polar.



Amino acids with polar, uncharged side chain: Ser, Thr, Asn, Gln

Ser and Thr both have an **-OH group** on the side chain.

Asn and Gln both have amide side chains **$-CONH_2$** derived from the corresponding carboxylates Asp and Glu.

All four are good **hydrogen bond** formers, with little hydrocarbon; hence they are **polar**.

Hydrogen bonds are electrostatic interactions between a **donor** consisting of the dipole of a polar **O-H or N-H bond** and an **acceptor,** consisting of an **available lone pair of electrons on a nearby N or O atom** (which may be on different molecule).

typical H-bond donors

$R_1N–H$ - - -:OR_2
$R_1O–H$ - - -:NR_2
$R_1O–H$ - - -:OR_2
$R_1N–H$ - - -:NR_2

typical H-bond acceptors

Typical hydrogen bonds (or **H-bonds**) are about 5-10% as strong as a normal covalent bond, and are not permanent bonds like covalent bonds. Instead they result in temporary attractive forces that help hold molecules together. Water molecules are excellent H-bond donors **and** acceptors; so polar amino acids interact well with H_2O. H-bonding attracts H_2O molecules to each other, and this makes water a liquid rather than a gas like methane, CH_4, whose molecules do not form H-bonds.

Biochemistry is concerned with **how molecules function in living cells**, which is an **aqueous environment**. Hence it is important to understand how various biochemical molecules will interact with H_2O.

Positively charged amino acids: Arg, Lys, His

These side chains are weak bases, fully protonated (Lys, Arg) or partly protonated (His) in normal biological conditions, pH 7.0-7.4.
Although the side chain has a hydrocarbon segment, the positive charge dominates over any hydrophobic effect. Charged amino acids are **very polar**.

e.g. **Lysine** side chain is **$-CH_2-CH_2-CH_2-CH_2-NH_3^+$**

Negatively charged amino acids: Asp, Glu

These side chains are carboxylate groups, normally deprotonated at pH 7, and very polar.

Aspartate side chain is **$-CH_2-COO^-$**

Glutamate side chain is **$-CH_2-CH_2-COO^-$**

Two amino acids with oppositely charged side chains can strongly attract each other by electrostatic interactions known as **salt bridges** or **ion pairs** (*see Lecture 7*). They also form strong H-bonds with uncharged H-bond donors or acceptors including H_2O.

Amino acids as weak electrolytes

Normal biochemical processes occur in aqueous solution close to neutral pH; typical physiological pH is about 7.2 to 7.4, and pH 7.0 is a close approximation. Certain functional groups found in biological molecules, in particular carboxylic acids or amino groups, can **gain or lose H^+** depending on the availability of hydrogen ions (or protons) in the solution.

pH expresses the availability of H^+; $pH = -\log_{10}[H^+]$

Each ionic functional group, e.g. amino groups or carboxylic acid groups, has a characteristic constant, **pKa**, which **expresses the tendency to gain or lose H^+**.

The Henderson Hasselbalch equation relates pH to pKa and the state of ionization:

$$pH = pK_a + \log_{10}\left\{\frac{[A^-]}{[HA]}\right\} \qquad \boxed{\frac{\text{deprotonated}}{\text{protonated}}}$$

The Henderson-Hasselbalch equation allows one to do the calculations needed:

1. to determine the pH given the ionic conditions of the surroundings; if pK_a and the concentrations are known, pH can be calculated.
2. to determine the **degree of protonation or deprotonation** of an ionizable functional group at a given pH. If the pH and pK_a are known, the **ratio of concentrations** can be calculated, and this means we can work out what the state of an "ionic" functional group actually is at a given pH.

For each amino acid, there is

- **an α-carboxylic acid (typical pK_a 2.4 ± 0.5)**
- **an α-amino group (typical pK_a 9.6 ± 0.5)**
- **certain amino acids also have a side chain which *may* be charged**
  - **Not all amino acids have a side chain that bears a charge!!**

Exact pK_a values for each of the 20 amino acids can be found in Lehninger p. 73

The Henderson-Hasselbalch equation is used to calculate the state of each group at pH 7.0

$$pH = pK_a + \log_{10}\left\{\frac{[A^-]}{[HA]}\right\}$$

$$10^{pH-pK_a} = \frac{[A^-]}{[HA]}$$

Given **pK_a = 2.4** for the **α-carboxylic acid group**, we can calculate the ratio of anionic carboxylate -COO^- to neutral carboxylic acid -COOH when **pH = 7.0**:

$$\frac{[A^-]}{[HA]} = 10^{7.0-2.4} = 40000$$

This indicates that the vast majority of α-carboxylic acid groups are fully deprotonated (i.e. have lost H^+) and **exist in the carboxylate ion state at pH 7.0.**

Given **pK_a = 9.6** for the **α-amino group,**

$$\frac{[\text{-}NH_2]}{[\text{-}NH_3^+]} = 10^{7.0-9.6} = 0.0025$$

This indicates that the α-amino group is essentially fully protonated (has gained H^+) and **exists in the $–NH_3^+$ state at pH 7.0.**

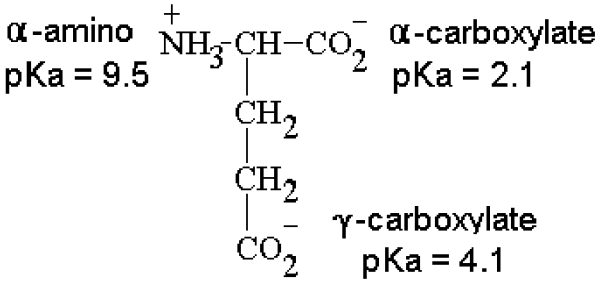
The backbone portion of a free amino acid at pH 7 is therefore best represented as

$^+NH_3$-CHR-COO^-

When an amino acid is linked up as part of a peptide chain, the situation is different. The α-amino groups and α-carboxylate groups combine to form amide or peptide bonds. **When combined as an amide, the α-amino groups and α-carboxylate groups are not free to protonate or deprotonate**, so in the peptide bonded state, the amino acid backbone is uncharged**: x**-NH-CHR-CO-**NH-CHR-CO-**NH-CHR-CO-**x**

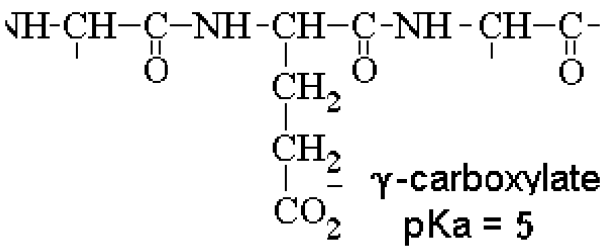
Meaning of pK_a

The value of pK_a tells you **where in the pH scale** a functional group undergoes protonation or deprotonation, **e.g. for glutamate:**

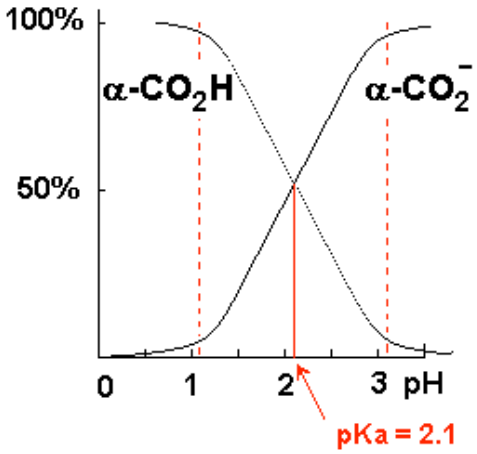


Each group has its **own pK_a value.** The exact value of a group's pK_a depends on its **chemical context.** The presence of the **positive NH_3^+** near the α-carboxylic acid position favours deprotonation to the negative **carboxylate,** hence this group is **more acidic (lower pKa)** than the **γ-carboxylate.**

When glutamate is part of a peptide chain, the α-amino group is part of a neutral amide bond, so in the absence of a nearby positive charge, the glutamate side chain carboxylate has $pK_a = 5$.

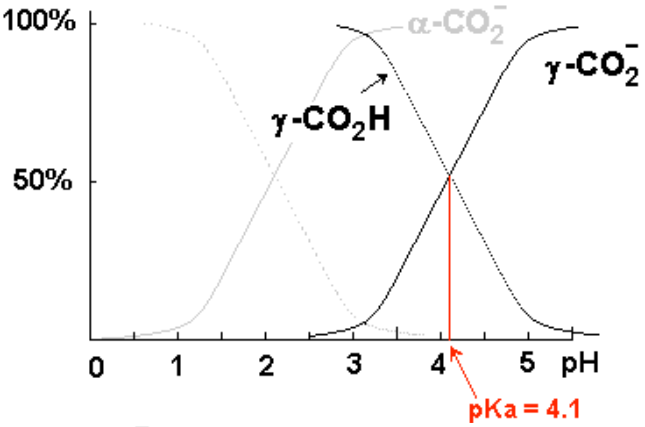


If we start with a sample of free glutamic acid initially at very low pH, all functional groups will be protonated. As pH is increased, each functional group will start to lose H^+ **when the pH approaches the pK_a of that group**. For example, deprotonation of the α-carboxylic acid group occurs around pH 2.1. When pH = 2.1, the α-carboxylic acid will be exactly 50% deprotonated.

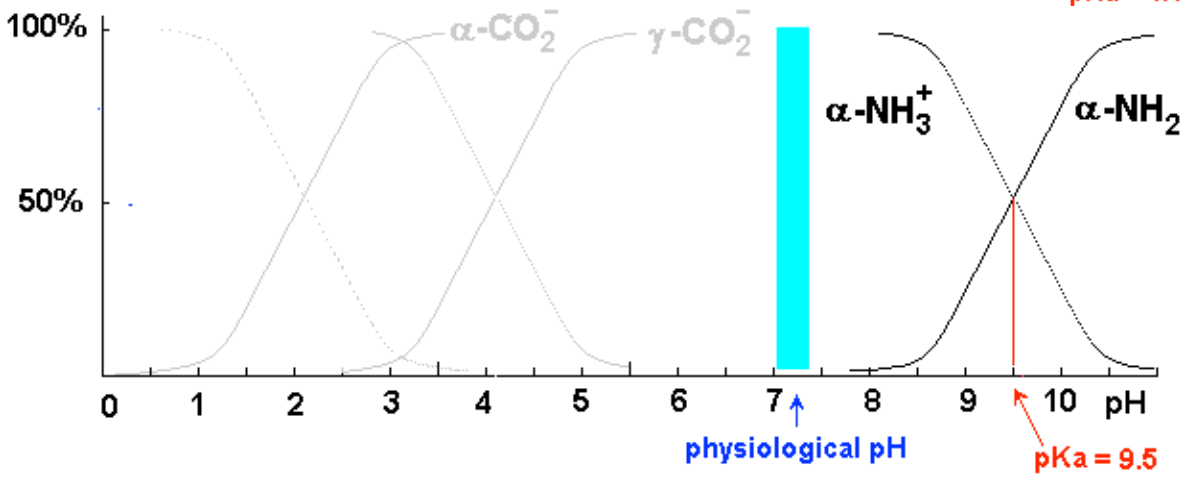


Above pH 3.1, the α-carboxylate group is essentially fully deprotonated.

If the pH continues to increase, as it approaches the pK_a of the γ-carboxylic acid group, this will deprotonate in turn to yield the γ-carboxylate.



Ultimately, if pH is increased much above pH 9, the α-amino group will start to be deprotonated.



Note that the transition of a given group from **fully protonated** to **fully deprotonated** occurs over a narrow range of pH, **essentially over a range of 1 pH unit on either side of pK_a**. This gives us a simple set of rules for determining the ionic state of the functional groups in a molecule at any given pH:

If **pH is one unit or more above its pK_a**, a group may be considered fully **deprotonated,**
e.g. carboxylate groups with pK_a = 2.4 exist as $-COO^-$ at pH 7.

If **pH is equal to pK_a,** the group is exactly 50% protonated and 50% deprotonated**.**

If **pH is one unit or more below its pK_a,** a group **may be considered fully protonated,**
e.g. amino groups with pK_a = 9.6 exist as $^+NH_3$- at pH 7.

Hence it is easy to determine by inspection that glutamate exists with its two carboxylate groups **deprotonated,** and its amino group **protonated** at physiological pH.

There are seven amino acids with side chains that undergo deprotonation

	pK_a	State at pH 7	State at low pH	State at high pH
Aspartate	**4.0**	**$-COO^-$**	**-COOH below pH 4**	**same as pH 7**
Glutamate	**5.0**	**$-COO^-$**	**-COOH below pH 5**	**same as pH 7**
Histidine	**6.5**	**76% ring N:**	**ring NH^+ below pH 6.5**	**ring N:**
Cysteine	**8.5**	**Neutral -SH**	**same as pH 7**	**$-S^-$ above pH 8.5**
Tyrosine	**10.0**	**Phenol-OH**	**same as pH 7**	**Phenolate-O^- above pH 10**
Lysine	**10.2**	**$-NH_3^+$**	**same as pH 7**	**$-NH_2$ above pH 10.2**
Arginine	**12.5**	**$-NH_3^+$**	**same as pH 7**	**=NH above pH 12.5**

pK_a values given are for the side chain of the amino acid **in a polypeptide**. Values are slightly different for the free amino acid (see Lehninger p. 73).

Cysteine and Tyrosine were not included as negative in the pyramid table of amino acids, because they are **neutral** at pH 7.

Important note: these rules can tell you whether a group is **protonated or deprotonated**, but not immediately whether the group is positively or negatively charged. To determine this, it is necessary to apply a little knowledge of the chemistry of the group:

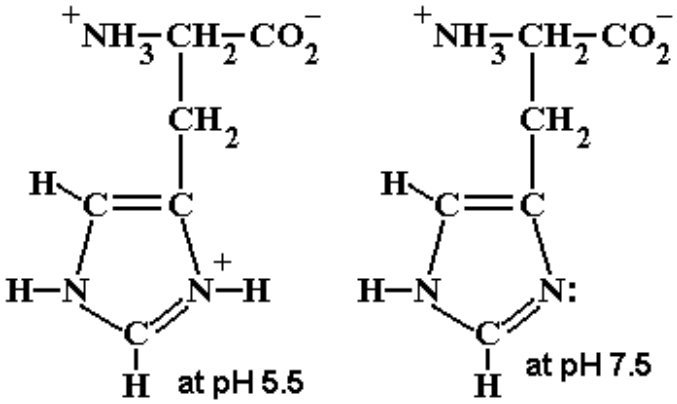
Groups that ionize on **O or S** are **neutral when protonated** and **negative when deprotonated.**
Groups that ionize on **N** are **positive when protonated** and **neutral when deprotonated.**
No group can go from positive to negative in a single deprotonation step.

Many functional groups contain O, N or S but do **NOT** undergo ionization, e.g. alcohols or amides. Simple alcohols **(R-OH)** such as the side chains of **serine or threonine**, and the amides **(R-CONH$_2$)** such as the side chains of **asparagine or glutamine**, do not deprotonate or protonate *appreciably* in aqueous solution and their side chains are neither weak acids nor weak bases.

Partial ionization

If pH is less than one unit greater or less than the pK$_a$, we can't say that a given group is fully protonated or deprotonated. It is still possible to say that the **major form** is deprotonated if pH is above the pKa, and protonated if pH is below the pK$_a$, but the minor form is still present in significant amounts. If we want to be more exact, we can use the Henderson-Hasselbalch equation to determine the exact state of deprotonation.

The **histidine** side chain has **pK$_a$ = 6.5.** At pH 7, the major form is deprotonated, but the pH is not a full unit higher, so it is nowhere near being "fully deprotonated", and the positively charged form is still present in significant amounts.



What is the **degree of deprotonation** of the side chain of **His (pK$_a$ = 6.5) at pH 7.0?**

Degree of deprotonation is the fraction of the total that is in the deprotonated state.

$$\alpha = \frac{[\text{deprotonated}]}{[\text{total}]} = \frac{[\text{deprotonated}]}{[\text{protonated} + \text{deprotonated}]}$$

The Henderson-Hasselbalch equation provides the ratio of deprotonated to protonated His :

$$pH = pK_a + \log_{10}\left\{\frac{[His]}{[HisH^+]}\right\} \qquad \frac{\text{deprotonated}}{\text{protonated}}$$

$$10^{7.0-6.5} = \frac{[His]}{[HisH^+]} = 3.2$$

$$\alpha = \frac{[His]}{[HisH^+] + [His]} \qquad \text{divide top and bottom by } [HisH^+]$$

$$\alpha = \frac{\frac{[His]}{[HisH^+]}}{1 + \frac{[His]}{[HisH^+]}} = \frac{3.2}{1 + 3.2} = 0.76$$

The histidine side chain is 76% deprotonated at pH = 7.0

Relating charge to degree of deprotonation

The degree of deprotonation α tells you what fraction is deprotonated. You need to know something about the **chemistry of the functional group** to deduce the charge, and since histidine has a N-base in its side chain, the protonated form is positive, and the deprotonated form is neutral.

If 76% is deprotonated, (charge 0) then 24% must be protonated (charge +1).

Now sum the contributions to average charge:

(0.76 × 0) + (0.24 × +1) = +0.24

fraction neutral fraction positive **average charge**

At pH 7.0, Histidine behaves as if it has a side chain charge of +0.24.

Although no one molecule can carry a fractional charge, the +ve $HisH^+$ and neutral His molecules **exchange H^+ very rapidly** (millions of times per second. Averaged over a period of time, each molecule behaves as if it has +0.24 charge.

A calculation similar to the above gives degree of deprotonation = 0.074 for cysteine (pK_a = 8.5) at pH 7.4. For cysteine, the side chain states are:

-SH ⟷ $-S^-$ + H^+

If 7.4 % is deprotonated single **negative $-S^-$,** then 92.6% is protonated **neutral –SH**

(0.074 × –1) + (0.926 × 0) = – 0.074

fraction negative fraction neutral **average charge**

Effective charge of cysteine side chain at pH 7.4 = – 0.074. This is low enough that it is usually ignored.

The following section **will not be covered in lecture**, but is provided to refresh your memory of buffers from first year Chemistry.

Buffers

A buffer is a substance present in solution in sufficiently high concentration to control the pH of its environment. This occurs by creating a mixture of the protonated (or weak acid) and deprotonated (or weak base) forms of the buffer such that the mixture is at equilibrium with the desired concentration $[H^+]$. The resulting pH will be close to the pK_a of the buffer. Buffering occurs in living cells due to the presence of **bicarbonate ions, phosphate ion and phosphate esters,** all of which have pK_a values close to 7**.**

In biochemistry labs, two buffers are widely used, chosen because their pK_a is close to physiological pH 7-7.4:

dihydrogen phosphate $H_2PO_4^-$ in equilibrium with HPO_4^{2-} (pK_a = 6.8).

trishydroxymethylaminomethane, or **Tris** for short**, in equilibrium with its protonated form $TrisH^+$ (pK_a = 8.08)**

Synopsis: Amino acid analysis usually involves two distinct processes, first **separation** of the individual amino acids from each other and from other contaminants, and then **detection** of the separated components.
Separation is based on the different properties of the side chains, such as polarity or charge. Separation is generally achieved by some form of **chromatography**, such as ion exchange, metal affinity, or gel filtration chromatography.
Detection is based on chemical reactions that generate coloured or fluorescent amino acid derivatives that can be seen and measured.

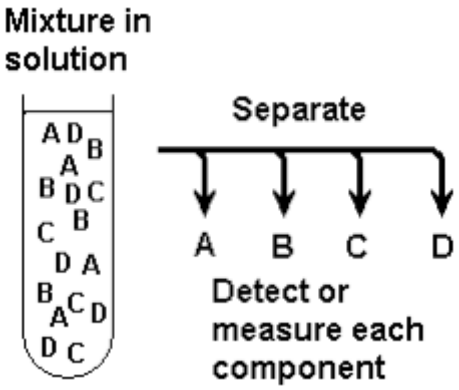
Amino acid analysis

Amino acid analysis is a necessary aspect of experiments to determine protein structure

Analysis involves two processes:

1. The mixture must be **separated** into individual components
2. The components of interest must be **detected**

- Detection can be **qualitative** and determine **what is present**
- Detection can be **quantitative** and measure how much is present.



Chromatography separates components of a mixture

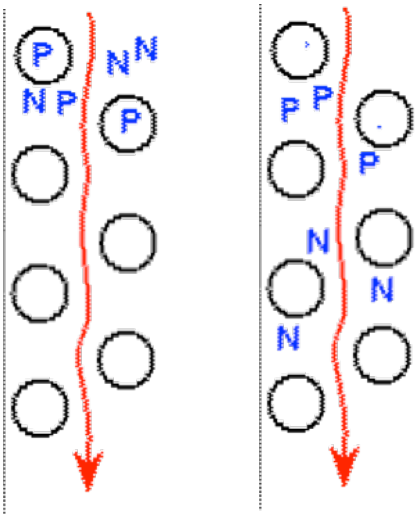
Particles of solid are chosen with a given property.

For example, silica gel contains HO-Si-OH groups that are effective in forming hydrogen bonds with polar amino acids. This makes up the **stationary phase**

Liquid solvent or buffer flows past the particles and is nonpolar. This makes up the **mobile phase.**

Amino acids rapidly exchange between phases.

Polar amino acids (**P**) spend more of their time hydrogen bonded to the stationary silica - they move more slowly Nonpolar amino acids (**N**) spend more time in the moving solvent, and move almost as fast as solvent.

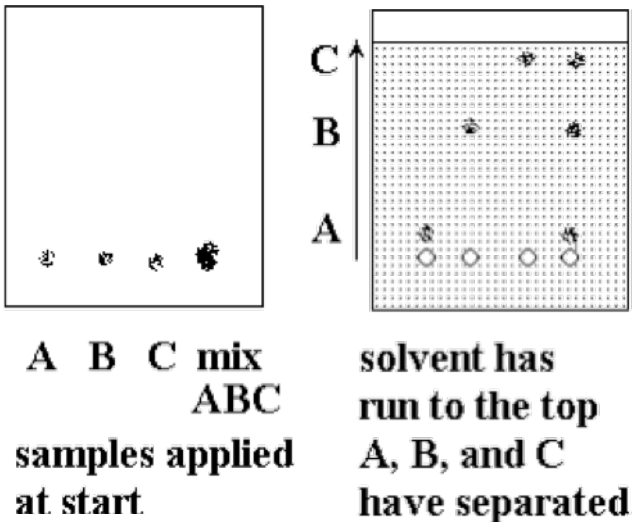


Thin layer chromatography

The silica gel is spread as a thin layer on a plastic or glass sheet.

Samples are applied to the silica gel in a small drop of solvent. Each sample forms a spot, and different sample spots are arranged in a row near the bottom edge of the sheet.

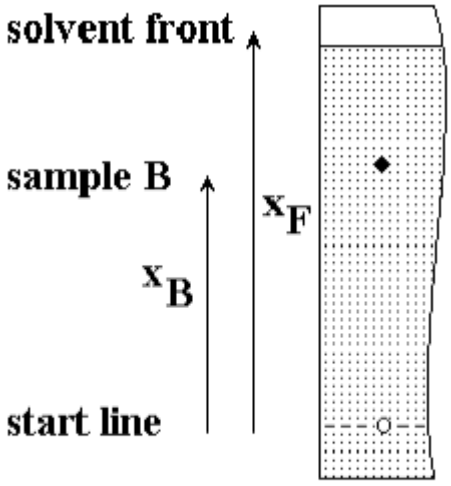
The lower edge of the sheet is dipped in solvent. As solvent soaks up the sheet, the sample spots shift as the solvent moves past.



Different substances move at different rates, so the components of an initial mixture are separated. Pure samples of substances suspected to be in the mixture are also applied. Spots in the mixture can be identified if they move the same distance as one of the pure samples.

Polarity is the basis for separation of substances by thin layer chromatography.

The rate at which a given sample, e.g. an amino acid, moves up the sheet depends on its **relative preference for stationary phase (silica gel) or mobile phase (nonpolar solvent)**. A very polar amino acid such as aspartate will spend most of its time stuck to the silica gel and will barely move. A very non-polar amino acid such as leucine will spend most of its time in the solvent, and will move up the sheet almost as fast as the solvent. Amino acids with intermediate polarity will be in equilibrium between the two phases, and will move part way up the sheet.



Relative mobility, R_F

The highest point the solvent reaches is called the solvent **front**. The ratio of distance moved by a sample and by the solvent **f**ront is called **r**elative mobility, R_F

Relative mobility $R_F = \frac{x_B}{x_F}$

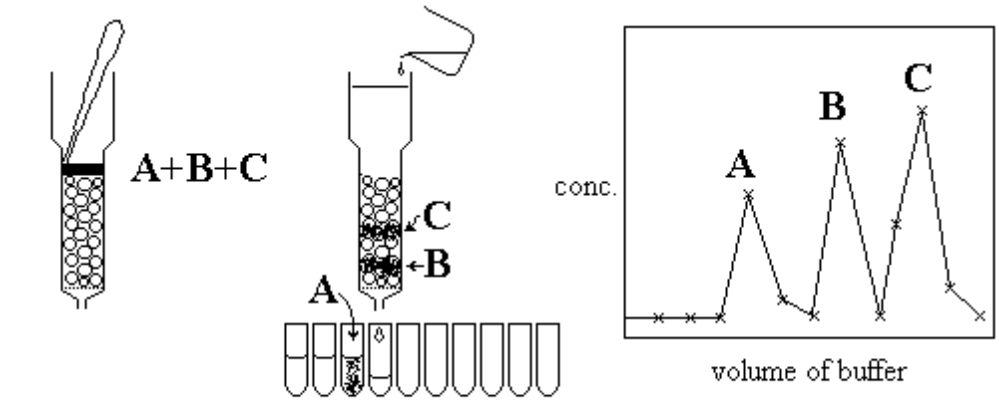
Very polar solutes will have R_F close to zero; Very non polar solutes will have R_F close to 1.0. Most substances will be spread out in between these two extremes.

The value of R_F will depend on the solvent used and solvent must be carefully chosen for a given mixture of compounds.

Other formats for chromatography

Column chromatography – Protein biochemistry work-horse

A granular solid such as silica gel is packed into a glass tube or **column**; silica is usually held in place by a porous disk at the bottom. A sample mixture is applied at the top, and then solvent or buffer solution is allowed to flow through. Sample solutes travel with the flow of buffer solution to the bottom of the column. Substances that bind more strongly to the solid phase require more buffer to pass through or be *eluted* from the column. The main advantage is that the **separated components of the mixture can be collected** allowing additional experiments to be performed on the individual components.



1) Sample mix ABC applied at top

2) Buffer added, carries sample mix through column to collection tubes.

3) Contents of collection tubes analysed, results plotted.

Collection tubes are changed after a fixed volume of buffer or solvent has passed through, whether sample has come through or not. When enough tubes have been collected, their contents are detected, and the quantity of sample in each tube is graphed as a function of volume buffer that has passed since the start. The volume of buffer needed to move a given sample through the column is called its **elution volume**.

High performance liquid chromatography (HPLC)
Column chromatography using specially designed columns and with solvent pumped through for greater efficiency. This is the usual method in research labs.

Detection of separated amino acids

All 20 amino acids are colourless substances, and quantities in an analysis may be anything from 10^{-6} to 10^{-10} moles; not enough to see, let alone weigh out.

The separated amino acids must be **detected** by special means, which involve reaction with a **colour** or **fluorescence-generating reagent.**

Ninhydrin: reacts with -amino N to a give purple colour (10^{-8} mol detectable).

Fluorescamine: reacts with -amino N to give yellow fluorescence (10^{-10} mol detectable). When illuminated with UV lamp, the sample emits a yellow glow.

Since fingerprint sweat contains significant traces of amino acids, ninhydrin and fluorescamine are both used by police investigators to detect otherwise invisible **fingerprints**.

The reagents are either sprayed onto the TLC plates, or added to the solvent as it emerges from the column. The **intensity** of colour or fluorescence is recorded and plotted as a graph. **Colour intensity is proportional to the number of moles of each amino acid.**

An alternative method often used in conjunction with HPLC is to **prelabel** the sample compound with a coloured or fluorescent dye **before** separation, and record the colour intensity as each amino acid emerges from the column. This method is often preferred for quantitative analysis, since the dye reaction can be allowed to go to completion ahead of time (*see Lecture 5*).

Dyes used for prelabelling include **fluorodinitrobenzene**, **dansyl chloride, dabsyl chloride, phenylisothiocyanate**, Lehninger Fig. 3-25 p.94.

Ninhydrin and fluorescamine can't be used to label amino acids before separation since the colour-forming reaction destroys the amino acid.

Different mechanisms of chromatography

Reversed phase chromatography:

Instead of polar silica gel, a **non-polar hydrocarbon silicon derivative** is used as the solid stationary phase; instead of non-polar solvent, **polar** solvent is used as mobile phase. The order of passage is reversed, since now **polar solutes don't bind** and have high R_F, while **non-polar solutes do bind** and have low R_F. (Used because it's better at distinguishing subtle differences in hydrocarbon side chains of amino acids.)

Ion exchange chromatography

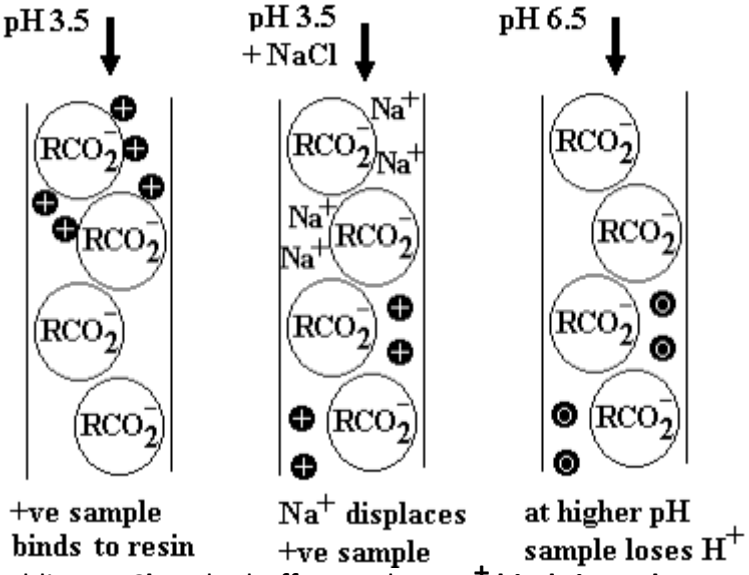
The silica gel stationary phase is replaced by **ionic resins**:

Cation exchange resins are based on polymers with **negative** carboxylate groups, and will bind positive ions or cations.

Anion exchange resins are made from polymers with **positive** amino groups, and will bind negative ions or anions.

Solutes will now bind according to their **charge** rather than polarity, e.g. **positive amino acids NH_3^+-CHR-CO_2H bind to negative charged resin.**

All amino acids can be made positive to some degree by **lowering the pH, e.g to pH 3.5**. The exact charge on a given amino acid will depend on its exact pKa value, and there is enough difference that all 20 amino acids are easily distinguished.

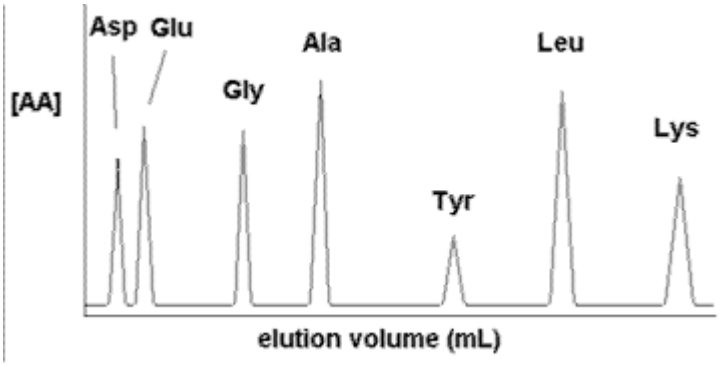


The amino acids can be eluted either by adding **NaCl** to the buffer, so that **Na^+ binds in exchange** for the +ve amino acid **NH_3^+-CHR-CO_2H**. More weakly bound amino acids are displaced by a low NaCl concentration, while tightly bound amino acids require a higher concentration of NaCl to be eluted.

Alternatively, pH may be increased. As pH rises, amino Amino acids become deprotonated giving a net neutral charge, **NH_3^+-CHR-CO_2^-**, and their binding to the negative resin is weakened.

Amino acids are detected and concentration measured as they come out of the column.

The volume of buffer needed to move a given amino acid from top to bottom of the column is also measured: this is the **elution volume** for that amino acid. Elution volumes may be normalized by comparing with elution volumes of a common standard such as Ala or Leu.



Elution volumes are characteristic for a given amino acid and this allows amino acids to be identified.

Separation of proteins from complex mixtures

Protein samples can be extremely complex since proteins are generally extracted from cellular sources. Bacteria such as ***Escherichia coli*** or yeasts such as ***Saccharomyces cerevisiae*** are often grown as sources of proteins and enzymes. Other sources include extracts of tissues such as animal liver. These extracts may contain about 1000 to 3000 different proteins, and three problems must be addressed.

- A typical single protein may represent only 0.03 to 0.1% by mass of the protein mixture; this can be increased by inducing **over-expression** of a gene inserted into yeast or bacteria.
- There may be several other proteins present in the extract with similar properties to the one you are trying to isolate.
- Proteins are easily damaged under harsh conditions such as extreme pH, non-aqueous solvents and temperature, and separation techniques must be carried out at 0-4°C and near neutral pH to minimize loss of sample.

Complete protein purification involves successive application of several chromatographic or other separation techniques. Since there may be other proteins with similar charge, separations based on other properties such as size are also applied.

Proteins are readily separated by ion exchange chromatography - separation based on charge of protein

Anion exchangers are **positive charged polymers** that bind and retain **negative charged solutes** (anions) including proteins.
Cation exchangers are **negative charged polymers** that bind **positive charged solutes** (cations) including proteins.

Since a protein is a long chain of many different amino acids, each protein species has a unique electric charge which is the algebraic **sum** of the number of negative and positive amino acid **side chains**; for histidine the contribution may be a fractional positive charge since histidine is only partly protonated (approximately +0.24) at neutral pH. The N-terminal (pKa = 8) and C-terminal (pKa = 3) may also contribute charge.

e.g. At pH 7:

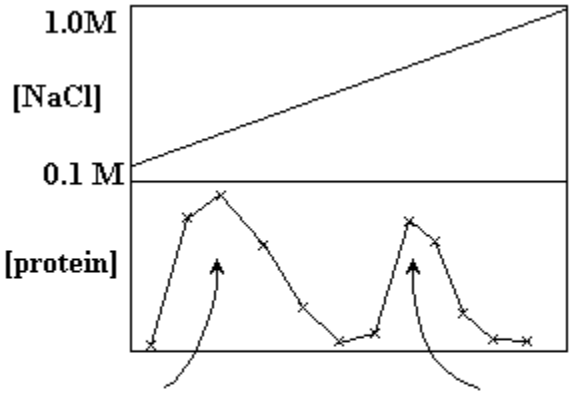
+1	**-1**	0	0	**+0.24**	0	0	0	0	**-1**	**+1**	0	**-1**
Ala –	Asp –	Leu –	Gly –	His –	Gln –	Tyr –	Cys –	Ile –	Glu –	Lys –	Ser –	Thr
Due to N-terminal												**Due to C-terminal**

Net charge = + 1 – 1 + 0.24 – 1 + 1 – 1 = –0.76

By changing the pH used in an ion exchange column it's also possible to change the net charge on a given peptide chain. However many proteins may not tolerate much pH change.

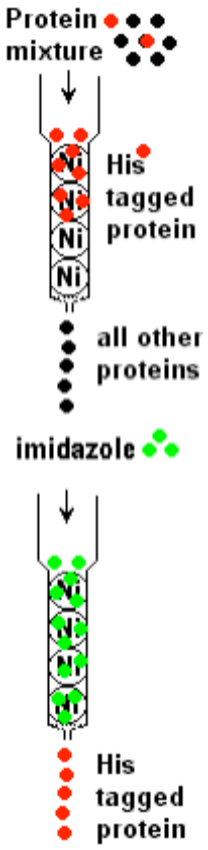
Proteins with the appropriate charge will bind to ion exchanger. They are released from the resin by **gradually increasing the concentration of neutral salt** such as NaCl or KCl, a technique known as **gradient** elution. The upper part of the graph on the right shows a gradually increasing NaCl concentration in the buffer.

The lower part shows the protein concentration measured in the buffer as it comes out of the column. Proteins that lack charge or have the same charge as the resin will not bind and are eluted quickly.
Proteins that have the opposite charge to the resin bind until the NaCl concentration has risen enough to release them. Proteins that have a higher charge will bind more tightly, and a higher NaCl concentration is needed to release or elute them. Resin and pH are chosen so that the desired protein binds moderately tightly.



Metal affinity chromatography

This is a widely used, modern technique because it can yield almost pure proteins in a single efficient step. It relies on the fact that clusters of **histidine** in a protein have high affinity for binding transition metals such as **Ni^{2+} or Co^{2+}**. A column is prepared with resin containing Ni bound to a chelating agent.

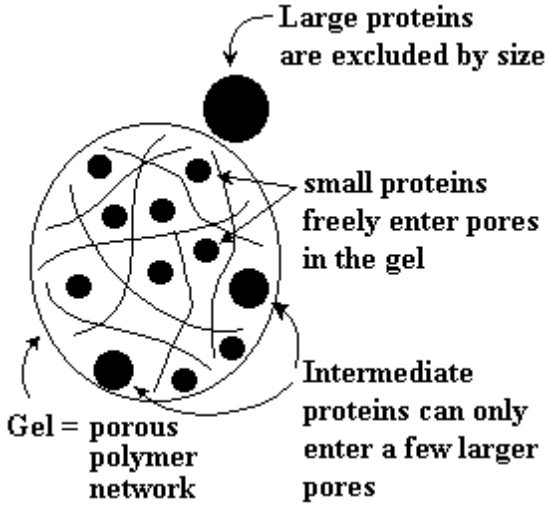


Most natural proteins do not contain histidine clusters, so will not bind to the column. However, if the protein of interest is being **artificially expressed** in yeast or bacteria, its gene can be modified to add a cluster of 6-8 histidines, known as a **His-tag**, either at the N- or the C- terminus of the polypeptide, where they are less likely to interfere with the protein's natural function.

Natural protein: Met-Pro-Ser-Leu-Ser-Tyr-*etc*
His-tagged protein: -His-His-His-His-His-His-Pro-Ser-Leu-Ser-Tyr-*etc*

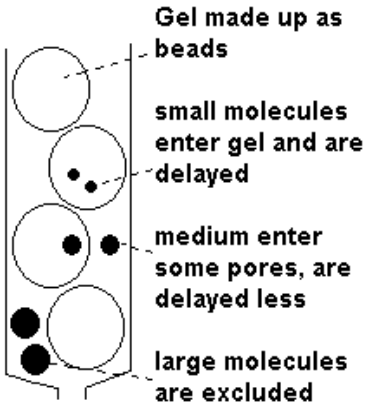
The protein of interest is now the only one that should bind to the column. The protein is then eluted by adding buffer containing **imidazole,** a molecule that resembles the histidine side chain. Imidazole will occupy the Ni^{2+} sites on the resin, allowing the His-tagged protein to pass out of the column and be collected.

Gel filtration chromatography



Another widely used chromatography technique is called **gel filtration** or **molecular exclusion** chromatography. Separation is on the basis of **molecular size**, with the largest molecules emerging first (the term **gel filtration** is a bit misleading because it seems to imply smaller molecules might pass through more easily). The stationary phase of the column consists of a hydrated **gel**, formed from a polymer, which absorbs water to form an aqueous network of open pores.

The gel is in the form of beads or granules. Molecules in the sample that are small enough can fit into the pores of the gel. Since the gel is the stationary phase, these molecules progress through the column more slowly.

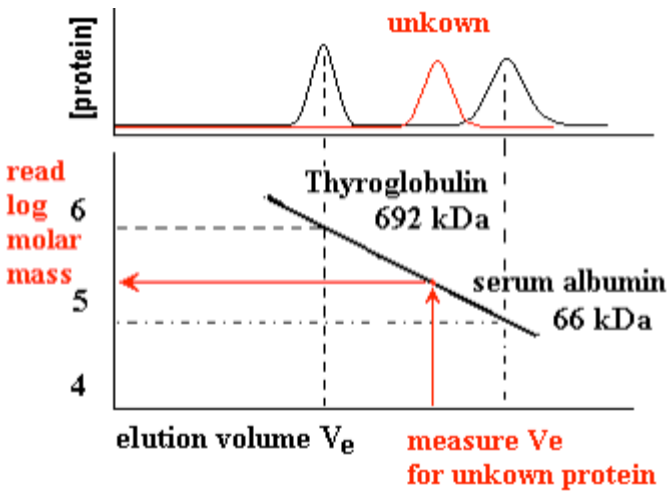


The larger molecules are excluded from the pores by their size, but this does not block them since they simply stay in the buffer that flows around the outside of the beads of gel. Therefore large molecules progress at a rate similar to the buffer, while smaller molecules are slowed down according to the extent that they can penetrate the gel.

Gel filtration can be used to estimate molecular mass of proteins

Elution volume of proteins is a **negative slope linear** function of the **log of molecular mass.**

If the elution volume is measured for two or more proteins of known molecular mass, it is then possible to **measure the elution volume of an unknown protein and estimate its molecular mass** either by interpolating the graph or by deriving the equation of the straight line.



Gels of various kinds are commercially available, and have different pore sizes for different sizes of protein.

Synopsis: Other methods used to separate amino acids and proteins also rely on the properties of the amino acids and include centrifugation, electrophoresis and mass spectrometry.

Protein structure is very much more complex than any simple organic chemical, but by eliminating detail, a pattern or hierarchy of organization emerges:

Methods for separating protein molecules other than chromatography

1. Centrifugation

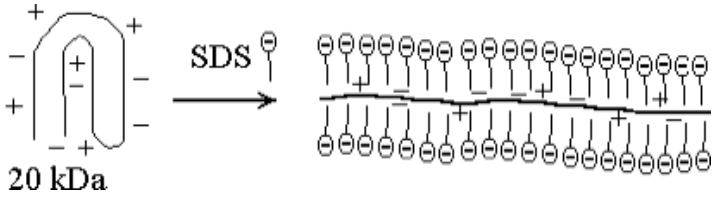
In **centrifugation**, a sample is spun in an *ultracentrifuge* at speeds from 10,000 to 75,000 rpm, producing a force from 10,000 to 500,000 x gravity. At these forces, individual molecules of proteins are large enough to **sediment** at a rate determined by their size. By measuring **sedimentation velocity**, it's possible to derive the molecular mass. It's theoretically possible, but time consuming to measure sedimentation rates of proteins down to 10 kDa.

2. Electrophoresis

Electrophoresis is separation based on movement of charged particles in an electric field. A mixture of proteins is placed between a pair of electrodes immersed in a conductive buffer solution, and a voltage of 100-1000 V applied. Positive molecules move towards the negative electrode and negative molecules move to the positive electrode. The rate of movement is a function of **size, shape, and charge**. See Lehninger Fig 3-18, p.89 for an illustration of electrophoresis apparatus.

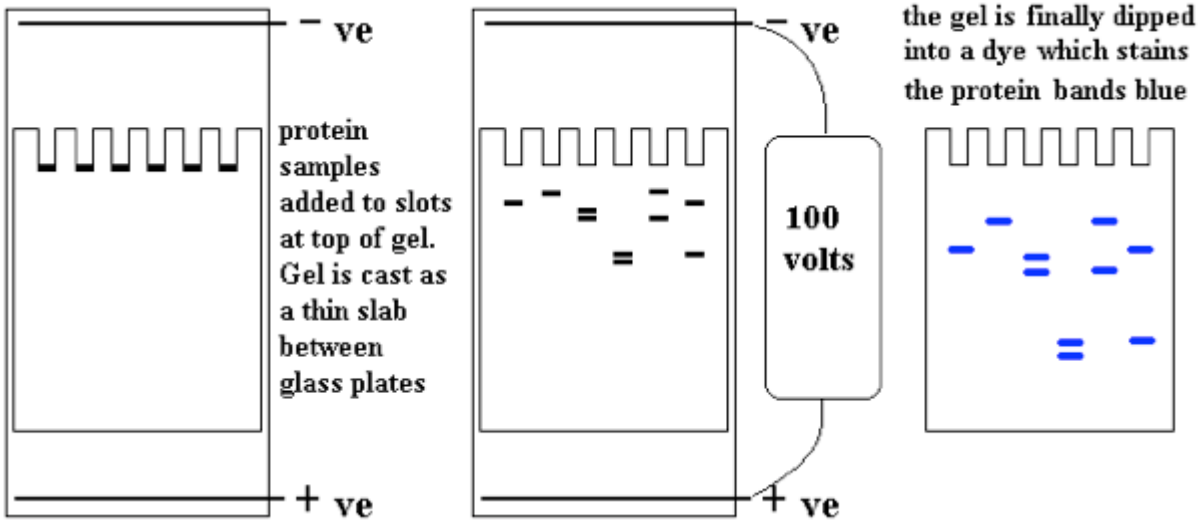
Since free solution is subject to disturbances by convection (local fluid motion caused by temperature differences), the buffer is immobilized in a **gel**. On the molecular scale, the gel is sufficiently porous to allow protein sized-molecules to pass through. Agarose gels are best for very high mass, especially DNA where molecular masses may be >10 MDa. Polyacrylamide gels are easily formed from simple chemicals in the lab, and are often used for proteins where molecular mass is in the range 10-1000 kDa. A typical polyacrylamide gel is 5-10% polymer, 90-95% buffer.

SDS-Poly**A**crylamide **G**el **E**lectrophoresis **(SDS-PAGE**) is a modified form of electrophoresis in which protein is treated with the ionic detergent **sodium dodecyl sulfate, SDS**. SDS ions **coat the protein molecules,** which adopt rodlike shapes, so that with SDS bound, all proteins have the same rodlike shape**.** The strong negative charge of the bound SDS ions overrides the somewhat variable charge of the polypeptide itself. The charge of the complex then depends on the **number** of SDS molecules bound, which in turn depends on the **size of the polypeptide**. As a result, all polypeptides now behave as if they had similar charge per unit length.

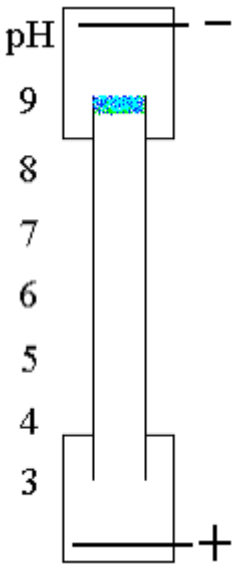


The one remaining factor that distinguishes protein in the presence of SDS is the **increased frictional resistance for larger proteins**. Hence separation is based on size, with smaller proteins being most mobile, and larger proteins being retarded (opposite to gel filtration).

We can use SDS gel electrophoresis to get information about the size or molecular mass of a polypeptide. In addition to the unknown protein samples, a set of proteins of known size are included in a separate lane of the gel. These proteins are used to create a calibration plot to match distance migrated to molecular mass. This is one of the standard laboratory methods for determining protein size or molecular mass. See Lehninger Fig 3-19, p. 90.



Another variant on electrophoresis is called **isoelectric focussing.** This is electrophoresis in a pH gradient, and separates proteins on the basis of charge.



Every protein has an **isoelectric point**, the specific pH at which the sum of negative charges is exactly equal to the sum of positive charges and its **net charge is zero**. If a protein starts at the high pH end of the gradient, it will have negative charge. If the positive electrode is placed at the low pH end of the gradient, the protein migrates towards the positive electrode, and passes through buffer of gradually decreasing pH. As the pH decreases, different side chains in the protein become protonated, causing the net negative charge to decrease. At some point the protein reaches the pH equal to its isoelectric point, where it has no charge and stops migrating since there is no attraction to either electrode. Separation occurs because each protein in a mixture has a different isoelectric point.

Two dimensional electrophoresis involves separation of a protein first by isoelectric focussing in a thin capillary tube. The spaghetti-like gel contains the partly separated proteins, and is then laid on the top edge of a conventional SDS-PAGE gel. A second separation by electrophoresis is then carried out at 90˚ to the original isoelectric focussing. See Lehninger Fig 3-20, p.90.

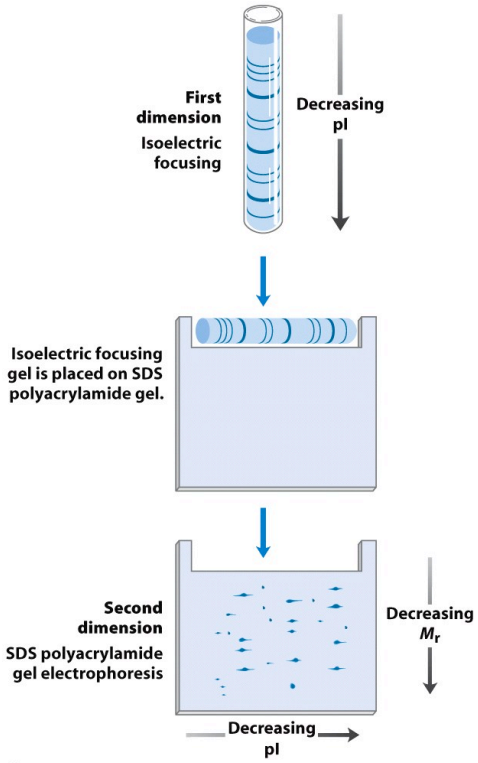
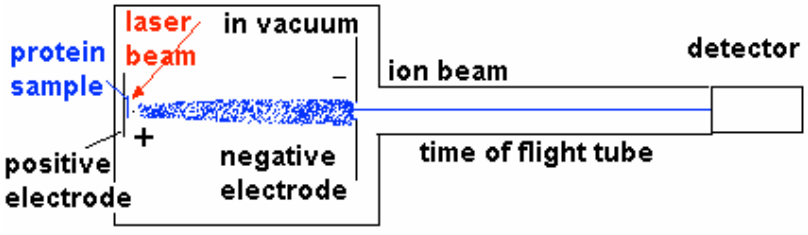


Figure 3-21a
Lehninger Principles of Biochemistry, Fifth Edition

Mass spectrometry is a technique often used in conjunction with electrophoresis to **identify** proteins (Lehninger Box 3-2, p. 98-100).

A pure protein sample is obtained as a band cut out from gel electrophoresis. The sample is either introduced into a high vacuum chamber as a superfine **spray**, or **vaporized by laser bombardment** on a positively charged electrode. This yields charged protein ionic particles, which are accelerated by attraction towards a negative electrode. A small hole in the negative electrode allows some of the ions to pass through, forming a **beam of positively charged protein ions**. The velocity of the ions depends inversely on ***m/z***, the **ratio of mass to charge**. Since the unit charge on an electron or proton is known, the **exact mass of the protein** can be calculated by measuring **the time of flight,** the time it takes the beam to travel down a tube of known length from negative electrode to detector. The protein mass can then be compared with a catalog of proteins of known mass.

The laser method is known as MALDI-TOF mass spectrometry, for Matrix-Adsorption Laser Desorption Ionization-Time of Flight mass spectrometry.



Levels of Protein Structure Hierarchy

Primary structure, the specific sequence of amino acids in the polypeptide chain

Secondary structure, the occurrence of regular repetitive patterns over short regions of the polypeptide

Tertiary structure, the overall folding of the complete polypeptide chain

Quaternary structure, the assembly of several protein molecules to form a larger complex with distinct properties

These four levels of protein structure depend in various ways on the amino acid sequence, so the chemistry for determining amino acid sequence forms our starting point for an exploration of protein structure.

Lehninger p. 82-84, 92-100

```
VAL LEU SER GLU GLY GLU TRP GLN LEU VAL  10
LEU HIS VAL TRP ALA LYS VAL GLU ALA ASP  20
VAL ALA GLY HIS GLY GLN ASP ILE LEU ILE  30
ARG LEU PHE LYS SER HIS PRO GLU THR LEU  40
GLU LYS PHE ASP ARG PHE LYS HIS LEU LYS  50
THR GLU ALA GLU MET LYS ALA SER GLU ASP  60
LEU LYS LYS HIS GLY VAL THR VAL LEU THR  70
ALA LEU GLY ALA ILE LEU LYS LYS LYS GLY  80
HIS HIS GLU ALA GLU LEU LYS PRO LEU ALA  90
GLN SER HIS ALA THR LYS HIS LYS ILE PRO 100
ILE LYS TYR LEU GLU PHE ILE SER GLU ALA 110
ILE ILE HIS VAL LEU HIS SER ARG HIS PRO 120
GLY ASP PHE GLY ALA ASP ALA GLN GLY ALA 130
MET ASN LYS ALA LEU GLU LEU PHE ARG LYS 140
ASP ILE ALA ALA LYS TYR LYS GLU LEU GLY 150
TYR GLN GLY
```

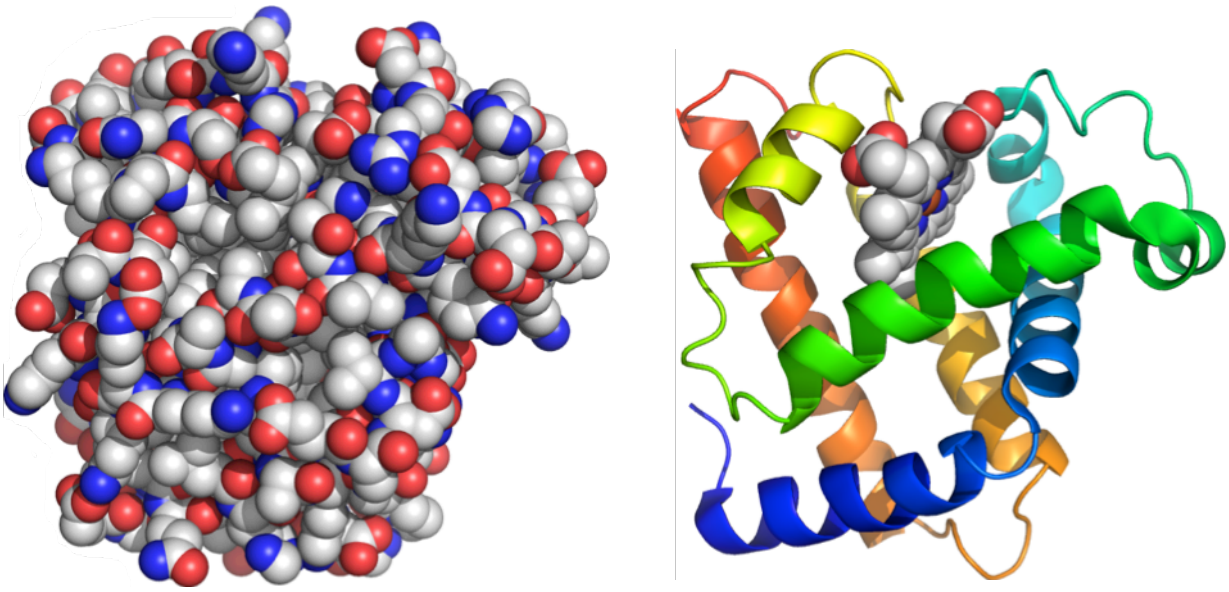
A protein consists of a long **linear chain of amino acids.**

Myoglobin, an oxygen binding protein found in muscle tissue, has 153 amino acids in its polypeptide chain (see sequence at left).

This is a relatively small protein; some proteins contain hundreds or even thousands of amino acids.

Describing the protein as a set of amino acids is one way to simplify the structure; however, it is not a complete or accurate view of the three-dimensional structure.

Protein structure is clearly **very different** from that of simple organic compounds.



We can't be unduly concerned with individual C-C or C-H bonds; attempting to view the myoglobin structure at the level of detail shown on the left makes it impossible to grasp the overall organization.

Instead, the structure of protein is viewed through a series of simplifications. Computer software can be used to **suppress detail** and make visual interpretation easier. The structure on the right is the **same myoglobin molecule seen from exactly the same viewpoint** as on the left. It **traces the path of the polypeptide backbone as a ribbon** and eliminates the side chains. Colour can be used to distinguish the sequence - the **N-terminus is shown in blue** and **progresses through the spectrum** until we reach **red at the C-terminal end**. (See also the simplified structure in Lehninger Fig 4-15 (a) compared with (d)).

Regular structural organization now becomes clearer. Protein structure can be subdivided into a hierarchy of three or four levels:

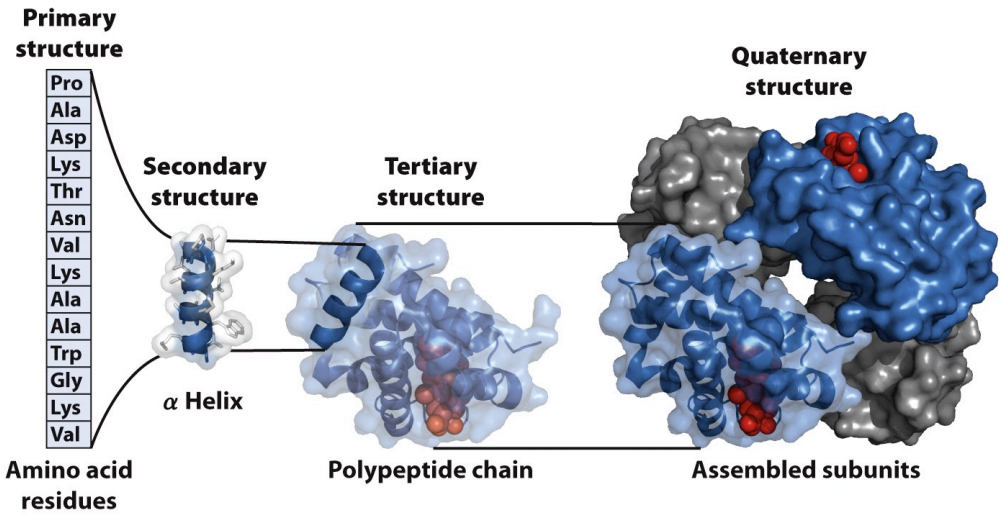


Figure 3-23
Lehninger Principles of Biochemistry, Fifth Edition

Primary structure is the linear sequence of amino acids in the polypeptide chain. By convention the **N-terminal amino acid is considered the start**, and **amino acids are numbered counting from the N-terminal end**.

The way that proteins function is dependent on **spatial organization and close placement in 3D space of particular amino acids** which might be **far apart on the linear polypeptide**. This means that each protein consists of a polypeptide that **folds up in a highly specific manner**, and the **pattern of folding** is **as important to the structure as the covalent bonds**.

Secondary structure: the polypeptide backbone is represented as a single ribbon, N terminus at lower left. The ribbon forms regular **helical** or spiral patterns in some parts, and irregular loops elsewhere. **Regular repetitive folding patterns over short sections** of the peptide chain (5-20 amino acids long), such as the helix sections appearing in myoglobin, are called secondary structure. (More details on secondary structure will follow in upcoming lectures.)

Tertiary structure is the **overall folding of the whole polypeptide**. For myoglobin, 8 helical secondary structure segments fold together to enclose a central cavity. (More details on tertiary structure will follow in upcoming lectures.)

Quaternary structure is the assembly of several individual polypeptide chains into a larger structure that has special properties. **Hemoglobin**, the O_2 binding protein of blood, consists of four independent molecules of globin, each similar in size and structure to myoglobin. The globin units associate by non-covalent forces, but behave in a **cooperative** manner to make the O_2 carrying function of hemoglobin more effective. (More details in BIOC*3560.) Only a few proteins have quaternary structure.

Investigation of structure

All higher order structure (secondary, tertiary, etc) arises from the primary structure, namely the amino acid sequence within the polypeptide. To find out how a polypeptide chain is made up, we need to find out **what amino acids are contained in it**, and **in what order or sequence they occur**. To do this it is necessary to break the peptide bonds (hydrolysis) so that the amino acids can be identified.

Practical aspects of peptide hydrolysis (attack by water)

H_2O itself hydrolyses peptides bonds extremely slowly, because neutral O: is a poor nucleophile. Although it has **two lone pairs**, **electronegative O: is less willing to share them than N: or S:**

Hydrolysis of peptides and proteins is usually done with a **catalyst**:

Acid hydrolysis is done in **6 M HCl at 110˚C;** it takes 24-72 hours to get complete breakdown of a peptide chain into single amino acids. Trp is destroyed during this process.

Base hydrolysis is done in **4 M NaOH at 110˚C**, and takes 16 hours for complete hydrolysis, but some amino acids are destroyed in strong base.

Hydrolysis may also be carried out by **digestive enzymes called proteases** (*more in upcoming Lectures)*. **Enzymes** are proteins that have a **catalytic function**, in this case to hydrolyse peptide bonds.

After hydrolysis, the amino acids present in a protein sample can be detected by chromatography e.g. **ion exchange** or **reversed phase**.

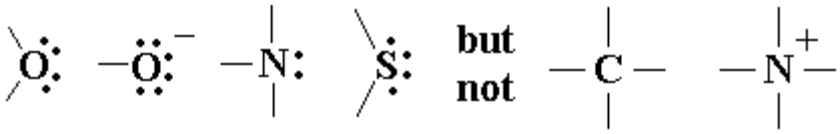
Underlying basis of chemical reactions

The chemical basis of the peptide bond-breaking reaction such as hydrolysis is a chemical process called **nucleophilic displacement**. Since many biochemical reactions that take place in living cells are initiated by **nucleophilic attack**, we need to understand what a nucleophile is, and why it can lead to peptide bond breakage.

The chemical reactivity of a molecule is a consequence of imbalances in the distribution of valence electrons of atoms. Parts of molecules that are primarily C-C and C-H bonded are **well balanced, non-polar and chemically inert**. However, where atoms seem to have **valence electrons to spare** or are **electron deficient**, or draw **electrons towards them**, these create imbalances where a reaction may occur as the atoms seek a better arrangement.

A **nucleophile** is simply an **atom with a lone pair of electrons** that is **available to share with another nucleus.** A nucleophile is "nucleus loving", and seeks out other atoms (nuclei) that are **electron-deficient**; this may be a fully or partially positively charged atom. By sharing the electron pair with another nucleus, a new bond is formed.

Atoms with lone pairs:



A **nucleophilic displacement** is a reaction in which an incoming **nucleophile X:** attacks a **target atom C** to displace another attached group. The group **Y** that detaches is called the **leaving group**:

An atom with a lone pair can use it in different ways:

1) The atom acts as **a hydrogen bond acceptor** if it simply **attracts an -OH or -NH dipole**

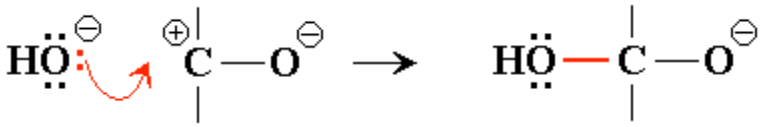
e.g. R-O: - - - H–N

2) The atom acts as **a base if it uses the lone pair to capture H^+**

e.g. H^+ + $:NH_2$-R → $^+NH_3$-R

3) The atom acts as **a nucleophile** when it **shares the lone pair, i.e. bonds to another nucleus**

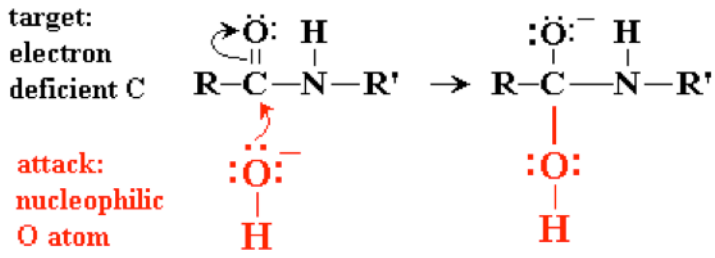
e.g.



In example 3, **O** is acting as a **nucleophile** because it has **shared** one of its lone pairs to bond to C, which is **electron-deficient.**

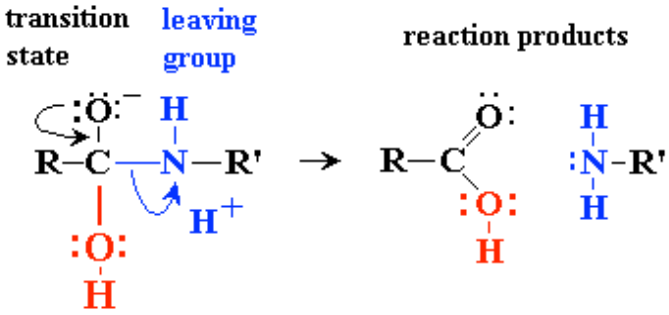
The curly arrow notation is commonly used **to indicate movement of a pair of electrons**, in this case from a non-bonded or lone pair position to form a new bond.

Hydrolysis is an **attack by H_2O, using O as a nucleophile, on a susceptible bond such as a peptide bond.**



The bond is **susceptible** because the **C atom** of the C=O is **electron deficient,** as electrons are drawn towards the electronegative O atom of C=O**.** Because the C is electron deficient, **it can accommodate the incoming electron pair** (the maximum number of valence electrons on C, N or O is 8).

This sequence then produces a **transition state**, a semi-stable halfway stage of the reaction. The transition state gives rise to **stable end products** by breaking the C–N bond. This happens because the N atom can serve as a good **leaving group,** i.e. it can hold the electrons from the breaking bond as a **lone pair.**



See Lehninger mechanism Fig 6-25, p. 215

BIOC*2580 Topic 5: Determining the amino acid sequence of a protein

Synopsis: Early methods to determine the amino acid sequence of a protein relied on cycling between basic and acidic environments to change the reactivity of the peptide. In practice, proteins need to **hydrolysed** into shorter peptides for sequencing. **Selective hydrolysis** of the polypeptide chain by **proteases,** or with **chemicals,** cuts very long polypeptides into specific fragments of more manageable sizes. Using **tandem mass spectrometry**, proteins can be sequenced and their identity determined by searching protein databases and using search tools like **BLAST**.

Determining amino acid sequence

Fred Sanger at Cambridge University was the first person to devise a method to determine the amino acid sequence of a polypeptide/protein. Over the period **1947-1953,** he worked out methods to find the **amino acid sequence of the protein hormone insulin**, and eventually won the Nobel Prize for this work**.**

Sanger introduced two important techniques:

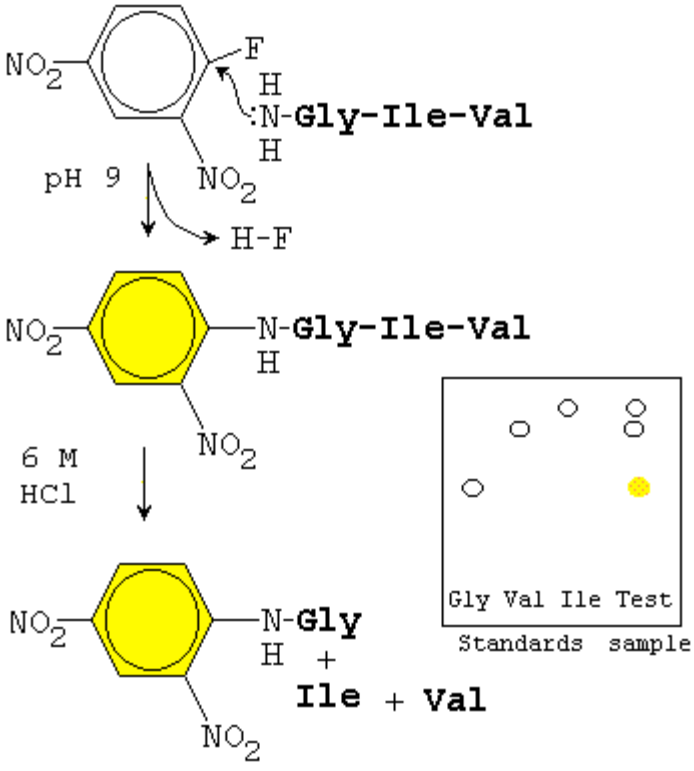
N-terminal tagging identifies the first amino acid in the chain

Limited hydrolysis breaks the chain into smaller, more manageable pieces

N-terminal tagging works because the N-terminal amino group becomes a nucleophile under mildly basic conditions (**$^{+}NH_3$-**, the normal state at pH 7 is not a nucleophile, but by increasing to pH 9 (pKa = 8 for the N-terminal of a peptide chain), it becomes deprotonated **$:NH_2$-.)** To avoid unwanted reactions at lysine, pKa 10.2, the pH should not be raised any further.

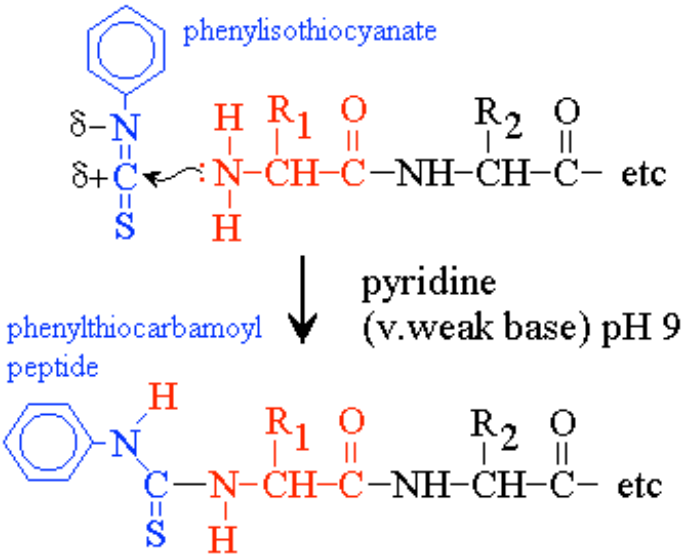
The nucleophilic N-terminal **$:NH_2$-** will then react by **displacing HF** from the reagent **fluorodinitrobenzene.**

Thus, the bright yellow dinitrophenyl group becomes bonded to the N-terminal amino acid. The tagged protein is then hydrolysed to its constituent amino acids, and the labeled (yellow) N-terminal amino acid can easily be separated and identified by chromatography.



Unfortunately, Sanger's method requires **complete hydrolysis of the peptide chain to recover the tagged amino acid**, and this destroys the rest of the peptide chain so that amino acids #2, #3 etc are not easily identified. Sanger proceeded by using **limited hydrolysis,** hydrolysis at lower temperature or for shorter time so that not all peptide bonds are broken. This creates a random mixture of dipeptides and tripeptides (short chains of 2-3 amino acids). By analyzing all the fragments, he was able to reconstruct the whole sequence, but it took 7 years to put together all the pieces of the puzzle. Luckily for Sanger, insulin is a very small protein with two chains of 21 and 30 amino acids respectively.

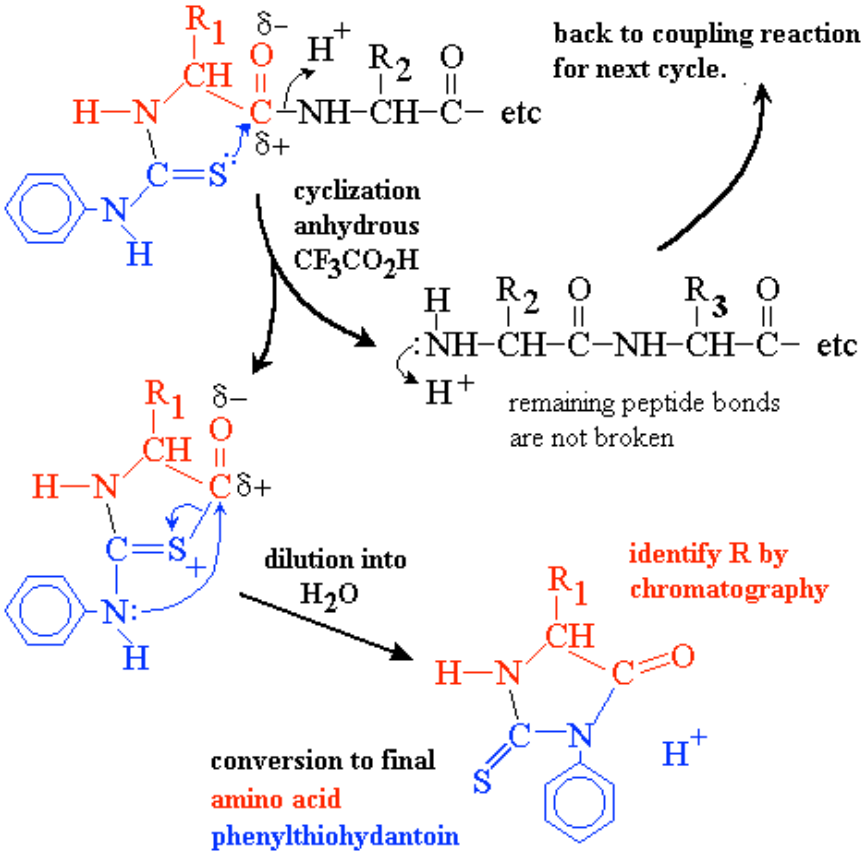
Per Edman solved the problem of hydrolyzing the complete peptide to recover the tagged amino acid in Sweden in 1956. He used the reagent **phenylisothiocyanate** to label the N-terminal end of the polypeptide sample. Lehninger Fig. 3-25, p. 94.



Phenylisothiocyanate reacts with a **deprotonated N-terminal amino group**.

Deprotonation exposes the lone pair of N, allowing it to react as a **nucleophile**, which can then attack an **electron deficient nucleus**, the C atom of isothiocyanate. This requires **mildly basic conditions, pH 9**, which is achieved by carrying out the reaction in a weak base such as pyridine.

The coupled product is called a **phenylthiocarbamoyl peptide**.



The phenylthiocarbamoyl peptide is transferred into **weak anhydrous (no H_2O) acid**, e.g. CF_3CO_2H, which causes the C=S to attack the nearest peptide bond, i.e. the one linking the N-terminal amino acid to the rest of the chain. The result is a **cyclization reaction** that splits off the first amino acid, leaving the rest of the chain intact.

Because the process is carried out in the absence of H_2O, there is no "hydrolysis".

The cyclized form of the first amino acid rearranges to the final product, an amino acid **p**henyl**t**hio**h**ydantoin, or **PTH amino acid**.

The released amino acid phenylthiohydantoin is then identified by chromatography or mass spectrometry. Because the rest of the chain is left intact, the cycle of reactions can be repeated many times, each cycle removing the currently exposed N-terminal amino acid, allowing each to be identified in sequence:

PTH-Gly + Ile-Val-Glu-Gln-Cys-Cys-Ala-Ser-Val
PTH-Ile + Val-Glu-Gln-Cys-Cys-Ala-Ser-Val
PTH-Val + Glu-Gln-Cys-Cys-Ala-Ser-Val etc.

An important factor for success is that **the two steps require contrasting conditions**:

1. **coupling** with phenylisothiocyanate occurs in **weak base**
2. **cyclization** to phenylthiohydantoin occurs in **anhydrous acid**

Because there are **two distinct phases to the reaction**, the reaction cycle remains strictly in phase. The coupling reaction at step 1 can be allowed to go to completion without any risk that some molecules of glycine make **cyclize early and expose Ile prematurely**, because cyclization requires acid. Similarly at step 2, the cyclization of Gly can proceed to completion without risk of **Ile coupling early**, since the conditions are acidic, not basic.

Another advantage is that the cycle of reactions is very easy to **automate**, and the whole process can be carried out by machine, producing one PTH amino acid every hour.

Although the Edman reaction can be repeated many times, and yields are high, there are **practical limits**. It's usual to read off sequences of 20-30 amino acids in one experiment. Sequences much over 50 or 60 amino acids are very hard to handle in a single run. Even if a reaction has 98% or 99% yield, there's a limit to the number of times you can repeat it.

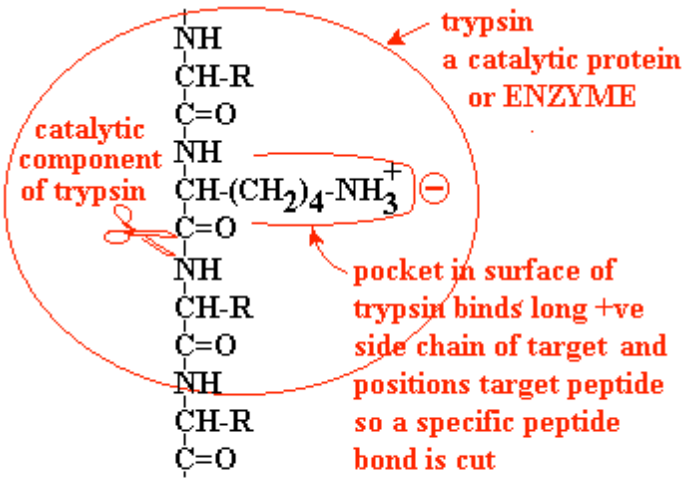
To overcome this limitation, proteins are **hydrolyzed** into **peptides** that can then be sequenced.

Selective hydrolysis

Selective hydrolysis of polypeptides allows a long polypeptide to be **cut at specific locations,** to give shorter **oligopeptides.** If the oligopeptides are no longer than 20-30 amino acids, their sequences can be determined by Edman's method or mass spectrometry.

Proteases

Selective hydrolysis can be achieved with the help of digestive enzymes called **proteases**. Enzymes are proteins that **catalyze a specific reaction**, in this case, hydrolysis of the targeted peptide bond. Lehninger p. 96-97 and Fig 3-27.



peptide chain bound to trypsin molecule

Trypsin is an enzyme that binds a polypeptide and cuts the peptide bond on the carboxylate side of the targets **Arg or Lys**.

Chymotrypsin cuts polypeptide on the carboxylate side of **Phe, Tyr or Trp**.

In both cases, if the next amino acid after the target is p**roline**, the polypeptide fails to bind to the enzyme and can't be cut at that point. Proline has an unusual conformation due to the side chain bonding to the α-amino N.

e.g. for **trypsin**
Gly------Lys-X-------Arg-Y------Lys-Pro------Asn
2 cuts give 3 peptides
Gly------Lys + X-------Arg + Y-----Lys-Pro------Asn

and for **chymotrypsin**

Gly------Phe-X-------Trp-Y------Phe-Pro------Asn
2 cuts give 3 peptides
Gly------Phe + X-------Trp + Y-----Phe-Pro------Asn

Selective chemical hydrolysis

The chemical reagent **cyanogen bromide**, CNBr, may also be used; Cyanogen bromide attacks on the carboxylate side of **methionine**, converting it to **homoserine, Hse**. Being a chemical reagent, not a catalyst, cyanogen bromide is consumed in the reaction:

Gly---Met-X----------Met-Y----------Asn
2 cuts give 3 peptides
Gly---Hse + X----------Hse + Y----------Asn

The overlap method

```
VAL LEU SER GLU GLY GLU TRP GLN LEU VAL  10
LEU HIS VAL TRP ALA LYS VAL GLU ALA ASP  20
VAL ALA GLY HIS GLY GLN ASP ILE LEU ILE  30
ARG LEU PHE LYS SER HIS PRO GLU THR LEU  40
GLU LYS PHE ASP ARG PHE LYS HIS LEU LYS  50
THR GLU ALA GLU MET LYS ALA SER GLU ASP  60
LEU LYS LYS HIS GLY VAL THR VAL LEU THR  70
ALA LEU GLY ALA ILE LEU LYS LYS LYS GLY  80
HIS HIS GLU ALA GLU LEU LYS PRO LEU ALA  90
GLN SER HIS ALA THR LYS HIS LYS ILE PRO 100
ILE LYS TYR LEU GLU PHE ILE SER GLU ALA 110
ILE ILE HIS VAL LEU HIS SER ARG HIS PRO 120
GLY ASP PHE GLY ALA ASP ALA GLN GLY ALA 130
MET ASN LYS ALA LEU GLU LEU PHE ARG LYS 140
ASP ILE ALA ALA LYS TYR LYS GLU LEU GLY 150
TYR GLN GLY
```

The sequence of myoglobin showing sites where the polypeptide chain can be cut: **red** for sites where chymotrypsin attacks; **blue** where trypsin attacks (note the underlined Lys-Pro is not cut); and **green** where cyanogen bromide attacks Met.

If myoglobin is digested in chymotrypsin, **all** the red labelled sites will be hydrolysed at the peptide bonds immediately following the target amino acid, since it's not possible to attack at only one location at a time. Similarly all the sites labelled in blue will be cut by trypsin.

This creates a series of oligopeptides with a characteristic pattern of molar masses that is unique to a given polypeptide.

In experiments to determine the complete amino acid sequence of a protein, selective hydrolysis is first carried out, and the resulting **oligopeptides** are separated by chromatography. Usually ion exchange, reversed phase or gel filtration techniques are used. The individual peptides can then be sequenced by Edman's method. Alternatively, the oligopeptide masses are easily measured by mass spectrometry (see earlier lecture) and this can be used to identify a particular protein.

After all oligopeptide sequences have been determined, the complete polypeptide sequence is deduced by the **overlap** method, see Lehninger Fig 3-27, p. 97.

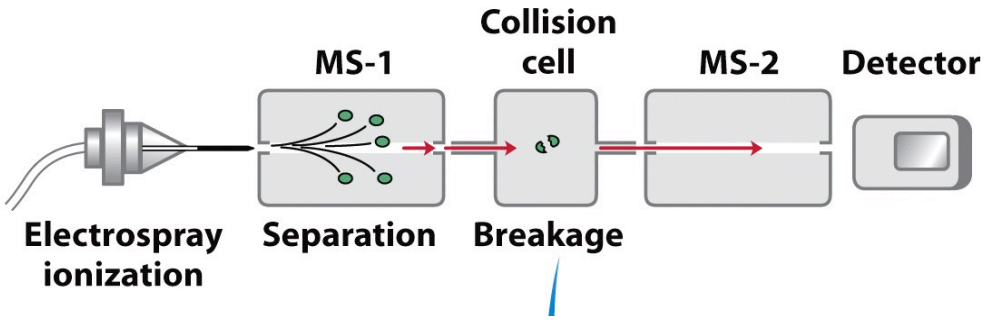
The **overlap method** is demonstrated in an animation separately from these notes. See the **Extra stuff** section on CourseLink.

Using mass spectrometry to sequence and identify proteins

Proteins can be sequenced directly using tandem mass spectrometry (t**andem MS or MS/MS**). This is a modern technique that is commonly used in the analysis of the entire protein complement of an organism or cell (an approach called **proteomics**). Since very small amounts of proteins are required, individual bands on 2D gels (see earlier lecture) can be cut out and sequenced without the need for complicated protein purification techniques.

The protein sample is **hydrolysed** into a mixture of shorter peptides using a **protease** or through chemical means. This mixture is then injected into a tandem MS; essentially two mass spectrometers in series.

In the **first MS chamber**, peptides of different masses are separated. Each of these peptides is then introduced into a **collision cell** where each peptide molecule fragments **only once**, usually at a peptide bond. In the **second MS chamber**, the masses of the peptide fragments are measured.



In one commonly used method, peptides are generated from the full-length protein with **trypsin;** each resulting peptide must therefore have a Lys or Arg at its C-terminus (see page 1 for the recognition site for trypsin).

The peptide is then moved into a **low pH** buffer. Under these conditions, acidic residues have **no charge** on the sidechain; basic residues have **+1 charge** on their sidechains.

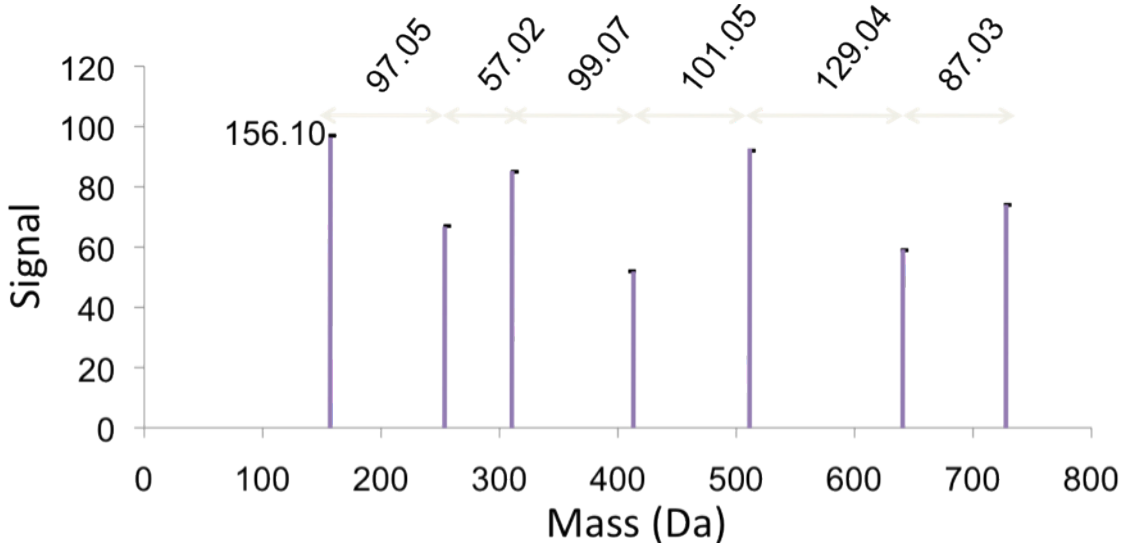
Here is an example peptide that we will use to illustrate the process of sequencing with MS:

Ser-Glu-Thr-Val-Gly-Pro-**Arg**
0 0 0 0 0 0 **+1**

This peptide then undergoes **fragmentation**, breaking one **peptide bond per molecule** on average, in a statistically random fashion. The example peptide might be fragmented at **one of six** possible break sites.

① ② ③ ④ ⑤ ⑥
↓ ↓ ↓ ↓ ↓ ↓
Ser-Glu-Thr-Val-Gly-Pro-**Arg**

These fragments then go through the **second MS**, where peptides with **charges** produce the highest signal. The resulting mass spectra would look something like this for our example peptide:



The mass of the peaks represents the mass of one charged fragment type. The difference in mass between the peaks presents the mass of one amino acid as you go from one fragment to the next. The data would be presented as follows:

Bond Broken	Uncharged Species	Charged Species	Mass Charged	Δmass	Amino Acid
none		SETVGPR+	726.37	87.03	**Ser**
①	S	ETVGPR+	639.33	129.04	**Glu**
②	SE	TVGPR+	510.29	101.05	**Thr**
③	SET	VGPR+	409.24	99.07	**Val**
④	SETV	GPR+	310.17	57.02	**Gly**
⑤	SETVG	PR+	253.15	97.05	**Pro**
⑥	SETVGP	R+	156.10		**Arg**

The **difference in mass** between fragments is used to identify the amino acid, using the following list of amino acid masses. The only **ambiguity** is leucine and isoleucine, which have exactly the same mass.

Amino acid residue masses (Da)

Glycine	57.02147	Isoleucine	113.08407	Methionine	131.04049
Alanine	71.03712	Leucine	113.08407	Histidine	137.05891
Serine	87.03203	Asparagine	114.04293	Phenylalanine	147.06842
Proline	97.05277	Aspartic acid	115.02695	Arginine	156.10112
Valine	99.06842	Glutamine	128.05858	Tyrosine	163.06333
Threonine	101.04768	Lysine	128.09497	Tryptophan	186.07932
Cysteine	103.00919	Glutamic acid	129.04264		

Using the **progression** of masses that then identify the amino acids involved, the original sequence of the peptide can be assembled. In the example, the charged amino acid (Arg) was at the **C-terminus**. Because this amino acid served as the "anchor" for the MS due to its charge under the low pH conditions of the fragmentation, we must begin assembling the sequence **from the C-terminus** and progress toward the N-terminus. The resulting sequence is:

The NCBI Database and BLAST searching

The sequence of a peptide can be compared with databanks of protein sequences of all known proteins. The NCBI (**National Center for Biotechnology Information)** database (www.ncbi.nlm.nih.gov) contains a vast amount of sequence information. With the explosion of DNA sequencing data available today, and our ability to convert those DNA sequences into the protein sequences that they encode, the amount of protein sequence information available is astronomical, and is growing every day.

One tool that is available through the NCBI is a search of protein sequences, called BLAST (**Basic Local Alignment Search Tool**). With this tool, one can enter a peptide or protein primary sequence and generate a list of sequences that contain the **highest similarity or identity.**

Several results can be obtained:

1. If the protein being analyzed is already in the protein database, then a 100% match will be generated.
2. If it is a protein that is not in the database, but has a close relative from another organism, or has a similar isoform in the same organism, then a very close match will result. These kinds of similar proteins are called **homologs;** e.g. myoglobin from horse and whale, or α-actin and β-actin in humans. This is often enough to identify the kind of protein.
3. If it is a completely new protein, then there will be very little homology in the database, and the identity of the protein will not be apparent.

Beyond sequence alignments, performing **biochemical tests** with a purified protein is the most definitive method of determining the identity of a protein.

This technique also has the added advantage of being able to identify the location and type of any **post-translational modifications** that may be present on the protein. These modifications are small molecules that are covalently bound to amino acid side chains. A good example of this kind of modification is protein phosphorylation, where the mass of an added PO_4 group would be added to a Ser, Thr, or Tyr side chain.