UNIT-II

Frederic Sanger first time achieved complete sequence of protein (bovine insulin) in 1953. For his work, he was awarded the Nobel Prize of Chemistry in (1958).
Protein sequencing refers to the techniques employed to determine the amino acid sequence of a protein. There are several applications of protein sequencing, which are:-
a) Identification of the protein family to which a particular protein belongs and finding the evolutionary history of that protein. Function prediction.

b) Prediction of the cellular localization of the protein based on its target sequence (sequence of amino acids at the N terminal end of the protein which determines the location of the protein inside the cell).

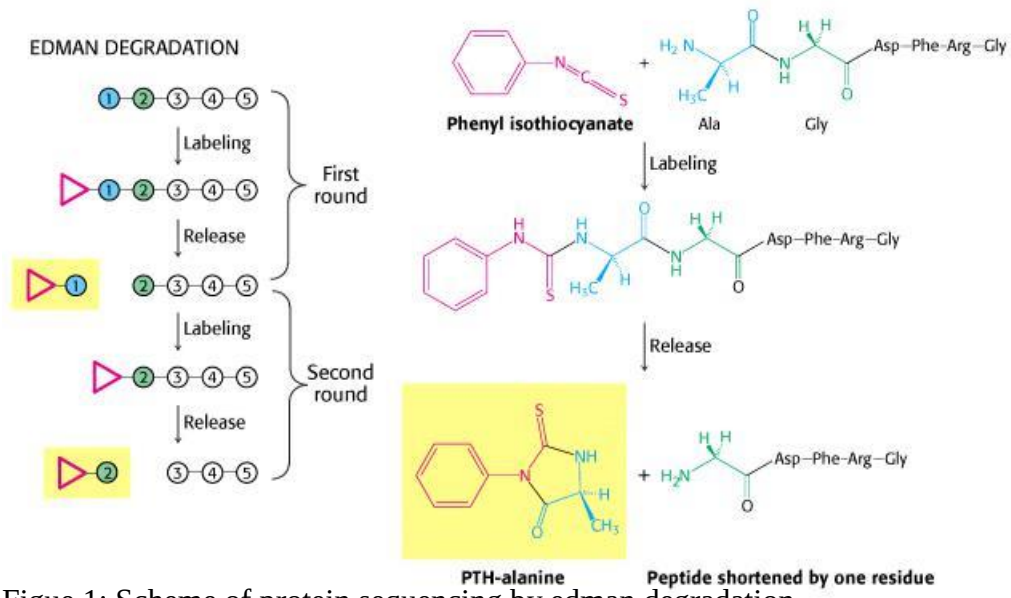
c) Prediction of the sequence of the gene encoding the particular protein.

d) Discovering the structure and function of a protein through various computational methods and experimental methods.

Till date several methods have been utilized for protein sequencing. Two main methods include Edman degradation and Mass Spectrometry. Protein sequence can also be generated from the DNA/mRNA sequence that codes for the protein, which has been explained in details in the recombinant DNA section. Here, we have discussed the most important methods used for protein sequencing and the pros and cons of each method.

Edman degradation

Before sequncingprocss is initiated, it is necessary to break all non-covalent interation by denaturants (like high concentration of urea or GuHCl). This process will also separate subunits, incase of oligomeric proteins. Ocassionally, subunits of aoligomeric protein are connected by covalent interations. In that case special treatments are required to separate subunits. The protein is treated with Edman‟s reagent (phenyl isothiocyanate) which reacts with the N-terminal amino acid and under mild acidic condition forms a cyclic compound Phenyl thiohydantoin derivative (PTH–amino acid) of N-terminal amino acid is released. Amino acid of PTH –amino acid derivative is identified by chromatographic property of the PTH –amino acid derivative. In this process N-terminal aminos acid is identified after first cycle. ***Since this method proceeds from the N terminal residue, the reaction will not work if that N-terminal of a protein is blocked (generally due to post-translational modification)***. After first cycle of the reaction, amino group of the second amino acid is free for reaction with Edman‟s reagent and at the end of reaction PTH derivative of second amino acid from N-terminal is released. The process continues till end of sequence or a disulfide bond is encountered in the sequence. PTH-cysteine derivative will remain attached with polypeptide and PTH-cystein will not be release (Fig. 1)



Figue 1: Scheme of protein sequencing by edman degradation

Thus, reduction of disulfide bond in the polypeptide sequnce needed before sequencing process can be initiated. Reduction of free cysteine can be done by use of -marcaptoethanol (Fig. 2)

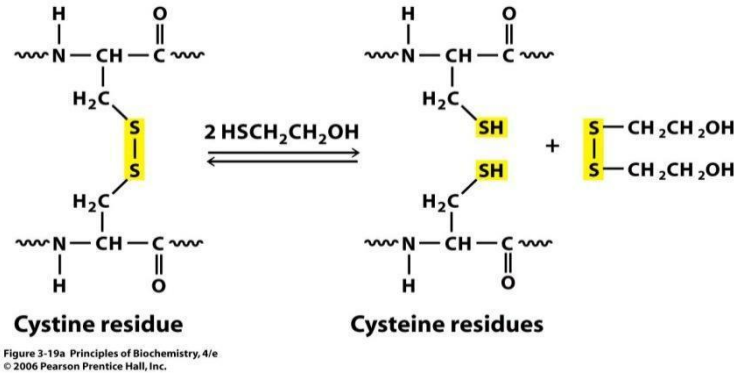


Figure 2

As free cystein can re-oxidize to form disulfide it is necessary to block free cystein. This may be done by use of iodoacetic acid or acrylonitrile (free cysteine modification) as shown in Fig3

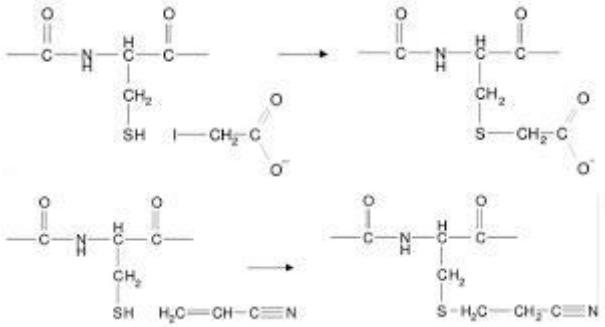


Figure 3

Other method for irreversible oxidation of disulfide bond is use of preformic acid. As shown in the figure below, performic acid oxidizes cysteine to negatively charge cysteic acid. Repulsion of negatively charged cysteic acid group prevents re-formation of disulfide and alkylation is not required. (Fig. 4)

Further, the accuracy of each cycle is 98%. So after 60 steps the accuracy is less than 30%. Thus, this method cannot be used for sequencing of proteins larger than 50 amino acids. In case of larger proteins it has to be broken down to short peptide fragments using cleavage proteases such as trypsin (cleaves a protein at carboxyl side of lysine and arginine residues) or chymotrypsin (cleaves at carboxyl side of tyrosine, tryptophan and phenylalanine). Specific cleavage can also be achieved by chemical methods like cynogen bromide, which always cleaves at carboxyl side of methionine residue (a protein with 12 methionine will yield 13 fragment polypeptide on cleavage with cynogen bromide (CNBr).

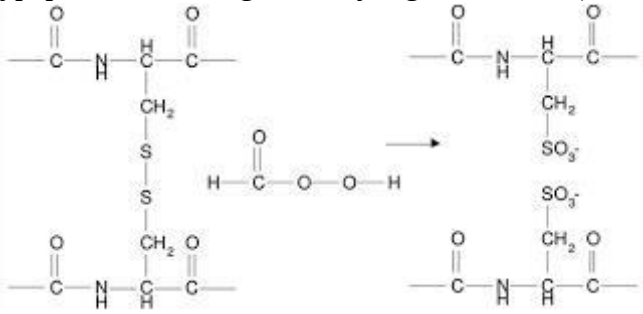


Figure 4

Protein fragments after a protease (for example trypsine) will be separated and sequenced. Let us assume that the following two peptide sequences are obtained.

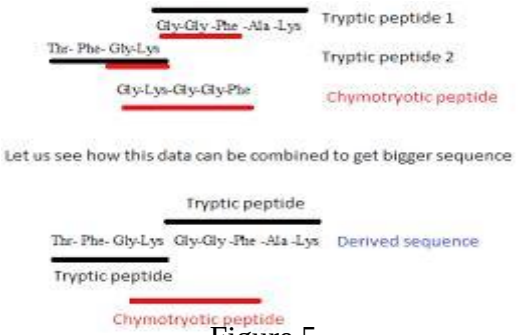


Figure 5

2) Protein sequencing using Sanger's reagent and dansyl chloride

Here, the N terminal amino acid of the protein is labeled by dyes like Sanger‟s reagent (fluoro-dinitrobenzene) or dansyl chloride. The labeled protein is then hydrolyzed by 6M HCl at 110 ºC by the above mentioned method and loaded in Dowex 50 column and the

elution profile is matched with the standard profile obtained from FNB or DNSCl derivative of all the amino acids, to obtain the N terminal amino acid. The reagents produce coloured derivatives which can be easily detected by absorbance (Fig. 6.)

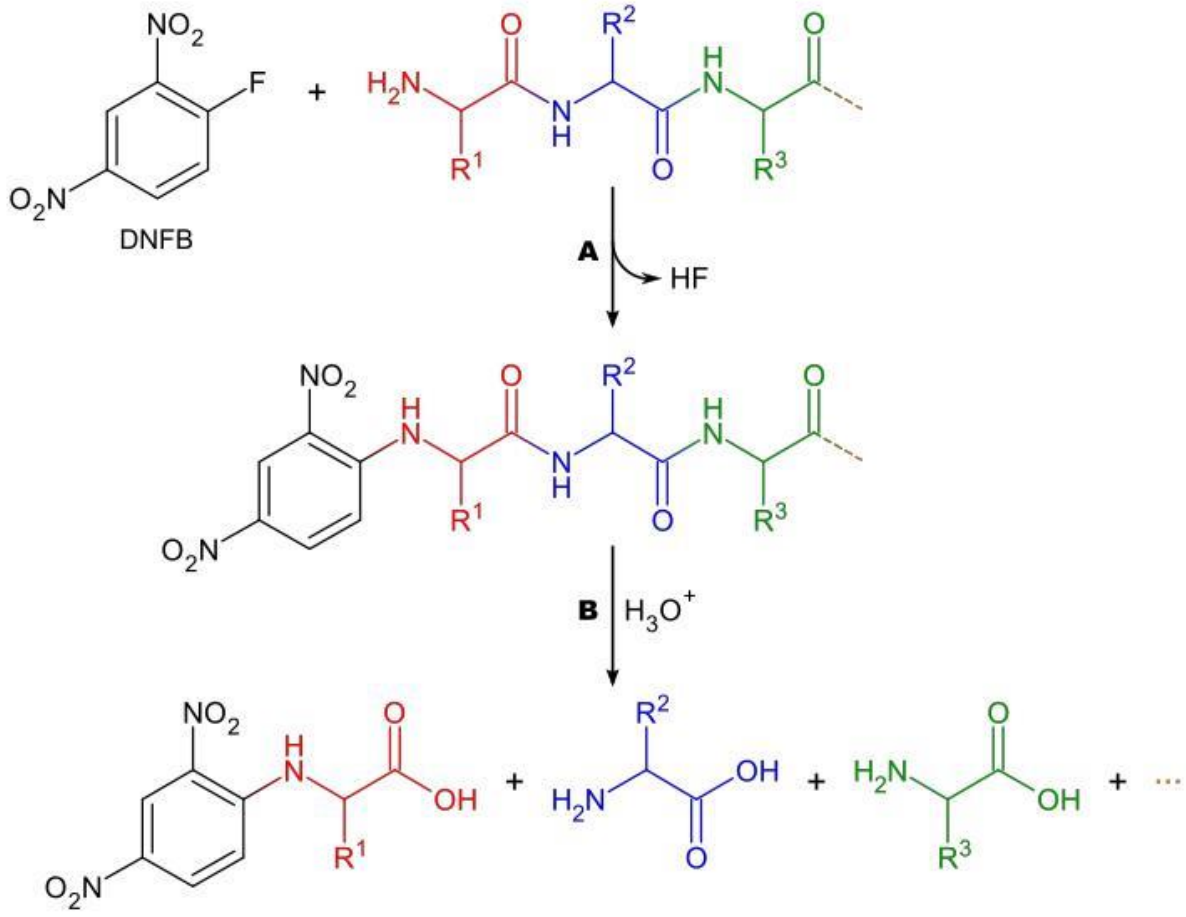


Figure 6

Disadvantages of this method include:

Once we get the N terminal amino acid, the protein is already hydrolyzed in constituent amino acids. Thus we cannot repeat the cycle with same sample. For second amino acid sequencing we require new stock of protein sample and the N-terminal residue need to be cleaved from the protein using an appropriate protease such as amino peptidase. This makes the process very tedious and complicated.

These dyes selectively labels the amine groups present in the protein and therefore can label the amine groups present in the side chains as well, which may give erroneous results.

Protein sequencing using Molecular Biology techniques

If first few N-terminal amino acid of a protein is known, complete aminoacid sequence can be derived using Molecular Biology techniques. A simple example is as follow:

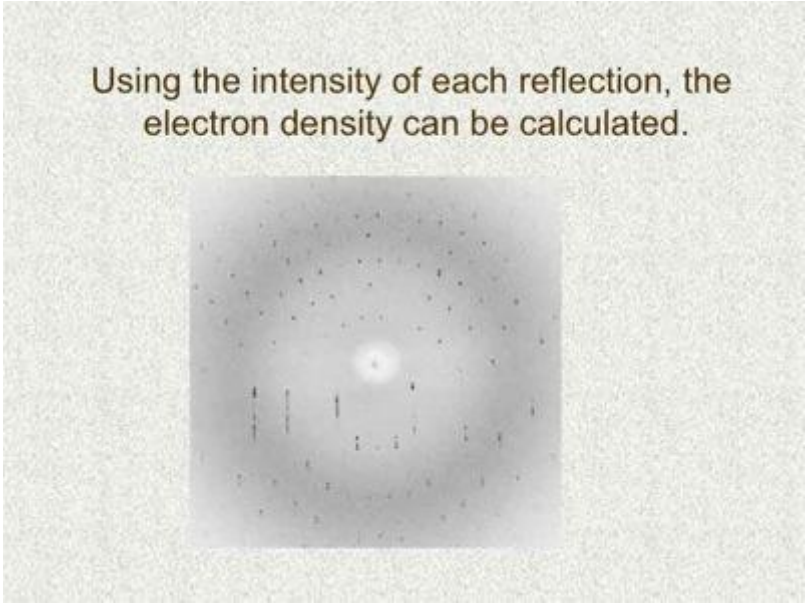
The genome sequence of *Calotropisprocera,* a plant, or the sequence of procerain B, a novel cystein protease from the plant, gene is not yet known. Thus, the only information for cloning of cDNA we have is the fifteen N-terminal amino acid residues. The double stranded cDNA can be amplified with help of degenerate primer (based of N-terminal amino acid sequence) and oligodT primer. Total RNA can be isolated from young leaf or latex of the plant and first strand of cDNA can be synthesised with oligodT primer by reverse transcription. The second strand of cDNA can be synthesised and the subsiquent amplification of double stranded cDNA can be achived by PCR with degenerate primer as forward and oligodT primer as reverse primmer. The amplified double stranded cDNA of expected size can be subjected to TA cloning and confirmed by sequencing. Once sequence of cDNA is availabe, it can be translated in protein sequnce.

PROTEIN STRUCTURE DETERMINATION

X-RAY DIFFRACTION

Historical outline

The method of protein crystallography originates from the discovery of X-rays by Conrad Röntgen, and the subsequent developments by Max von Laue, who was first to observe diffraction of X-rays and revealed the wave nature of X-rays. These discoveries were followed by the experiments by the Brags (father and son), who showed that X-ray diffraction could be used in the determination of the atomic structure of matter. However, the world had to wait for additional 45 years before the first protein structure was determined by protein crystallography. This was the structure of myoglobin, which gave the authors, Max Perutz and John Kendrew the Chemistry Nobel Prize in 1962. Since then several other protein crystallographic structures have been awarded the Nobel Prize. Among these is the prize awarded to Dorothy Hodgkin for the structures of vitamin B12 and insulin (Chemistry Prize of 1964); Johann Deisenhofer, Robert Huber and Hartmut Michel for the determination of the structure of the first membrane protein, the photosynthetic reaction center (Chemistry Prize of 1988); John E Walker for his role in the determination of the structure of ATP synthase (Chemistry Prize of 1997). Recent prizes related to protein crystallography include those awarded to Peter Agre& Roderick MacKinnon (Chemistry Prize of 2003), Roger Kornberg (Chemistry Prize of 2006), VenkiRamakrishnan, Thomas A. Steitz, Ada Yonath for the elucidation of the ternary structure of the ribosome (Chemistry Prize of 2009), and recently Brian Kobilka and Robert Lefkowitz for functional and structural studies of GPCR proteins (Chemistry Prize, 2012). **Protein X-ray crystallography** and NMR spectroscopy are currently the only two methods, which provide atomic resolutiontertary protein structures. Although, with around 90 000 entries in the Protein Data Bank (PDB), of which almost 80 000 were determined by diffraction methods, one could say that the method dominates the field of structural biology. The use of protein structure information is currently widely spread within many areas of science and industry, among which are biotechnology and pharmaceutical industry.



X-ray crystallography makes use of the diffraction pattern of X-rays that are shot through an object. The pattern is determined by the *electron density* within the crystal. The diffraction is the result of an interaction with the high energy X-rays and the electrons in the atom. The electrons get activated and their relaxation to the initial energy state emits new X-rays. Bundles of such waves can be enhanced if they are in phase, and they get canceled out if they are out of phase. Therefore the diffraction of parallel X-rays from an object containing thousands of unit molecules arranged in a regular lattice results in the enhancement and cancellation of the diffracted waves and a resulting pattern of this vectorial process can be correlated with the distribution of the electrons in the crystal.

X-ray crystallography requires the growth of protein crystals up to 1 mm in size from a highly purified protein source. Crystal growth is an experimental technique and there exists no rules about the optimal conditions for a protein solution to result in a good protein crystal. The protocol has to be established for every new type of protein. Water soluble proteins are easier to crystallize than membrane proteins. The latter tend to precipitate out of solution due to unfavorable protein-protein and protein-solute interactions. To be kept soluble in aqueous solution, membrane proteins need the addition of detergents. The presence of detergents, however, often interferes with regular arrangements of the protein complexes in the crystal resulting in diffuse diffraction pattern. If membrane proteins contain large extra-membranous domains, these water soluble domains can be cleaved off from the membrane buried domain and crystallized individually.

X-rays have a wavelength of 0.2Å to 2.0Å. The wave length, as in an optical microscope, determines the resolution limit of half the applied wave length. X-rays are therefore suited for the atomic distances which reside in the angstrom range. X-rays are high energy electromagnetic radiation and can be recorded on X-ray sensitive film, the normal technique to record diffraction patterns of protein crystals.

X-rays that interact with an electron cause it to oscillate. Oscillating electrons serve as a new source of X-rays that propagate away from the stimulated electron. The waves of

neighboringelectrons super impose and depending on their being in-phase or out of phase result in a signal or in no signal at all. Diffraction by a crystal can be regarded as the reflection of the primary beam by sets of parallel planes that define the dimensions of the unit cell (the smallest repetitive pattern) of the crystal. The relationship between reflection angle, , the distance between the planes, d, and the wavelength, , is given by Bragg's law:

2dsin = Bragg's Law

The 2-dimensional distribution of the diffraction pattern can be calculated back into a 3-dimensional space of the electron distribution causing the diffraction. The mathematical formalism to do this is called *Fourier transformation*. The distances between the spots inversely correlates with the distances of the unit cell in the crystal and the intensity of the spots with the density of electrons in the molecular structure. The exact location of the electrons, however, is lost in a single diffraction pattern, because the information of the phase of the diffracted beams is not given. This is called the *phase problem* and is the hardest obstacle to overcome. The phase problem requires at least 3 different protein crystals with identical unit cell geometry and the inclusion of evenly spaced *heavy metals* or derivatives in the protein structure that give information about the relative phase in the individual crystal. The diffraction spots originating from the electron shell of the heavy metals can easily be identified and distinguished from other electron dens centers in the crystal. From the heavy metal location in the unit cell and the phase shift can be determined. The method to solve the phase problem using different crystals with identical protein structures containing regularly but infrequently spaced heavy metals or protein isoforms is known as *multiple isomorphous replacement*.

The amplitudes and phases of the diffraction data are used to calculate an *electron-density map*of the repeating unit of the crystal. This is a step that involves the *interpretation of the raw data*. This step is sensitive to the resolution of the diffraction data, which in turn is determined by the *quality of the protein crystal*, i.e., the regularity of the lattice of the protein in the unit cell and the regularity of the distribution of the heavy atom inclusions. The interpretation of the diffraction data needs information about the amino acid sequence of the protein because depending on the resolution of the data different amino acids can have indistinguishable electron densities (e.g. Tyr and Phe, or Leu and Ile).

Initial models of protein structures due to limits in the resolution have to be refined. This is often achieved by comparing the experimental data with the optimal structure obtained by computer modeling. The difference in experimental structure and hypothetical structure is given as *R-factor*.

Nuclear magnetic resonance, or NMR

Nuclear magnet resonance obtains the same high resolution using a very different strategy. NMR measures the distances between atomic nuclei, rather than the electron density in a molecule. With NMR, a *strong, high frequency magnetic field* stimulates atomic nuclei of the isotopes H-1, D-2, C-13, or N-15 (they have a magnetic spin) and measures the frequency of the magnetic field of the atomic nuclei during its oscillation period back to the initial state. The important step is to determine which resonance comes from which spin. The distance and type of neighboring nuclei determines the resonance frequency of the stimulated atomic

nuclei. This dependence on next neighbors known as *chemical shift* (or spin-spin coupling constant) andreflects the local electronic environment and the information contained in *1-D NMR spectra*. For proteins, NMR usually measures the spin of protons. The following reasons make the H-1 NMR spectroscopy the method of choice for biological macromolecules:

- H are present at many sites in proteins, nucleic acids, and polysaccharides

- H have a high abundance for each site

- H nuclei is the most sensitive to detect

1-D spectra contain the information about all the chemical shifts of all the H in the protein. The frequency resolution is often not enough to distinguish individual chemical shifts. 2-D NMR solves this problems by containing information about the relative position of H in molecular structures. 2-D NMR spectra contain information about interaction between H that are covalently linked through one or two other atoms (COSY or *correlation spectroscopy*). Alternatively, pairs of H that can be close in space, even if they are from residues that are not close in sequence (NOE spectra, or *Nuclear Overhauser Effect*). A complete structure can thus be calculated by sequentially assigning cross peak correlations in 2-D spectras. Currently, the size limit for proteins amenable to NMR solution structure analysis is about 200 amino acids. An important feature of the identification of cross peaks is that regular patterns can be recognized that stem from secondary structure elements such as alpha helices and parallel or anti-parallel beta sheets because they contain typical hydrogen bonding networks.

Fig. Observed NOEs in antiparallel and parallel b sheets

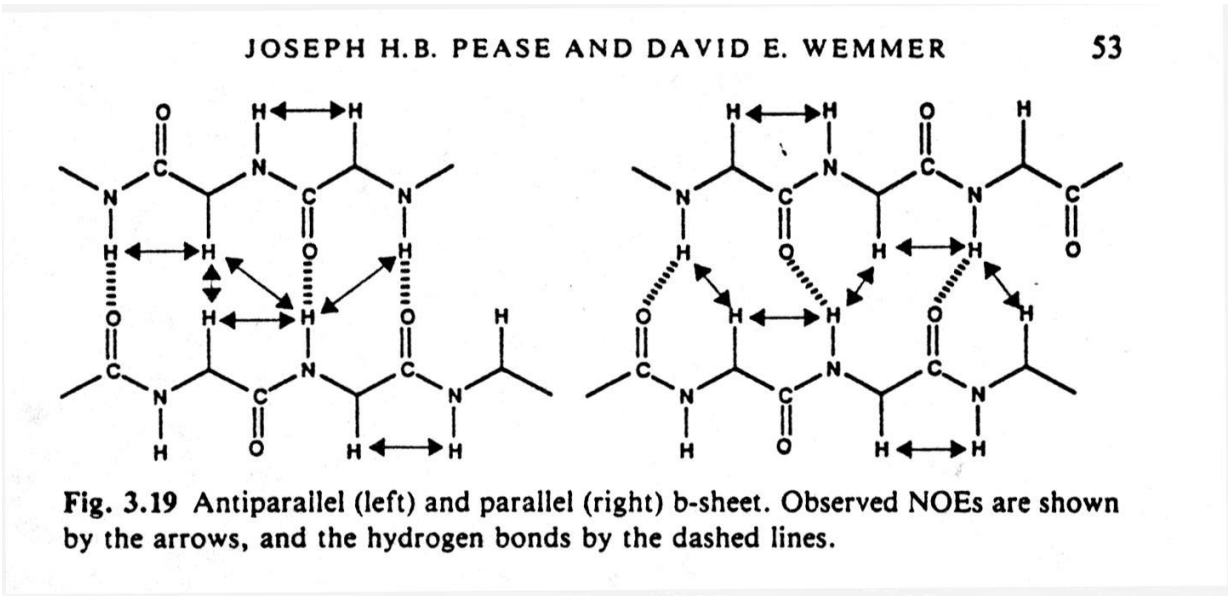


Fig. 3.19 Antiparallel (left) and parallel (right) b-sheet. Observed NOEs are shown by the arrows, and the hydrogen bonds by the dashed lines.

NMR also requires the knowledge of the *amino acid sequence*, but the protein does not have to be in an ordered crystal, yet high concentrations of solubilized protein must be available (NMR structures are therefor also called *solution structures*). In biopolymers, the primary structure (sequence) logically breaks up the molecule into groups of coupled spins normally

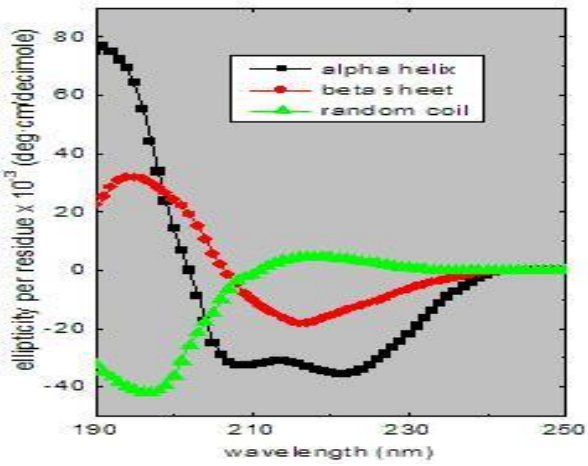
one or two groups per residue. This is true not only for proteins, but also for nucleic acids and polysaccharides.

4. X-ray crystallography and NMR are complementary techniques

NMR	**X-ray crystallography**
short time scale, protein folding	long time scale, static structure
solution, purity	single crystal, purity
< 20kD, domain	any size, domain, complex
functional active site	active or inactive
domains	domains
atomic nuclei, chemical bonds	electron density
resolution limit 2-3.5Å	resolution limit 2-3.5Å
primary structure must be known	primary structure must be know (except if resolution is 2Å or better for every single residue)

CIRCULAR DICHORISM SPECTROSCOPY

Circular dichroism (CD) spectroscopy measures differences in the absorption of left-handed polarized light versus right-handed polarized light which arise due to structural asymmetry. The absence of regular structure results in zero CD intensity, while an ordered structure results in a spectrum which can contain both positive and negative signals.



Circular dichroism spectroscopy is particularly good for:

- determining whether a protein is folded, and if so characterizing its secondary structure, tertiary structure, and the structural family to which it belongs
- comparing the structures of a protein obtained from different sources (*e.g.* species or expression systems) or comparing structures for different mutants of the same protein
- demonstrating comparability of solution conformation and/or thermal stability after changes in manufacturing processes or formulation
- studying the conformational stability of a protein under stress -- thermal stability, pH stability, and stability to denaturants -- and how this stability is altered by buffer composition or addition of stabilizers and excipients
  - CD is excellent for finding solvent conditions that increase the melting temperature and/or the reversibility of thermal unfolding, conditions which generally enhance shelf life
- determining whether protein-protein or protein-ligand interactions alter the conformation of protein.
  - If there are any conformational changes, this will result in a spectrum which will differ from the sum of the individual components. Small conformational changes have been seen, for example, upon formation of several different receptor/ligand complexes.

Determination of Protein Secondary Structure by Circular Dichroism:

Secondary structure can be determined by CD spectroscopy in the "far-UV" spectral region (190-250 nm). At these wavelengths the chromophore is the peptide bond, and the signal arises when it is located in a regular, folded environment.

Alpha-helix, beta-sheet, and random coil structures each give rise to a characteristic shape and magnitude of CD spectrum. This is illustrated by the graph to the right, which shows spectra for poly-lysine in these three different conformations. The approximate fraction of each secondary structure type that is present in any protein can thus be determined by analyzing its far-UV CD spectrum as a sum of fractional multiples of such reference spectra for each structural type.

Like all spectroscopic techniques, the CD signal reflects an average of the entire molecular population. Thus, while CD can determine that a protein contains about 50% alpha-helix, it cannot determine which specific residues are involved in the alpha-helical portion.

Far-UV CD spectra require 20 to 200 µl of solution containing 1 mg/ml to 50 µg/ml protein, in any buffer which does not have a high absorbance in this region of the spectrum. (High concentrations of DTT, histidine, or imidazole, for example, cannot be used in the far-UV region.) Note that for many formulated protein samples the absorbance due to the excipients prevents collecting spectra below 200 nm (and even 200 nm is often not possible). When that is true the accuracy reliability of secondary structure calculations (the actual percentages of different structures) is compromised, but the validity of spectral comparisons is not.

Information About Protein Tertiary Structure from Circular Dichroism:

The CD spectrum of a protein in the "near-UV" spectral region (250-350 nm) can be sensitive to certain aspects of tertiary structure. At these wavelengths the chromophores are the aromatic amino acids and disulfide bonds, and the CD signals they produce are sensitive to the overall tertiary structure of the protein.

Signals in the region from 250-270 nm are attributable to phenylalanine residues, signals from 270-290 nm are attributable to tyrosine, and those from 280-300 nm are attributable to tryptophan. Disulfide bonds give rise to broad weak signals throughout the near-UV spectrum.

If a protein retains secondary structure but no defined three-dimensional structure (*e.g.* an incorrectly folded or "molten-globule" structure), the signals in the near-UV region will be nearly zero. On the other hand, the presence of significant near-UV signals is a good indication that the protein is folded into a well-defined structure.

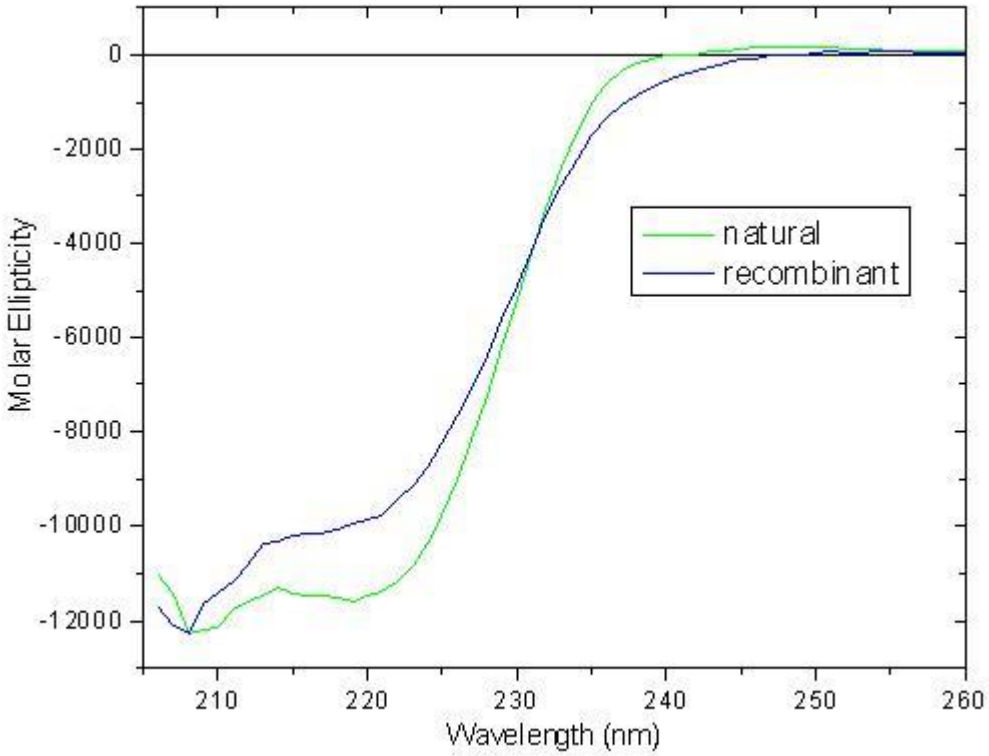
The near-UV CD spectrum can be sensitive to small changes in tertiary structure due to protein-protein interactions and/or changes in solvent conditions.

The signal strength in the near-UV CD region is much weaker than that in the far-UV CD region. Near-UV CD spectra require about 1 ml of protein solution with an OD at 280 nm of 0.5 to 1 (which corresponds to 0.25 to 2 mg/ml for most proteins).

Demonstrating Comparability of Conformation

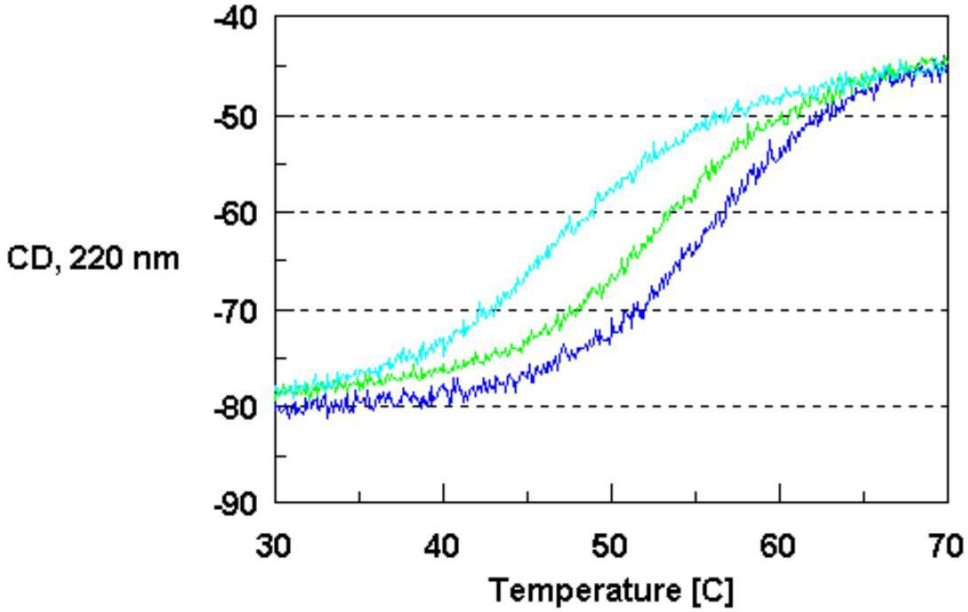
Often it is necessary to demonstrate that different lots of a protein have equivalent conformations, for example after a scale-up in the purification process or to qualify a new manufacturing site, and CD can be a good tool for this.

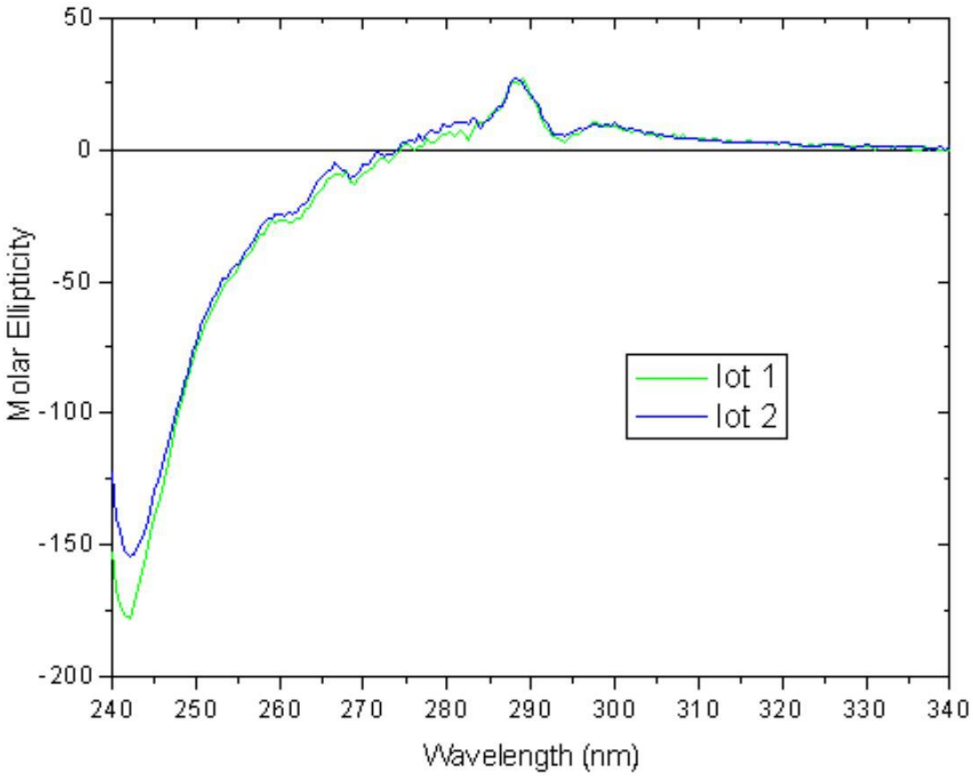
The data below show a case where the far-UV spectra show that the recombinant form of an enzyme clearly does <u>not</u> have the same secondary structure as the natural protein (*i.e.* the recombinant protein is not properly folded).



Such cases of significant differences in secondary structure are, however, unusual. More typically subtle differences in conformation do not produce a detectable difference in far-UV CD, but may produce a difference in near-UV CD. One such example, for different lots of a monoclonal antibody, is shown below. This small but reproducible difference at ~240 nm correlates with differences in the stability of different lots of this antibody.

Thermal Stability by Circular Dichroism


CD, 220 nm
-40
-50
-60
-70
-80
-90
30
40
50
60
70
Temperature [C]

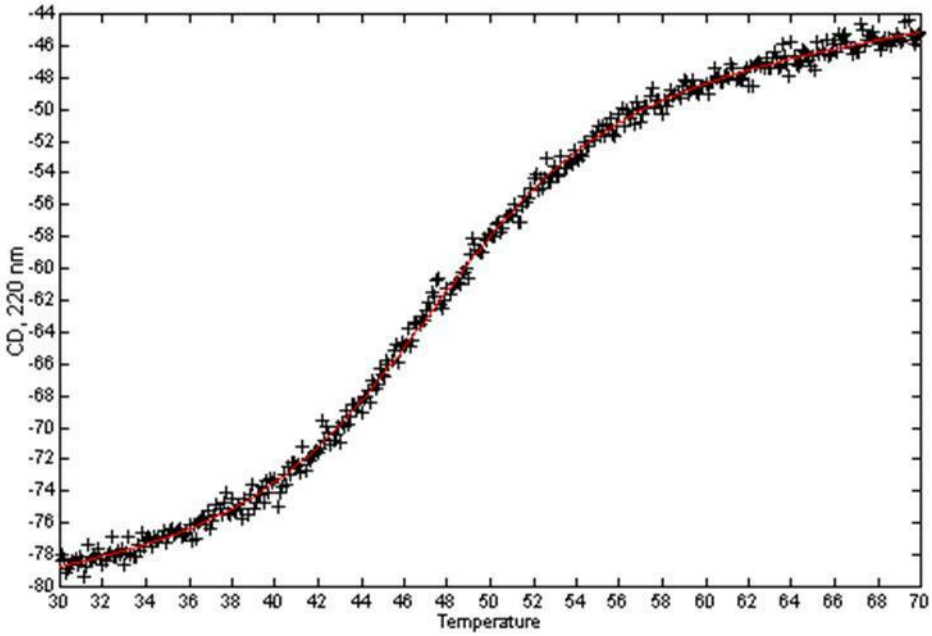

Molar Ellipticity
50
0
-50
-100
-150
-200
240
250
260
270
280
290
300
310
320
330
340
Wavelength (nm)
lot 1
lot 2

Thermal stability is assessed using CD by following changes in the spectrum with increasing temperature. In some cases the entire spectrum in the far- or near-UV CD region can be followed at a number of temperatures. Alternatively, a single wavelength can be chosen which monitors some specific feature of the protein structure, and the signal at that wavelength is then recorded continuously as the temperature is raised. CD is often used to assess the degree to which solution pH, buffers, and additives such as sugars, amino acids or salts alter the thermal stability.

This graph illustrates thermal scans done in our lab for the same recombinant protein in 3 different buffers. While unfolding is completely reversible under all these conditions, clearly there are quite significant differences in thermal stability.

Many proteins aggregate or precipitate quickly after they are unfolded ("melted"), making unfolding irreversible. The reversibility of the unfolding reaction can be assessed by cooling the sample and then heating again to see if the unfolding reaction is duplicated. Finding solvent conditions that make unfolding reversible may be actually be more important for long-term stability (shelf life) than raising the melting temperature.

If (and only if) the melting is fully reversible, the melting temperature is directly related to conformational stability, and the thermodynamics of protein folding can be extracted from the data. The fact that thermal unfolding can generally be measured by CD at much lower concentrations than by DSC increases the probability of reversible reactions and of thermodynamically interpretable data.



This graph illustrates a detailed analysis of one of the data sets shown above, using custom software developed in our lab. The data (+) were fitted to a simple thermodynamic unfolding model (solid line). The fit returns the melting temperature (midpoint of the transition) as 47.3 +/- 0.1 °C. The width of the transition region is related to the enthalpy of unfolding, *H*, which the fit returns as 52 +/- 2 kcal/mol. Fitting the data also allows a more reproducible measurement of the onset of unfolding, a temperature which is often more relevant for formulation and shelf-life considerations than the midpoint. The onset (defined as the temperature at which 5% of the protein is unfolded) occurs at 36.1 +/- 0.3 °C in this case.

If the protein precipitates or aggregates as it is unfolded, the melting reaction will be irreversible, and the melting temperature will reflect the kinetics of aggregation and the solubility of the unfolded form of the molecule as well as the intrinsic conformational stability.

The cooperativity of the unfolding reaction is measured qualitatively by the width and shape of the unfolding transition. A highly cooperative unfolding reaction indicates that the protein existed initially as a compact, well-folded structure, while a very gradual, non-cooperative melting reaction indicates that the protein existed initially as a very flexible, partially unfolded protein or as a heterogeneous population of folded structures.

Melting of Secondary Structure

Changes in secondary structure, monitored in the far-UV CD region, can be determined with as little as 50 µg of protein, at concentrations of 0.2 mg/ml. By following changes over the entire far-UV CD region we can determine whether at high temperatures the protein is losing all of its secondary structure, loses only a portion of its secondary structure, or simply undergoes conformational change involving a change in secondary structure. Occasionally the unfolded form of a protein will possess a defined but totally different secondary structure than the native form (*e.g.*, TNF-alpha contains beta-sheet when folded, but alpha-helix when melted, and many proteins form amyloid-like aggregates following a transition from alpha-helix to beta-strand).

Melting of Tertiary Structure

Changes in tertiary structure can be followed by monitoring changes in the near-UV CD region. Due to the weaker signal in this region this requires 1-3 mg of protein. Such studies will reveal whether the melting of a protein occurs in a single step (with concurrent loss of both secondary and tertiary structure), or in a two-step reaction.

Melting of Protein Complexes

The effect of forming a protein-protein complex (*e.g.* ligand/receptor or antigen-antibody) on the thermal stability of the individual proteins in the complex can also be determined. This works best if the individual proteins have CD spectra which are quite different from each

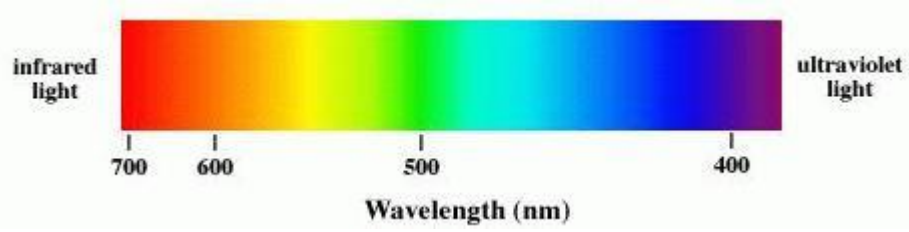
other, such that changes at specific wavelengths can be monitored to follow changes in the corresponding protein. In such cases it is possible to determine whether there is an increase in stability of one or both of the proteins following complex formation.

Infrared spectroscopy of proteins

During the last years the use of Fourier Transform Infrared spectroscopy (FTIR) to determine the structure of biologicalmacromolecules has dramatically expanded. The complete three-dimensional structure of a protein at high resolution can be determined by X-ray crystallography. This technique requires the molecule to form a well ordered crystal which is not possible for all proteins. An alternative to X-ray crystallography is multidimensional nuclear magnetic resonance (NMR) spectroscopy. Using NMR spectroscopy structures of the proteins can be determined in solution. The interpretation of the NMR spectra of large proteins is very complex, so its present application is limited to small proteins (~15-25 kDa). These limitations have led to the development of alternative methods that are not able to generate structures at atomic resolution but provide also structural information on proteins (especially on secondary structure). These methods include circular dichroism (CD) and vibrational (infrared and RAMAN) spectroscopy.
The new technique of FTIR spectroscopy requires only small amounts of proteins (1mM) in a variety of environments. Therefore, high quality spectra can be obtained relatively easy without problems of background fluorescence, light scattering and problems related to the size of the proteins. The omnipresent water absorption can be subtracted by mathematical approaches. Methods are now available that can separate subcomponents that overlap in the spectra of proteins. These facts have made practical biological systems amenable to studies by FTIR spectroscopy.

Basic principles of infrared (IR) absorption

We will focus on very few aspects here, because many textbooks present excellent descriptions of the basis of IR spectroscopy (see for example *Campbell &Dwek, in Biological Spectroscoy, Benjamin Cummings, Menlo Park, CA 1984* and *Brey, Physical Chemistry and its Biological Applications, Academic Press, New York, 1984, p.133*). IR spectroscopy is the measurement of the wavelength and intensity of the absorption of infrared light by a sample. Infrared light is energetic enough to excite molecular vibrations to higher energy levels.



Electromagnetic spectrum

frequencyrange (Hz)	wavelength range	type of radiation	type of transition
10^{20} - 10^{24}	10^{-12} - 10^{-16} m	gamma rays	nuclear
10^{17} - 10^{20}	1 nm - 1 pm	x-rays	inner electrons
10^{15} - 10^{17}	400 - 1 nm	ultraviolet light	outer electrons
$4.3x10^{14}$ - $7.5x10^{14}$	700 - 400 nm	visible light	outer electrons
10^{12}-10^{14}	2.5 um - 700 nm	infrared light	vibrations
10^{8} - 10^{12}	1 mm - 2.5 um	microwaves	rotations
10^{0} - 10^{8}	10^{8} - 1 m	radio waves	spin flips

The infrared spectra usually have sharp features that are characteristic of specific types of molecular vibrations, making the spectra useful for sample identification.

Table of characteristic IR bands

X-H vibrations	***bond***	***wavenumbers (cm^{-1})***
hydroxyl	O-H	3610-3640
amines	N-H	3300-3500
aromatic rings	C-H	3000-3100
alkenes	C-H	3020-3080
alkanes	C-H	2850-2960
triple bonds		2500-1900
double bonds		1900-1500
deformation/heavy atoms		1500-

For a molecule of N atoms, 3N-6 fundamental vibrations (or normal modes) exist (3N-5 if the molecule is linear). Therefore, for the linear CO2 molecule 4 normal modes have to be expected.

Normal modes for CO_2

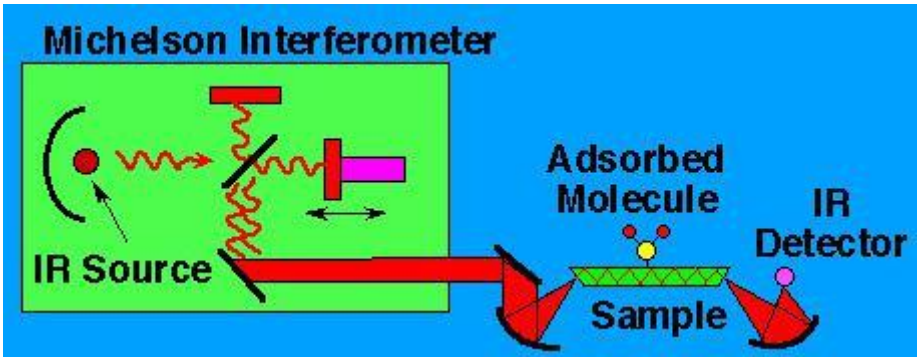
		cm^{-1}	*IR*	*RAMAN*
stretching (sym.)	-> <- O==C==O	1340	-	+
stretching (asym.)	-> <- <- O==C==O	2349	+	-
deformation	/| /| O==C==O |	667	+	-
deformation	+ - + O==C==O	667	+	-

Fourier Transform Infrared (FTIR) spectroscopy

To use the Fourier Transform Infrared Spectroscopy, a continuum source of light (such as a Nernst Globar) is used to produce light over a broad range of infrared wavelengths. Light coming from this continuum source is split into two paths using a half-silvered mirror; this light is then reflected from two mirrors back onto the beamsplitter, where it is recombined. One of these mirrors is fixed, and the second is movable. If the distance from the beamsplitter to the fixed mirror is not exactly the same as the distance from the beamsplitter to the second mirror, then when the two beams are recombined, there will be a small difference in the phase of the light between these two paths. Because of the "superposition principle" constructive and destructive interference exist for different wavelengths depending of the relative distances of the two mirrors from the beamsplitter.

It can be shown that if the intensity of light is measured and plotted as a function of the position of the movable mirror, the resultant graph is the Fourier Transform of the intensity of light as a function of wavenumber . In FTIR spectroscopy , the light is directed onto the

sample of interest, and the intensity is measured using an infrared detector. The intensity of light striking the detector is measured as a function of the mirror position, and this is then Fourier-transformed to produce a plot of intensity vs. wavenumber.
As radiation source a Michelson Interferometer is used (see the drawing below).

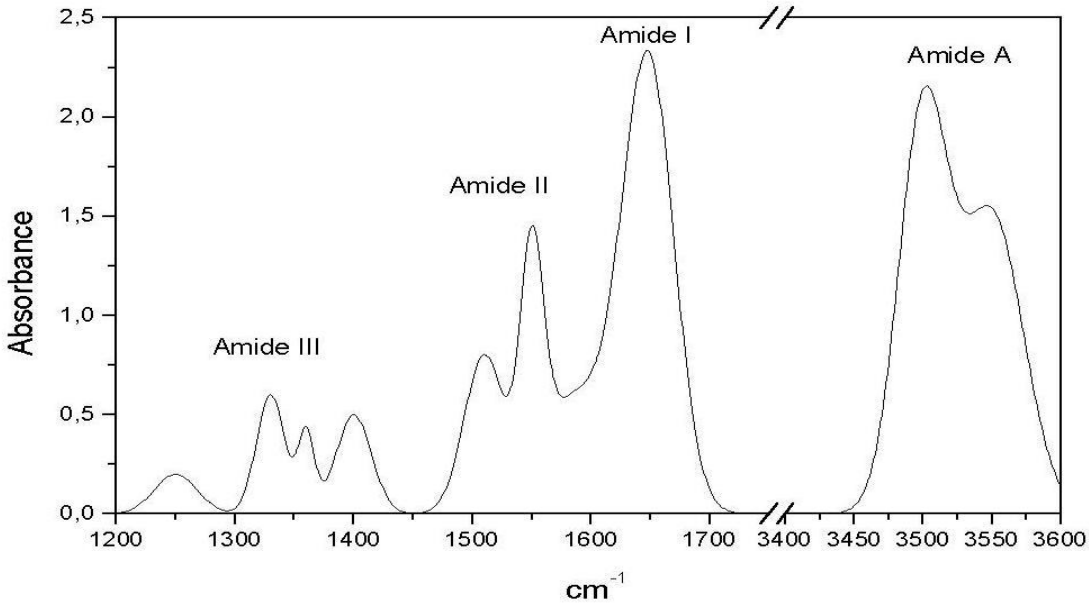


It is necessary to increase the sensitivity somehow, because the absorption due to one monolayer of molecules typically results in a change in intensity of only about one part in 10^5. For semiconductors, one way of increasing the sensitivity is to use multiple internal reflection. In this technique, the edges of the sample are polished, and the light is sent in at an angle. The light bounces around inside the sample, making about 30-50 bounces. This increases the sensitivity by about a factor of 30-50, making it possible to measure the absorption of less than one monolayer of molecules on a surface.

Band assignments

Amide vibrations

The peptide group, the structural repeat unit of proteins, gives up to 9 characteristic bands named amide A, B, I, II ... VII. The amide A band (about 3500 cm^{-1}) and amide B (about 3100 cm^{-1}) originate from a Fermi resonance between the first overtone of amide II and and the N-H stretching vibration. Amide I and amide II bands are two major bands of the protein infrared spectrum. The amide I band (between 1600 and 1700 cm^{-1}) is mainly associated with the C=O stretching vibration (70-85%) and is directly related to the backbone conformation. Amide II results from the N-H bending vibration (40-60%) and from the C-N stretching vibration (18-40%). This band is conformationally sensitive. Amide III and IV are very complex bands resulting from a mixture of several coordinate displacements. The out-of-plane motions are found in amide V, VI andVIII.



Amide A is with more than 95% due to the the N-H stretching vibration. This mode of vibration does not depend on the backbone conformation but is very sensitive to the strength of a hydrogen bond. It has wavenumbers between 3225 and 3280 cm^{-1} for hydrogen bond lengths between 2.69 to 2.85 Å, ***Amide I*** is the most intense absorption band in proteins. It is primilarygoverend by the stretching vibrations of the C=O (70-85%) and C-N groups (10-20%). Its frequency is found in the range between 1600 and 1700 cm^{-1}. The exact band position is determined by the
backbone conformation and the hydrogen bonding pattern. ***Amide II*** is found in the 1510 and 1580 cm^{-1} region and it is more complex than amide I. Amide II derives mainly from in-plane N-H bending (40-60% of the potential energy). The rest of the potential energy arises from the C-N (18-40%) and the C-C (about 10%) stretching vibrations.

Amide III, V are very complex bands dependent on the details of the force field, the nature of side chains and hydrogen bonding. Therefore these bands are only of limited use for the extraction of structural information.

Amino acid side chain vibrations

The presence of bands arising from amino acid side chains must be recognized before attempting to extract structural information from the shapes of amide I and amide II bands. The contribution of the side chain vibrations in the region between 1800 and 1400 cm^{-1} (amide I and amide II region) has been thor. Among the 20 proteinogenous amino acids only 9 (Asp, Asn, Glu, Gln, Lys, Arg, Tyr, Phe, His) show a significant absorbance in the region discussed above. The contribution of the different amino acid side chains were fitted by a sum of Gaussian and Lorentzian components.

AS	vibration		cm^{-1}	A_0 (l/mol/cm)	FWHH (cm^{-1})	surface ($x10^{-4}$ l/mol/cm)
Asp	-COO st as	pH>pK (~4.5)	1574	380	44	5.5
	-COOH st	pH<pK (~4.5)	1716	280	50	4.1
Glu	-COO st as	pH>pK (~4.4)	1560	470	48	7.1
	-COOH st	pH<pK (~4.4)	1712	220	56	3.6
Arg	$-CN_3H_5^+$ st as		1673	420	40	4.3
	st s		1633	300	40	3.6
Lys	$-NH_3^+$ bd as		1629	130	46	1.8
	bd s		1526	100	48	1.3
Asn	-C=O st		1678	310	32	2.7
	$-NH_2$ bd		1622	160	44	2.5
Gln	-C=O st		1670	360	32	3.1
	$-NH_2$ bd		1610	220	44	3.5
Tyr	ring-OH	pH<pK (~10)	1518	430	8	1.0
	ring-O	pH>pK (~10)	1602	160	14	0.7
			1498	700	10	2.5
His	ring		1596	70	14	0.3
Phe	ring		1494	80	6	0.2
terminal						
	-COO st as		1598	240	47	3.5
	-COOH st		1740	170	50	2.1

	$-NH_3^+$ bd as		1631	210	54	3.8

	bd s		1515	200	60	4.3
	-NH_2 bd		1560	450	46	7.5

frequency, absorbance at the maximum (Ao), full width at half height (FWHH), surface of Gaussian band
st=stretching vibration
bd=bending
s=symetrical
as=asymetrical

Secondary structure of peptide model compounds

A large number of synthetic polypeptides has been used for the characterization of infrared spectra for proteins with a defined secondary structure content. For example, polylysine may adopt both beta-sheet or alpha-helical structures in dependence on temperature and pH of the solution. Experimental and theoretical work on a large number of synthetic polypeptides has provided insights into the variability of the frequencies for particular secondary structure conformations

Beta sheet structures (beta strand)

The frequencies of the main absorption bands from synthetic polypeptides adopting an antiparallel chain structure have been compiled by Chirgadze&Nevskaya . From these data it this follows, that the amide I absorption is primarily determined by the backbone conformation and independent of the amino acid sequence, its hydrophilic or hydrophobic properties and charge. The average frequency of the main component is about 1629 cm^{-1} with
a minimum of 1615 cm^{-1} and a maximum of 1637 cm^{-1}. The average value for the second frequency is 1696 cm^{-1} (lowest value 1685 cm^{-1}). The parallel beta sheet structure that is not common in synthetic polypeptides leads to an amide I abosrption near 1640 cm^{-1}

Helical structures

The alpha-helix*: *For alpha-helical structures the mean frequency was found to be 1652 cm-1 for the amide I and 1548 cm^{-1} for the amid II absorptions. The half width of the alpha-helix band depends on the stability of the helix. For the most stable helices, the half-width of about 15 cm^{-1} corresponds to a helix-coil transition free energy of more than 300 cal/mole. Other helices display half-widths of 38 cm^{-1} and helix-coil transition free energies of about 90 cal/mole.

The 3_{10}-helix differs from the alpha-helix in that the internal hydrogen bonding occurs between residues i and i+3 instead of i and i+4 in alpha helices.

Turn structures

The beta turn structure involves 4 amino acid residues which form a loop so that the two chain segments separated by the turn adopt an antiparallel orientation and form an i to i+3

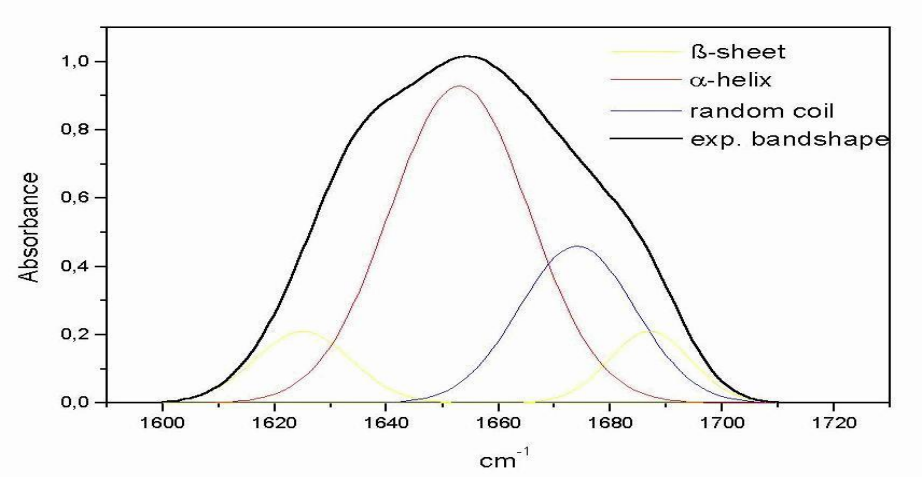
hydrogen bond. A number of turn structures have been identified from protein structures: type I (42%, non-helical), type II (15%, non-helical, requires Gly in position 3) and type III (18%, corresponds to one turn of 3_{10} helix). Assignment of beta turns by means of a normal mode analysis for insulin demonstrates a strong overlapping of the different types of beta turns with the alpha-helical absorption. However, an absorption near 1680 cm^{-1} is now clearly assigned to beta turns.

Secondary structure in proteins

The shape of the amide I band of globular proteins is characteristic of their secondary structure. With a publication by Byler & Susi The determination of secondary structures in proteins from FTIR spectra actually started. This had become possible by the availability of high signal-to-noise ratio digitalised spectra obtained by the FTIR spectrometer and by the access to computers and software able to perform many operations on the spectra in a short time.

Deconvolution of the amide I band

The concept of Fourier self deconvolution is based on the assumption, that a spectrum of single bands (each narrow band is characteristic for a secondary structure) is broadened in the liquid or solid state. Therefore, the bands overlap and can not be distinguished in the amide envelope. A curve fitting procedure can be applied to estimate quantitatively the area of each component representing a type of secondary structure. In the pioneering work by Susi & Byler the amide I was deconvoluted with a Lorentzian line shape function and a resolution enhancement factor of 2.4 was applied. The deconvoluted spectrum was fitted with Gaussian band shapes by an iterative curve fitting procedure. The results are in good agreement with with the secondary structure information obtained from X-ray crystallographic structures of the proteins under study.



	a)			***b)***			
sec. structure	**Mean (cm^{-1})**	**RMS (cm^{-1})**	**Max (cm^{-1})**	**Mean (cm^{-1})**	**RMS (cm^{-1})**	**Max (cm^{-1})**	**Region (cm^{-1})**
turns	1694	1.7	2	-	-	-	
	1688	1.1	2	-	-	-	
	1683	1.5	2	1678	2.1	5	**1682-1662**
	1670	1.4	2	1670	2.9	5	
	1663	2.2	4	1664	1.0	3	
alpha-helix	1654	1.5	3	1656	1.5	3	
				1648	1.6	3	**1662-1645**
unordered	1645	1.6	4	1641	2.0	3	**1645-1637**
beta sheet	1624	2.4	4	1624	2.5	5	
	1631	2.5	3	1633	2.1	4	**1637-1613**
	1637	1.4	3	-	-	-	
	1675	2.6	4	1685	2.1	4	**1689-1682**

MASS SPECTROMETRY IN PROTEIN IDENTIFICATION AND SEQUENCE ANALYSIS

Mass spectrometry is an analytical tool used for measuring the **molecular mass** of a sample.

For large samples such as **biomolecules,** molecular masses can be measured to within an accuracy of **0.01%** of the total molecular mass of the sample *i.e.* within a 4 Daltons (Da) or atomic mass units (amu) error for a sample of 40,000 Da. This is sufficient to allow minor mass changes to be detected, *e.g.* the substitution of one amino acid for another, or a post-translational modification.

For small **organic molecules** the molecular mass can be measured to within an accuracy of **5 ppm** or less, which is often sufficient to confirm the molecular formula of a compound, and is also a standard requirement for publication in a chemical journal.

Structural information can be generated using certain types of mass spectrometers, usually those with multiple analysers which are known as **tandem mass spectrometers.** This is achieved by fragmenting the sample inside the instrument and analysing the products generated. This procedure is useful for the structural elucidation of **organic compounds** and for **peptide** or **oligonucleotide** sequencing.

Where are mass spectrometers used?

Mass spectrometers are used in industry and academia for both routine and research purposes. The following list is just a brief summary of the major mass spectrometric applications:

- **Biotechnology:** *the analysis of proteins, peptides, oligonucleotides*
- **Pharmaceutical:** *drug discovery, combinatorial chemistry, pharmacokinetics, drug metabolism*
- **Clinical:** *neonatal screening, haemoglobin analysis, drug testing*
- **Environmental:** *PAHs, PCBs, water quality, food contamination*
- **Geological:** *oil composition*

3. How can mass spectrometry help biochemists?

- **Accurate molecular weight measurements:**
  sample confirmation, to determine the purity of a sample, to verify amino acid substitutions, to detect post-translational modifications, to calculate the number of disulphide bridges
- **Reaction monitoring:**
  to monitor enzyme reactions, chemical modification, protein digestion
- **Amino acid sequencing:**
  sequence confirmation, de novo *characterisation of peptides, identification of proteins by database searching with a sequence "tag" from a proteolytic fragment*
- **Oligonucleotide sequencing:**
  the characterisation or quality control of oligonucleotides
- **Protein structure:**
  protein folding monitored by H/D exchange, protein-ligand complex formation under physiological conditions, macromolecular structure determination

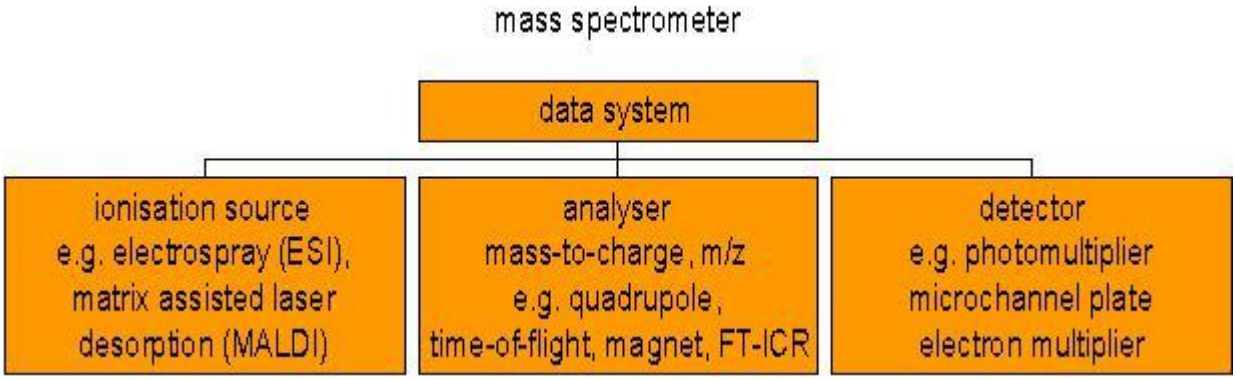
How does a mass spectrometer work?

Introduction

Mass spectrometers can be divided into three fundamental parts, namely the **ionisation source** , the **analyser** , and the **detector.**

The sample has to be introduced into the ionisation source of the instrument. Once inside the ionisation source, the sample molecules are ionised, because ions are easier to manipulate than neutral molecules. These ions are extracted into the analyser region of the mass spectrometer where they are separated according to their **mass (m) -to-charge (z) ratios (m/z)** . The separated ions are detected and this signal sent to a data system where the m/z ratios are stored together with their relative abundance for presentation in the format of a **m/z spectrum** .

The analyser and detector of the mass spectrometer, and often the ionisation source too, are maintained under high vacuum to give the ions a reasonable chance of travelling from one end of the instrument to the other without any hindrance from air molecules. The entire operation of the mass spectrometer, and often the sample introduction process also, is under complete **data system** control on modern mass spectrometers.



Simplified schematic of a mass spectrometer

Sample introduction

The method of sample introduction to the ionisation source often depends on the ionisation method being used, as well as the type and complexity of the sample.

The sample can be inserted directly into the ionisation source, or can undergo some type of chromatography *en route* to the ionisation source. This latter method of sample introduction usually involves the mass spectrometer being coupled directly to a high pressure liquid chromatography (HPLC), gas chromatography (GC) or capillary electrophoresis (CE) separation column, and hence the sample is separated into a series of components which then enter the mass spectrometer sequentially for individual analysis.

Methods of sample ionisation

Many ionisation methods are available and each has its own advantages and disadvantages (**"Ionization Methods in Organic Mass Spectrometry",** Alison E. Ashcroft, The Royal Society of Chemistry, UK, 1997; and references cited therein).

The ionisation method to be used should depend on the type of sample under investigation and the mass spectrometer available.

Ionisation methods include the following:
Atmospheric Pressure Chemical Ionisation (APCI)
Chemical Ionisation (CI)
Electron Impact (EI)
Electrospray Ionisation (ESI)
Fast Atom Bombardment (FAB)
Field Desorption / Field Ionisation (FD/FI)
Matrix Assisted Laser Desorption Ionisation (MALDI)
Thermospray Ionisation (TSP)

The ionisation methods used for the majority of biochemical analyses are **Electrospray Ionisation (ESI)** and **Matrix Assisted Laser Desorption Ionisation (MALDI)** , and these are described in more detail in Sections 5 and 6 respectively.

With most ionisation methods there is the possibility of creating both positively and negatively charged sample ions, depending on the proton affinity of the sample. Before

embarking on an analysis, the user must decide whether to detect the positively or negatively charged ions (see section 7).

Analysis and Separation of Sample Ions
The main function of the **mass analyser** is to **separate** , or **resolve** , the ions formed in the ionisation source of the mass spectrometer according to their **mass-to-charge (m/z)** ratios. There are a number of mass analysers currently available, the better known of which include **quadrupoles**, **time-of-flight (TOF)** analysers, **magnetic sectors** , and both **Fourier transform** and **quadrupole ion traps** .

These mass analysers have different features, including the m/z range that can be covered, the mass accuracy, and the achievable resolution. The compatibility of different analysers with different ionisation methods varies. For example, all of the analysers listed above can be used in conjunction with electrospray ionisation, whereas MALDI is not usually coupled to a quadrupole analyser.

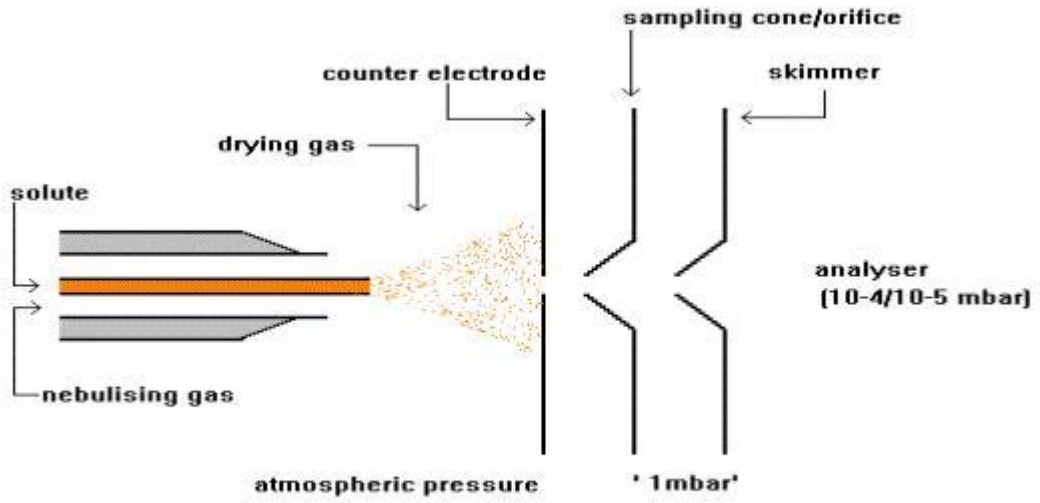
Tandem (MS-MS) mass spectrometers are instruments that have more than one analyser and so can be used for structural and sequencing studies. Two, three and four analysers have all been incorporated into commercially available tandem instruments, and the analysers do not necessarily have to be of the same type, in which case the instrument is a **hybrid** one. More popular tandem mass spectrometers include those of the **quadrupole-quadrupole, magnetic sector-quadrupole**, and more recently, the **quadrupole-time-of-flight** geometries.

Detection and recording of sample ions.
The **detector** monitors the ion current, amplifies it and the signal is then transmitted to the data system where it is recorded in the form of **mass spectra** . The **m/z** values of the ions are plotted against their **intensities** to show the **number of components** in the sample, the **molecular mass** of each component, and the **relative abundance** of the various components in the sample.

The type of detector is supplied to suit the type of analyser; the more common ones are the **photomultiplier** , the **electron multiplier** and the **micro-channel plate** detectors.

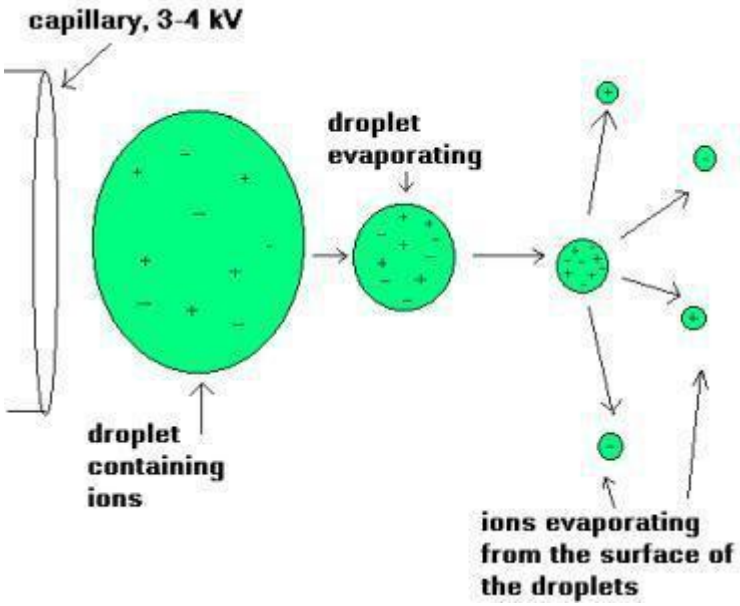
Electrospray ionisation

Electrospray ionisation
Electrospray Ionisation (ESI) is one of the **Atmospheric Pressure Ionisation (API)** techniques and is well-suited to the analysis of polar molecules ranging from less than 100 Da to more than 1,000,000 Da in molecular mass.



Standard electrospray ionisation source (Platform II)

During standard electrospray ionisation (J. Fenn, J. Phys. Chem., 1984, 88, 4451), the sample is dissolved in a polar, volatile solvent and pumped through a narrow, **stainless steel capillary** (75 - 150 micrometersi.d.) at a flow rate of between 1 ﷺL/min and 1 mL/min. A **high voltage** of 3 or 4 kV is applied to the tip of the capillary, which is situated within the ionisation source of the mass spectrometer, and as a consequence of this strong electric field, the sample emerging from the tip is dispersed into an **aerosol of highly charged droplets**, a process that is aided by a co-axially introduced **nebulising gas** flowing around the outside of the capillary. This gas, usually nitrogen, helps to direct the spray emerging from the capillary tip towards the mass spectrometer. The charged droplets diminish in size by **solvent evaporation**, assisted by a warm flow of nitrogen known as the **drying gas** which passes across the front of the ionisation source. Eventually charged **sample ions**, free from solvent, are released from the droplets, some of which pass through a **sampling cone** or orifice into an **intermediate vacuum** region, and from there through a small aperture into the analyser of the mass spectrometer, which is held under **high vacuum**. The lens voltages are optimised individually for each sample.



The electrospray ionisation process

Nanospray ionisation

Nanospray ionisation (M. Wilm, M. Mann, Anal. Chem., 1996, 68, 1) is a **low flow rate** version of electrospray ionisation. A **small volume** (1-4 microL) of the sample dissolved in a suitable volatile solvent, at a concentration of ca. **1 - 10 pmol/microL**, is transferred into a miniature **sample vial**. A reasonably **high voltage** (ca. 700 - 2000 V) is applied to the specially manufactured gold-plated vial resulting in **sample ionisation** and spraying. The flow rate of solute and solvent using this procedure is very low, **30 - 1000 nL/min**, and so not only is far less sample consumed than with the standard electrospray ionisation technique, but also a small volume of sample lasts for several minutes, thus enabling **multiple experiments** to be performed. A common application of this technique is for a **protein digest** mixture to be analysed to generate a list of **molecular masses** for the components present, and then each component to be analysed further by **tandem mass spectrometric (MS-MS) amino acid sequencing** techniques (see Section 8).

ESI and **nanospray ionisation** are very sensitive analytical techniques but the sensitivity deteriorates with the presence of non-volatile buffers and other additives, which should be avoided as far as possible.

In **positive ionisation** mode, a trace of formic acid is often added to aid protonation of the sample molecules; in **negative ionisation** mode a trace of ammonia solution or a volatile amine is added to aid deprotonation of the sample molecules. **Proteins and peptides** are usually analysed under **positive ionisation** conditions and **saccharides and oligonucleotides** under **negative ionisation** conditions. In all cases, the **m/z** scale must be **calibrated** by analysing a standard sample of a similar type to the sample being analysed (e.g. a protein calibrant for a protein sample), and then applying a mass correction.

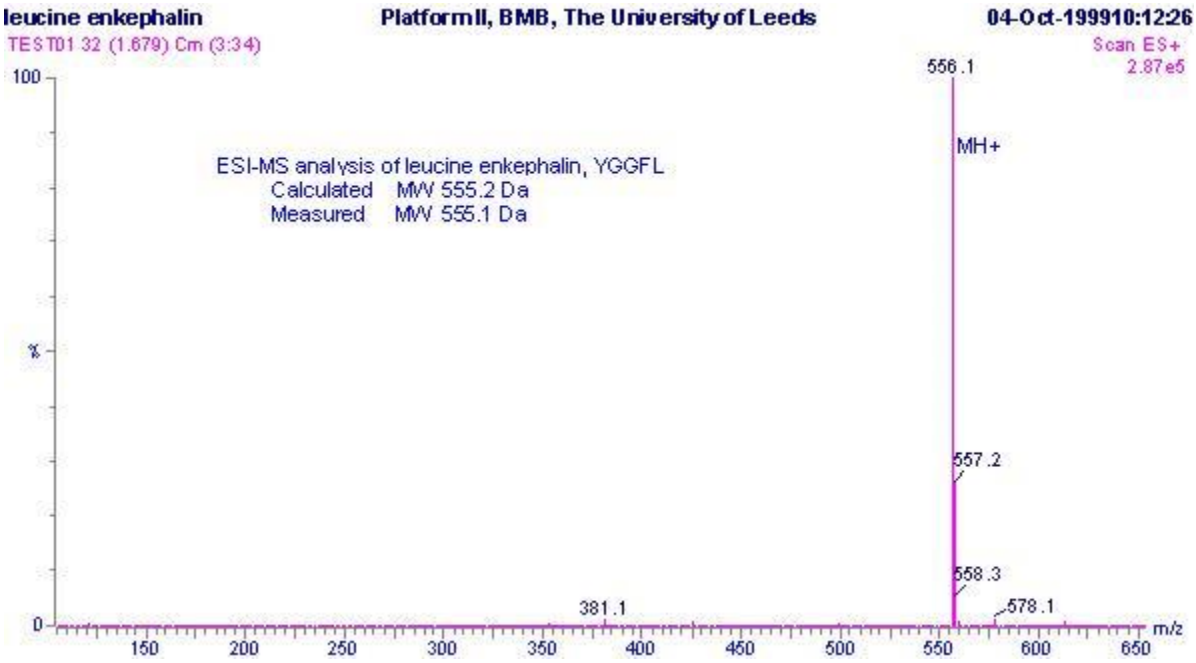
Data processing

ESI and **nanospray ionisation** generate the same type of spectral data for samples, and so the data processing procedures are identical.

In ESI, samples (M) with **molecular masses up to ca. 1200 Da** give rise to **singly charged molecular-related ions**, usually **protonated molecular ions** of the formula $(M+H)^+$ in **positive ionisation mode**, and **deprotonated molecular ions** of the formula $(M-H)^-$ in **negative ionisation** mode.

An example of this type of sample analysis is shown in the m/z spectrum of the pentapeptideleucineenkephalin, YGGFL. The molecular formula for this compound is $C_{28}H_{37}N_5O_7$ and the calculated monoisotopic molecular weight is 555.2692 Da.

The m/z spectrum shows dominant ions at m/z 556.1, which are consistent with the expected protonated molecular ions, $(M+H^+)$. Protonated molecular ions are expected because the sample was analysed under positive ionisation conditions. These m/z ions are **singly charged**, and so the m/z value is consistent with the molecular mass, as the value of z (number of charges) equals 1. Hence the measured molecular weight is deduced to be 555.1 Da, in good agreement with the theoretical value.



Positive ESI-MS m/z spectrum of leucineenkaphalin, YGGFL.

The m/z spectrum also shows other ions of lower intensity (ca. 25 % of the m/z 556.1 ions) at m/z 557.2. These represent the molecule in which one ^{12}C atom has been replaced by a ^{13}C atom, because carbon has a naturally occurring isotope one atomic mass unit (Da) higher. The intensity of these isotopic ions relates to the relative abundance of the naturally occurring isotope multiplied by the total number of carbon atoms in the molecule. Additionally the fact that the ^{13}C ions are one Da higher on the m/z scale than the ^{12}C ions is an indication that z = 1, and hence the sample ions are singly charged. If the sample ions had been doubly charged, then the m/z values would only differ by 0.5 Da as z, the number of charges, would then be equal to 2.

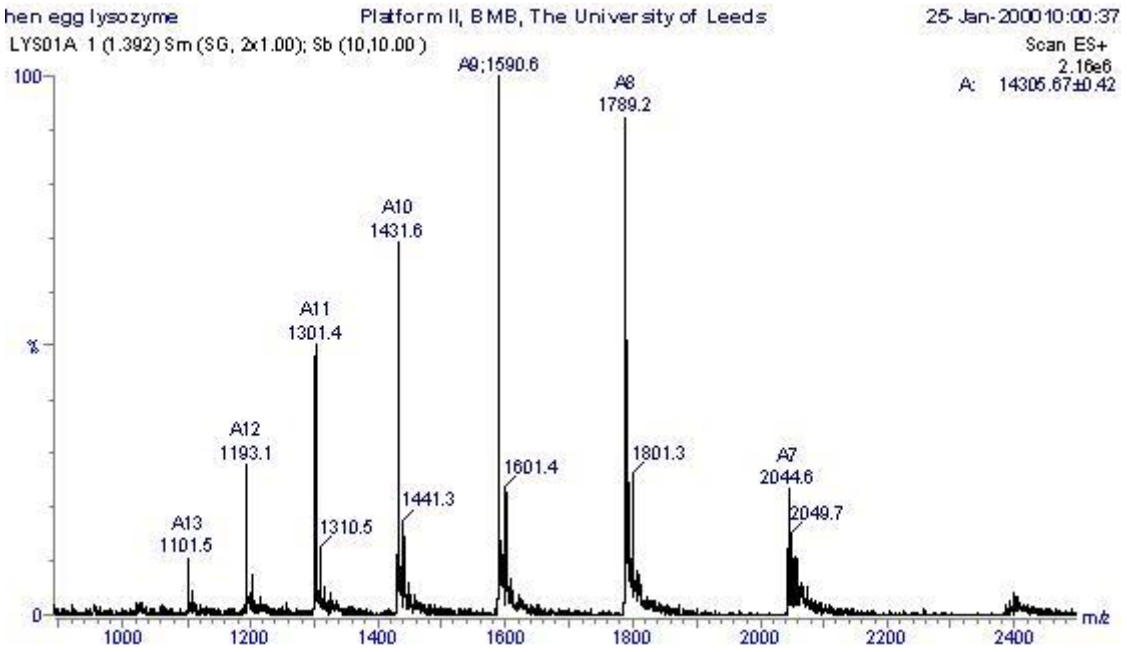
The m/z spectrum also contains ions at m/z 578.1, some 23 Da higher than the expected molecular mass. These can be identified as the sodium adduct ions, $(M+Na)^+$, and are quite common in electrospray ionisation. Instead of the sample molecules being ionised by the addition of a proton H^+, some molecules have been ionised by the addition of a sodium

cationNa^+. Other common adduct ions include K^+ (+39) and NH_4^+ (+18) in positive ionisation mode and Cl^- (+35) in negative ionisation mode.

Electrospray ionisation is known as a "soft" ionisation method as the sample is ionised by the addition or removal of a proton, with very little extra energy remaining to cause fragmentation of the sample ions.

Samples (M) with molecular weights greater than ca. 1200 Da give rise to multiply charged molecular-related ions such as **$(M+nH)^{n+}$ in positive ionisation mode** and **$(M-nH)^{n-}$ in negative ionisation mode**. Proteins have many suitable sites for protonation as all of the backbone amide nitrogen atoms could be protonated theoretically, as well as certain amino acid side chains such as lysine and arginine which contain primary amine functionalities.

An example of multiple charging, which is practically unique to electrospray ionisation, is presented in the positive ionisation m/z spectrum of the protein hen egg white lysozyme.



Positive ESI-MS m/z spectrum of the protien hen egg white lysozyme.

The sample was analysed in a solution of 1:1 (*v/v*) acetonitrile : 0.1% aqueous formic acid and the m/z spectrum shows a Gaussian-type distribution of multiply charged ions ranging from m/z 1101.5 to 2044.6. Each peak represents the intact protein molecule carrying a different number of charges (protons). The peak width is greater than that of the singly charged ions seen in the leucineenkephalin spectrum, as the isotopes associated with these multiply charged ions are not clearly resolved as they were in the case of the singly charged ions. The individual peaks in the multiply charged series become closer together at lower m/z values and, because the molecular weight is the same for all of the peaks, those with more charges appear at lower m/z values than do those with fewer charges (M. Mann, C. K. Meng, J. B. Fenn, *Anal. Chem.*, 1989, **61**, 1702).

The m/z values can be expressed as follows:

$m/z = (MW + nH^+)/n$

where m/z = the mass-to-charge ratio marked on the abscissa of the spectrum;
MW = the molecular mass of the sample
n = the integer number of charges on the ions
H = the mass of a proton = 1.008 Da.

If the number of charges on an ion is known, then it is simply a matter of reading the m/z value from the spectrum and solving the above equation to determine the molecular weight of the sample. Usually the number of charges is not known, but can be calculated if the assumption is made that any two adjacent members in the series of multiply charged ions differ by one charge.

For example, if the ions appearing at m/z 1431.6 in the lysozyme spectrum have "n" charges, then the ions at m/z 1301.4 will have "n+1" charges, and the above equation can be written again for these two ions:

$$1431.6 = (MW + nH^+)/n \text{ and } 1301.4 = [MW + (n+1)H^+] /(n+1)$$

These simultaneous equations can be rearranged to exclude the MW term:

$$n(1431.6) - nH^+ = (n+1)1301.4 - (n+1)H^+$$
and so:
$$n(1431.6) = n(1301.4) +1301.4 - H^+$$
therefore:
$$n(1431.6 - 1301.4) = 1301.4 - H^+$$
and so:
$$n = (1301.4 - H^+) / (1431.6 - 1301.4)$$

hence the number of charges on the ions at m/z 1431.6 = 1300.4/130.2 = 10.

Putting the value of n back into the equation:

$$1431.6 = (MW + nH^+) n$$
gives $1431.6 \times 10 = MW + (10 \times 1.008)$
and so MW = 14,316 - 10.08
therefore **MW = 14,305.9 Da**

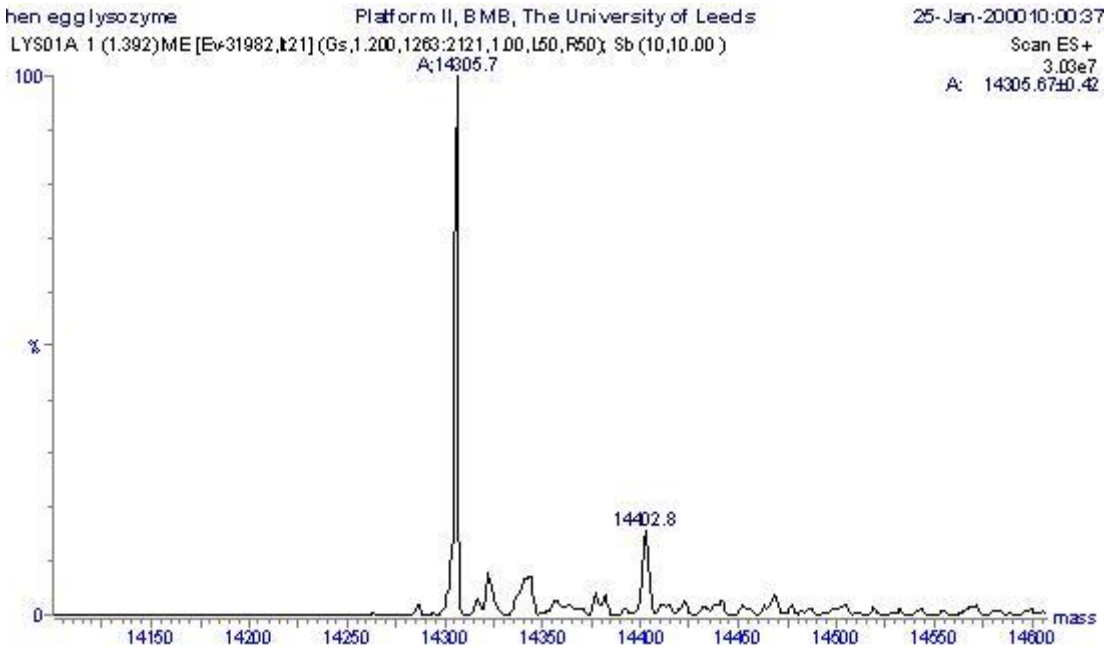
The observed molecular mass is in good agreement with the theoretical molecular mass of hen egg lysozyme (based on average atomic masses) of 14305.14 Da. The individual isotopes cannot be resolved when the ions have a large number of charges, and so for proteins the average mass is measured.

This may seem long-winded but fortunately the molecular mass of the sample can be calculated automatically, or at least semi-automatically, by the processing software associated with the mass spectrometer. This is of great help for multi-component mixture analysis where the m/z spectrum may well contain several overlapping series of multiply charged ions, with each component exhibiting completely different charge states.

Using **electrospray** or **nanospray ionisation,** a **mass accuracy of within 0.01%** of the molecular mass should be achievable, which in this case represents +/- 1.4 Da.

In order to clarify electrospray/nanospray data, **molecular mass profiles** can be generated from the m/z spectra of high molecular mass, multiply charged samples. To achieve this, all the components are transposed onto a true molecular mass (or **zero charge state**) profile from which molecular masses can be read directly without any amendments or calculations.

The m/z spectrum of lysozyme has been converted to a molecular mass profile using Maximum Entropy processing and the data are shown. The mass profile is dominated by a component of molecular mass 14,305.7 Da, with a series of minor peaks at higher mass, which is usually indicative of salt adducting e.g. Na (M+23), K (M+39), H_2SO_4 or H_3PO_4 (M+98). The molecular masses can be read easily and unambiguously, and a good idea of the purity of the protein is obtained on inspection of the molecular mass profile.



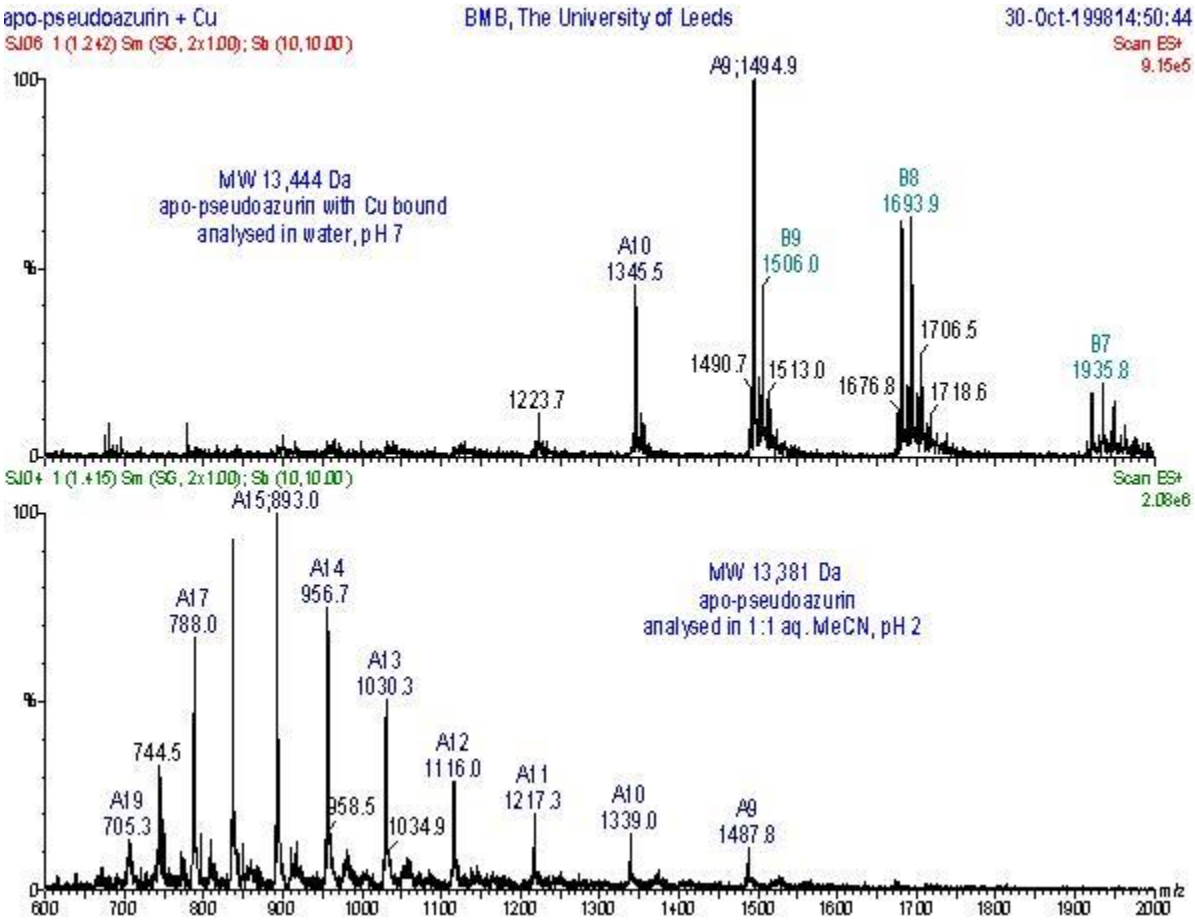
Molecular mass profile of lysozyme obtained by maximum entropy processing of the m/z spectrum

Proteins in their **native state**, or at least containing a significant amount of folding, tend to produce multiply charged ions covering a smaller range of charge states (say two or three). These charge states tend to have fewer charges than an unfolded protein would have, due to the inaccessibility of many of the protonation sites. In such cases, increasing the **sampling cone voltage** may provide sufficient energy for the protein to begin to unfold and create a wider charge state distribution centering on more highly charged ions in the lower m/z region of the spectrum.

The differences in m/z spectra due to the folded state of the protein are illustrated with the m/z spectra of the protein apo-pseudoazurin acquired under different solvent conditions.

Analysis of the protein in 1:1 acetonitrile : 0.1% aqueous formic acid at pH2 gave a Gaussian-type distribution with multiply charged states ranging from n = 9 at m/z 1487.8 to n = 19 at m/z 705.3, centering on n = 15 (lower trace). The molecular mass for this protein was 13,381 Da. Analysis of the protein in water gave fewer charge states, from n = 7 at m/z 1921.7 to n = 11 at m/z 1223.7, centering at n = 9 (upper trace). Not only has the charge state distribution changed, the molecular weight is now 13,444 Da which represents an increase of

63 Da and indicates that copper is remaining bound to the protein. Many types of **protein complexes** can be observed in this way, including **protein-ligand, protein-peptide, protein-metal and protein-RNA macromolecules**.



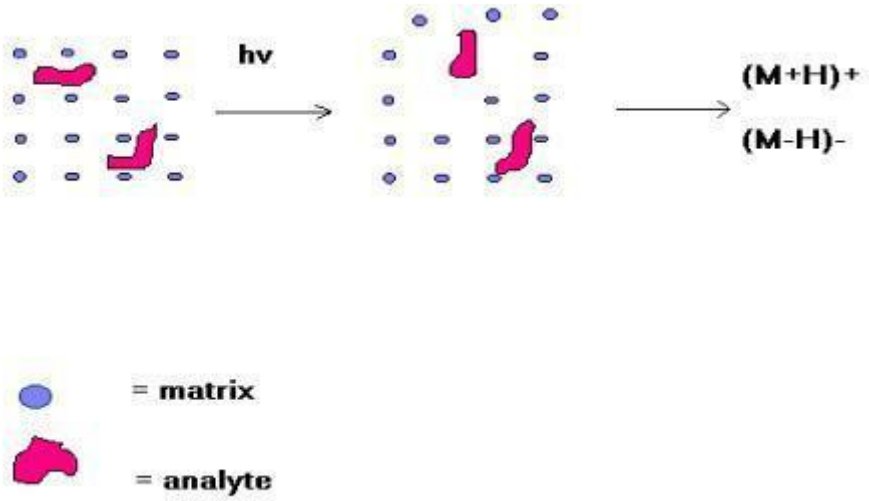
Positive ESI-MS m/z spectra of the protein apo-pseudoazurin analysed in water at pH7 (upper trace) and in 1:1 acetonitrile:0.1% aq. formic acid at pH2 (lower trace).

Matrix assisted laser desorption ionisation

Matrix Assisted Laser Desorption Ionisation (MALDI) (F. Hillenkamp, M. Karas, R. C. Beavis, B. T. Chait, *Anal. Chem*., 1991, **63**, 1193) deals well with thermolabile, non-volatile organic compounds especially those of high molecular mass and is used successfully in biochemical areas for the analysis of **proteins, peptides, glycoproteins, oligosaccharides, and oligonucleotides**. It is relatively straightforward to use and reasonably tolerant to buffers and other additives. The mass accuracy depends on the type and performance of the analyser of the mass spectrometer, but most modern instruments should be capable of measuring masses to within 0.01% of the molecular mass of the sample, at least up to ca. 40,000 Da.

MALDI is based on the **bombardment** of sample molecules with a **laser** light to bring about **sample ionisation**. The sample is pre-mixed with a highly absorbing **matrix** compound for the most consistent and reliable results, and a low concentration of sample to matrix works best. The matrix transforms the laser energy into **excitation energy** for the sample, which leads to sputtering of analyte and matrix ions from the surface of the mixture. In this way energy transfer is efficient and also the analyte molecules are spared excessive direct energy

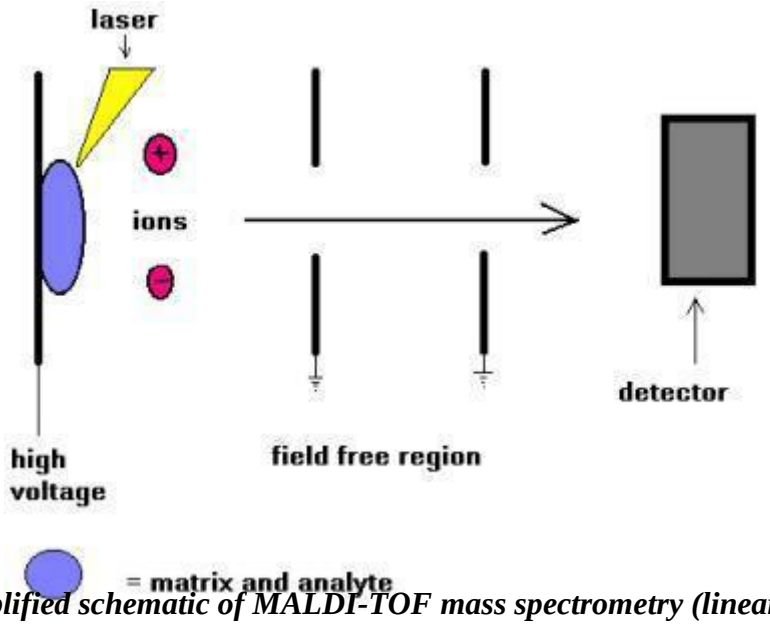
that may otherwise cause decomposition. Most commercially available MALDI mass spectrometers now have a pulsed nitrogen laser of wavelength 337 nm.



Matrix assisted laser desorption ionisation (MALDI)

The sample to be analysed is dissolved in an appropriate volatile solvent, usually with a trace of trifluoroacetic acid if positive ionisation is being used, at a concentration of ca. **10 pmol/**L and an aliquot (**1-2** L) of this removed and mixed with an equal volume of a solution containing a vast excess of a matrix. A range of compounds is suitable for use as matrices: **sinapinic acid** is a common one for **protein** analysis while **alpha-cyano-4-hydroxycinnamic acid** is often used for **peptide** analysis. An aliquot (1-2 L) of the final solution is applied to the sample target which is allowed to dry prior to insertion into the high vacuum of the mass spectrometer. The laser is fired, the energy arriving at the sample/matrix surface optimised, and data accumulated until a m/z spectrum of reasonable intensity has been amassed. The time-of-flight analyser separates ions according to their**mass(m)-to-charge(z) (m/z)** ratios by measuring the time it takes for ions to travel through a field free region known as the flight, or drift, tube. The heavier ions are slower than the lighter ones.

The m/z scale of the mass spectrometer is **calibrated** with a known sample that can either be analysed independently (external calibration) or pre-mixed with the sample and matrix (internal calibration).

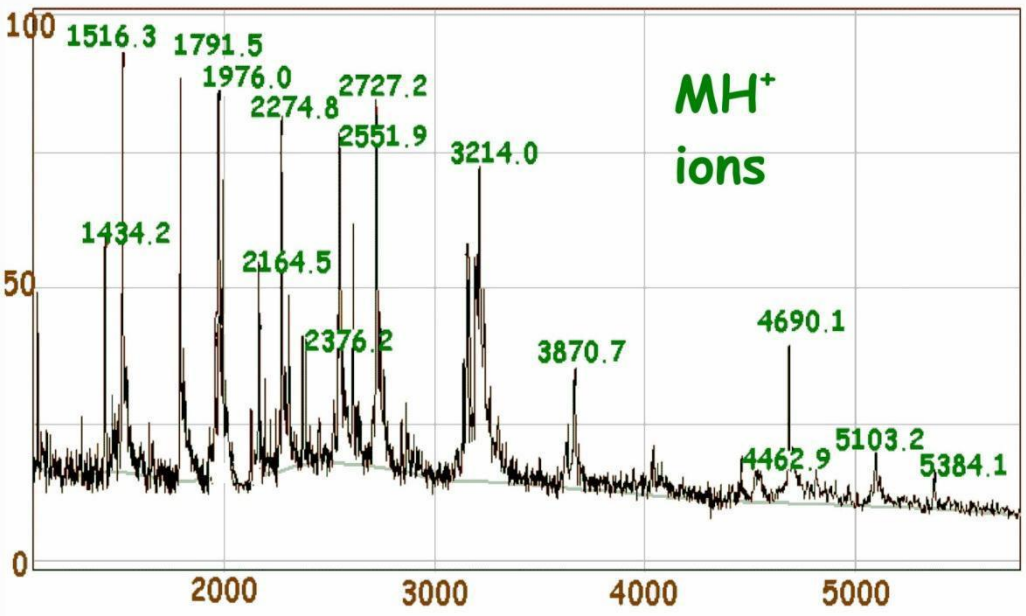


Simplified schematic of MALDI-TOF mass spectrometry (linear mode)

MALDI is also a "soft" ionisation method and so results predominantly in the generation of **singly charged molecular-related ions** regardless of the molecular mass, hence the spectra are relatively easy to interpret. Fragmentation of the sample ions does not usually occur.

In **positive ionisation** mode the **protonated molecular ions ($M+H^+$)** are usually the dominant species, although they can be accompanied by salt adducts, a trace of the doubly charged molecular ion at approximately half the m/z value, and/or a trace of a dimeric species at approximately twice the m/z value. Positive ionisation is used in general for **protein** and **peptide** analyses.

In **negative ionisation** mode the **deprotonated molecular ions ($M-H^-$)** are usually the most abundant species, accompanied by some salt adducts and possibly traces of dimeric or doubly charged materials. Negative ionisation can be used for the analysis of **oligonucleotides** and **oligosaccharides**.



Positive ionisation MALDI m/z spectrum of a peptide mixture using alpha-cyano-4-hydroxycinnamic acid as matrix

Positive or negative ionisation?

If the sample has functional groups that readily accept a proton (H^+) then positive ion detection is used

e.g. amines R-NH_2 + H^+ = R-NH_3^+ as in proteins or peptides.

If the sample has functional groups that readily lose a proton then negative ion detection is used

e.g. carboxylic acids R-CO_2H = R-CO_2^- and alcohols R-OH = R-O- as in saccharides or oligonucleotides

Tandem mass spectrometry (MS-MS): Structural and sequence information from mass spectrometry.

Tandem mass spectrometry

Tandem mass spectrometry (MS-MS) is used to produce **structural information** about a compound by fragmenting specific sample ions inside the mass spectrometer and identifying the resulting fragment ions. This information can then be pieced together to generate structural information regarding the intact molecule. Tandem mass spectrometry also enables specific compounds to be detected in complex mixtures on account of their specific and characteristic fragmentation patterns.

A **tandem mass spectrometer** is a mass spectrometer that has more than one analyser, in practice usually two. The two analysers are separated by a collision cell into which an inert gas (e.g. argon, xenon) is admitted to collide with the selected sample ions and bring about

their fragmentation. The analysers can be of the same or of different types, the most common combinations being:

- quadrupole - quadrupole
- magnetic sector - quadrupole
- magnetic sector - magnetic sector
- quadrupole - time-of-flight.

Fragmentation experiments can also be performed on certain single analyser mass spectrometers such as ion trap and time-of-flight instruments, the latter type using a post-source decay experiment to effect the fragmentation of sample ions.

Tandem mass spectrometry analyses.
The basic modes of data acquisition for tandem mass spectrometry experiments are as follows:

Product or daughter ion scanning:
the first analyser is used to select user-specified sample ions arising from a particular component; usually the molecular-related (i.e. $(M+H)^+$ or $(M-H)^-$) ions. These chosen ions pass into the collision cell, are bombarded by the gas molecules which cause fragment ions to be formed, and these fragment ions are analysed i.e. separated according to their mass to charge ratios, by the second analyser. All the fragment ions arise directly from the precursor ions specified in the experiment, and thus produce a fingerprint pattern specific to the compound under investigation.

This type of experiment is particularly useful for providing structural information concerning **small organic molecules** and for generating **peptide sequence** information.

Precursor or parent ion scanning:
the first analyser allows the transmission of all sample ions, whilst the second analyser is set to monitor specific fragment ions, which are generated by bombardment of the sample ions with the collision gas in the collision cell. This type of experiment is particularly useful for monitoring groups of compounds contained within a mixture which fragment to produce common fragment ions, e.g. **glycosylated peptides** in a tryptic digest mixture, **aliphatic hydrocarbons** in an oil sample, or **glucuronide conjugates** in urine.

Constant neutral loss scanning:
this involves both analysers scanning, or collecting data, across the whole m/z range, but the two are off-set so that the second analyser allows only those ions which differ by a certain number of mass units (equivalent to a neutral fragment) from the ions transmitted through the first analyser. e.g. This type of experiment could be used to monitor all of the carboxylic acids in a mixture. Carboxylic acids tend to fragment by losing a (neutral) molecule of carbon dioxide, CO_2, which is equivalent to a loss of 44 Da or atomic mass units. All ions pass through the first analyser into the collision cell. The ions detected from the collision cell are those from which 44 Da have been lost.

Selected/multiple reaction monitoring:
both of the analysers are static in this case as user-selected specific ions are transmitted through the first analyser and user-selected specific fragments arising from these ions are measured by the second analyser. The compound under scrutiny must be known and have

been well-characterised previously before this type of experiment is undertaken. This methodology is used to confirm unambiguously the presence of a compound in a matrix e.g. drug testing with blood or urine samples. It is not only a highly specific method but also has very high sensitivity.

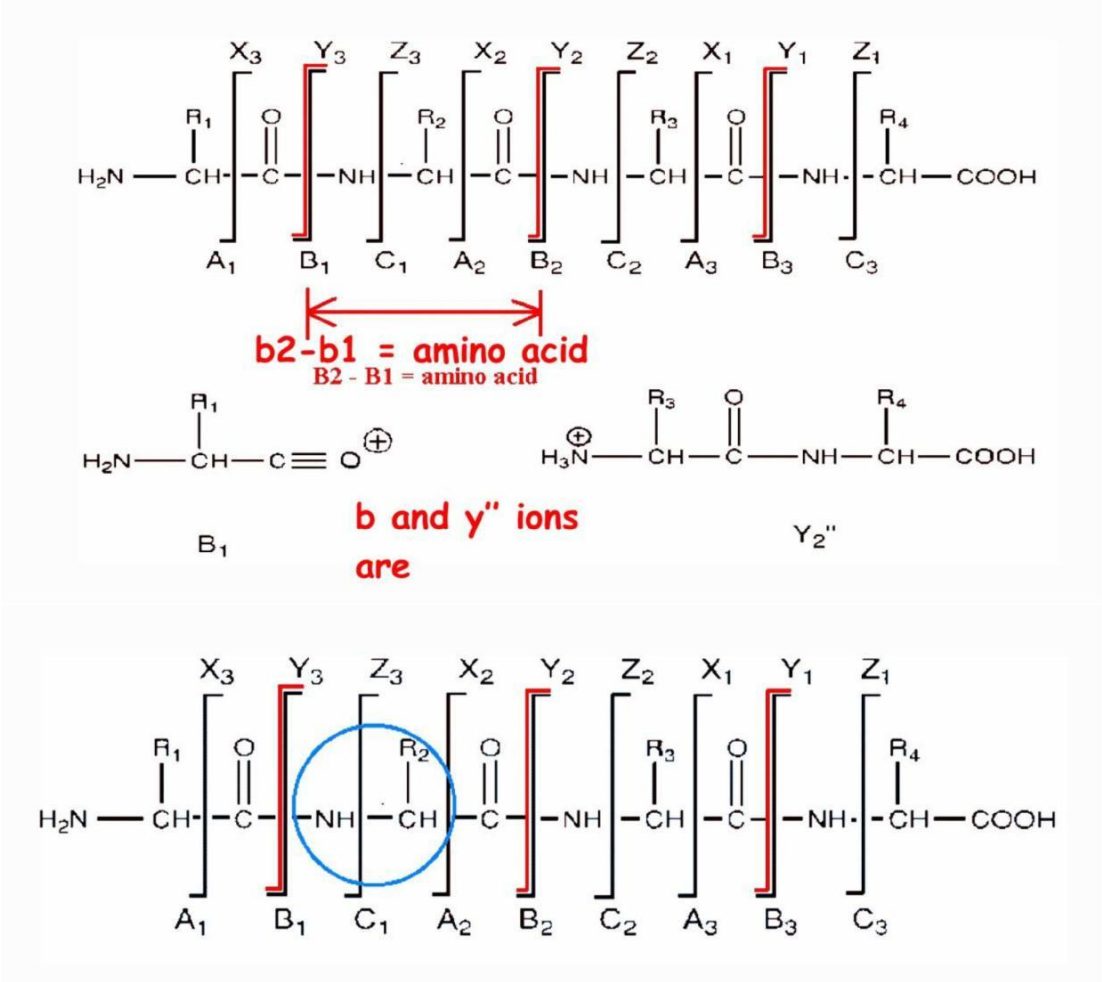
8.3 Peptide Sequencing by Tandem Mass Spectrometry.

The most common usage of MS-MS in biochemical areas is the **product or daughter ion scanning** experiment which is particularly successful for **peptide** and **nucleotide sequencing**.

Peptide sequencing: H_2N-CH(R')-CO-NH-CH(R")-CO_2H

Peptides fragment in a reasonably well-documented manner (P. Roepstorrf, J. Fohlmann, *Biomed. Mass Spectrom*., 1984, **11**, 601; R. S. Johnson, K. Biemann, *Biomed. Environ. Mass Spectrom*., 1989, **18**, 945). The protonated molecules fragment along the **peptide backbone** and also show some **side-chain fragmentation** with certain instruments (Four-Sector Tandem Mass Spectrometry of Peptides, A. E. Ashcroft, P. J. Derrick in "Mass Spectrometry of Peptides" ed. D. M. Desiderio, CRC Press, Florida, 1990).

There are three different types of bonds that can fragment along the amino acid backbone: the **NH-CH**, **CH-CO**, and **CO-NH** bonds. Each bond breakage gives rise to two species, one neutral and the other one charged, and only the charged species is monitored by the mass spectrometer. The charge can stay on either of the two fragments depending on the chemistry and relative proton affinity of the two species. Hence there are six possible fragment ions for each amino acid residue and these are labelled as in the diagram, with the **a, b, and c" ions** having the charge retained on the **N-terminal fragment**, and the **x, y", and z ions** having the charge retained on the **C-terminal fragment**. The most common cleavage sites are at the CO-NH bonds which give rise to the b and/or the y" ions. The mass difference between two adjacent b ions, or y"; ions, is indicative of a particular amino acid residue (see Table of amino acid residues at the end of this document).



Peptide sequencing by tandem mass spectrometry - backbone cleavages

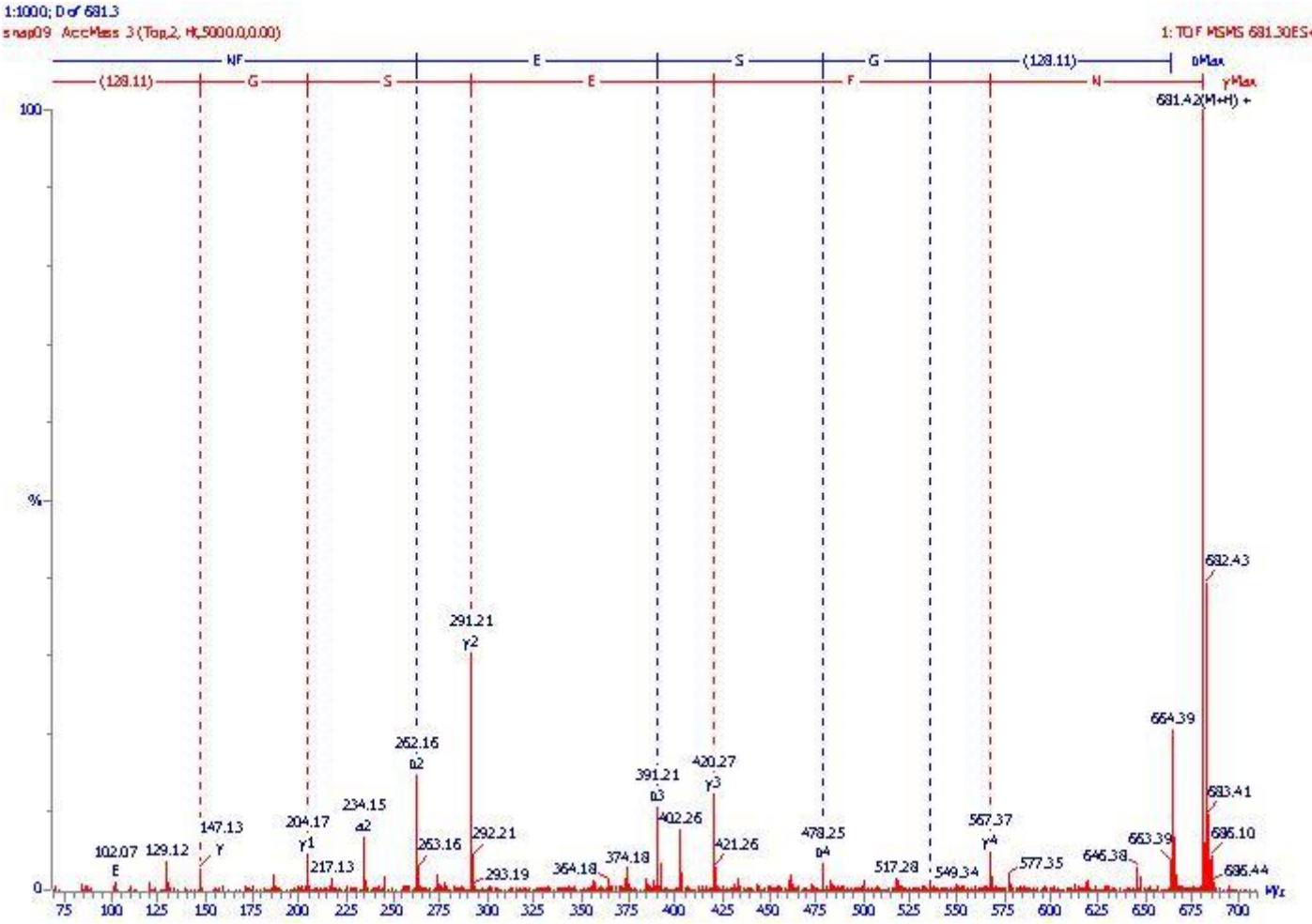
The extent of **side-chain fragmentation** detected depends on the type of analysers used in the mass spectrometer. A magnetic sector - magnetic sector instrument will give rise to **high energy** collisions resulting in many different types of side-chain cleavages. Quadrupole - quadrupole and quadrupole - time-of-flight mass spectrometers generate **low energy** fragmentations with fewer types of side-chain fragmentations.

Immonium ions (labelled "i") appear in the very low m/z range of the MS-MS spectrum. Each amino acid residue leads to a diagnostic immonium ion, with the exception of the two pairsleucine (L) and iso-leucine (I), and lysine (K) and glutamine (Q), which produce immonium ions with the same m/z ratio, i.e. m/z 86 for I and L, m/z 101 for K and Q. The immoniumions are useful for detecting and confirming many of the amino acid residues in a peptide, although no information regarding the position of these amino acid residues in the peptide sequence can be ascertained from the immonium ions.

An example of an **MS/MS daughter or product ion spectrum** is illustrated below. The molecular mass of the peptide was measured using standard mass spectrometric techniques and found to be 680.4 Da, the dominant ions in the MS spectrum being the protonated molecular ions ($M+H^+$) at m/z 681.4. These ions were selected for transmission through the first analyser, then fragmented in the collision cell and their fragments analysed by the

second analyser to produce the following MS/MS spectrum. The **sequence (amino acid backbone) ions** have been identified, and in this example the peptide fragmented

predominantly at the **CO-NH** bonds and gave both b and y" ions. (Often either the b series or the y" series predominates, sometimes to the exclusion of the other). The b series ions have been labelled with blue vertical lines and the y" series ions have been labelled with red vertical lines. The mass difference between adjacent members of a series can be calculated e.g. b3-b2 = 391.21 - 262.16 = 129.05 Da which is equivalent to a glutamine (E) amino acid residue; and similarly y4 - y3 = 567.37 - 420.27 = 147.10 Da which is equivalent to a phenylalanine (F) residue. In this way, using either the b series or the y" series, the amino acid sequence of the peptide can be determined and was found to be NFESGK (n.b. the y" series reads from right to left!). The immonium ions at m/z 102 merely confirm the presence of the glutamine (E) residue in the peptide.

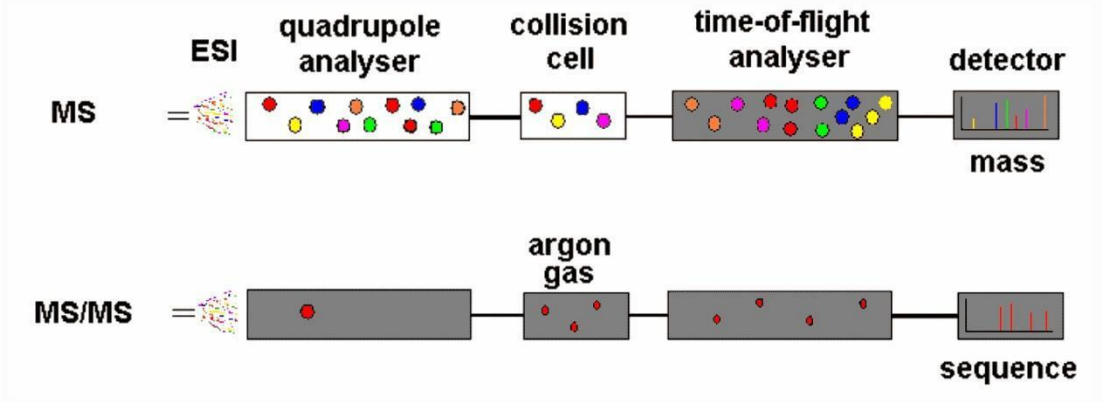


Peptide sequencing by tandem mass spectrometry - an MS-MS daughter or product ion spectrum.

A **protein identification study** would proceed as follows:

- a. The **protein** under investigation would be analysed by mass spectrometry to generate a molecular mass to within an accuracy of 0.01%.
- b. The protein would then be **digested** with a suitable enzyme. **Trypsin** is useful for mass spectrometric studies because each proteolytic fragment contains a basic **arginine (R)** or **lysine (K)** amino acid residue, and thus is eminently suitable for positive ionisation mass spectrometric analysis. The digest mixture is analysed -

without prior separation or clean-up - by mass spectrometry to produce a rather complex spectrum from which the molecular weights of all of the proteolytic fragments can be read. This spectrum, with its molecular weight information, is called a **peptide map**. (If the protein already exists on a **database**, then the peptide map is often sufficient to confirm the protein.)
For these experiments the mass spectrometer would be operated in the **"MS" mode**, whereby the sample is sprayed and ionised from the nanospray needle and the ions pass through the sampling cone, skimmer lenses, Rfhexapole focusing system, and the first (quadrupole) analyser. The quadrupole in this instance is not used as an analyser, merely as a lens to focus the ion beam into the second (time-of-flight) analyser which separates the ions according to their mass-to-charge ratio.



Q-TOF mass spectrometer operating in MS (upper) and MS/MS mode (lower) modes.

- c. With the digest mixture still spraying into the mass spectrometer, the Q-Tof mass spectrometer is switched into **"MS/MS" mode**. The protonated molecular ions of each of the digest fragments can be independently selected and transmitted through the quadrupole analyser, which is now used as an analyser to transmit solely the ions of interest into the **collision cell** which lies inbetween the first and second analysers. An inert gas such as argon is introduced into the collision cell and the sample ions are bombarded by the collision gas molecules which cause them to fragment. The optimum collision cell conditions vary from peptide to peptide and must be optimised for each one. The **fragment (**or **daughter or product**) **ions** are then analysed by the second (time-of-flight) analyser. In this way an **MS/MS spectrum** is produced showing all the **fragment ions** that arise directly from the chosen **parent** or **precursor ions** for a given peptide component.

  An **MS/MS daughter** (or **fragment**, or **product**) ion spectrum is produced for each of the components identified in the proteolytic digest. Varying amounts of sequence information can be gleaned from each fragmentation spectrum, and the spectra need to be interpreted carefully. Some of the processing can be automated, but in general the **processing** and **interpretation** of spectra will take longer than the data acquisition if accurate and reliable data are to be generated.

The amount of sequence information generated will vary from one peptide to another, Some peptide sequences will be confirmed totally, other may produce a partial sequence of, say, 4 or 5 amino acid residues. Often sequence "tag" of 4 or 5 residues is sufficient to search a protein database and confirm the identity of the protein.

Peptide Mass Fingerprinting (PMF)

This is another method of protein identification. In this method, 2-D gel electrophoresis is used for protein separation. The separated spots are obtained from the gel and then identified by PMF. The technique is based on the use of a proteolytic enzyme to digest the protein into smaller peptides. The most commonly used enzyme is trypsin, which cleaves lysine and arginine sites. When the digestion is complete, a set of peptides are produced of varying masses that are unique to that protein. The mass of each peptide will be the sum of amino acids present, inducing any modifications that amino acids might have undergone. Once the set of peptides have been obtained, one has to search for peptide sequences.