Chapter 3

3.4 The structure of proteins

Investigating proteins with Mass Spectrometry

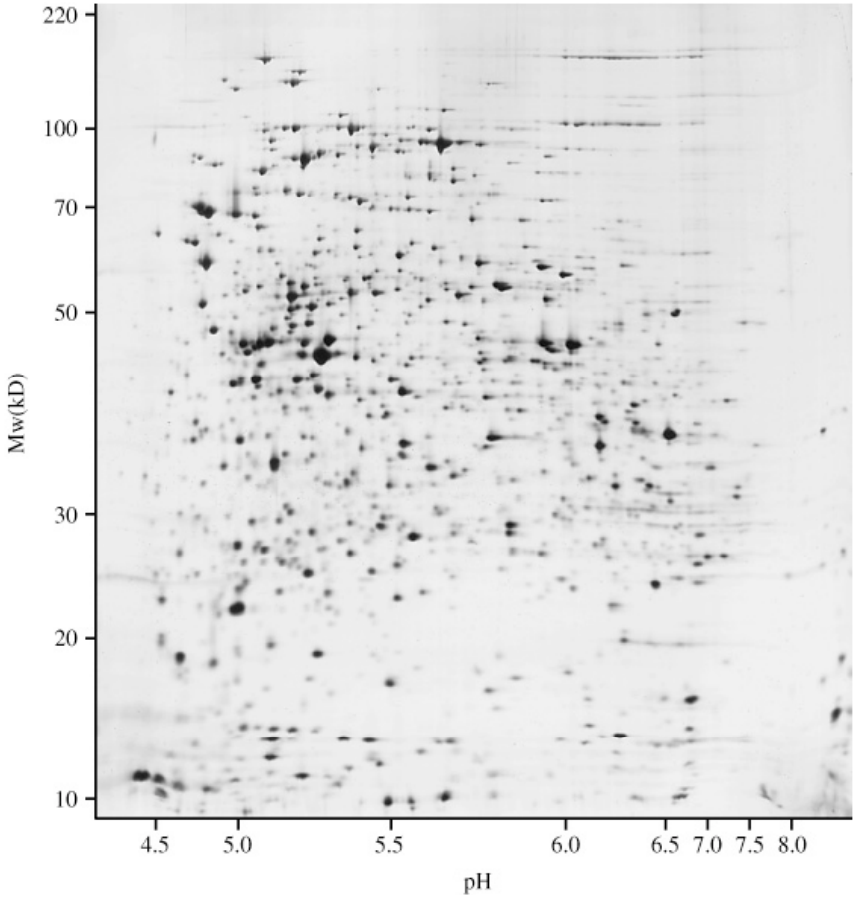
Proteome

was coined by **Marc Wilkins** in 1994

- **Proteomics (蛋白質體學)** was first coined in 1997 by Peter James to make an analogy with **genomics (基因體學)**, the study of the genes.
- **Proteomics** is the study of **large sets of proteins, such as the entire complement of proteins, including the modifications made to a particular set of proteins, expressed by a genome, or by a cell/organism or tissue/system type**.
- ***E. coli*** has about **4,000** different polypeptides (average size 300 amino acids, M_r 33,000)
- Fruit fly (***Drosophila*** *melanogaster*) about **16,000**, **humans, other mammals** about **40,000** different polypeptides

The Proteome of *E. coli*

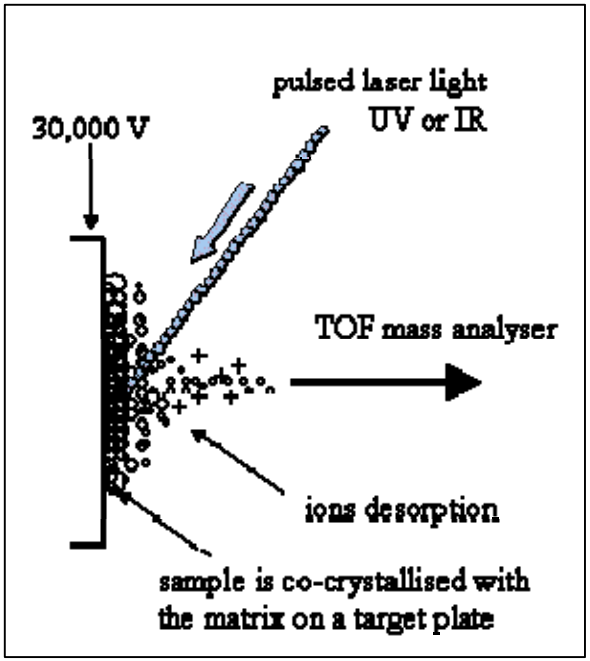
A <u>subset</u> of *E. coli* proteins on 2D gel electrophoresis



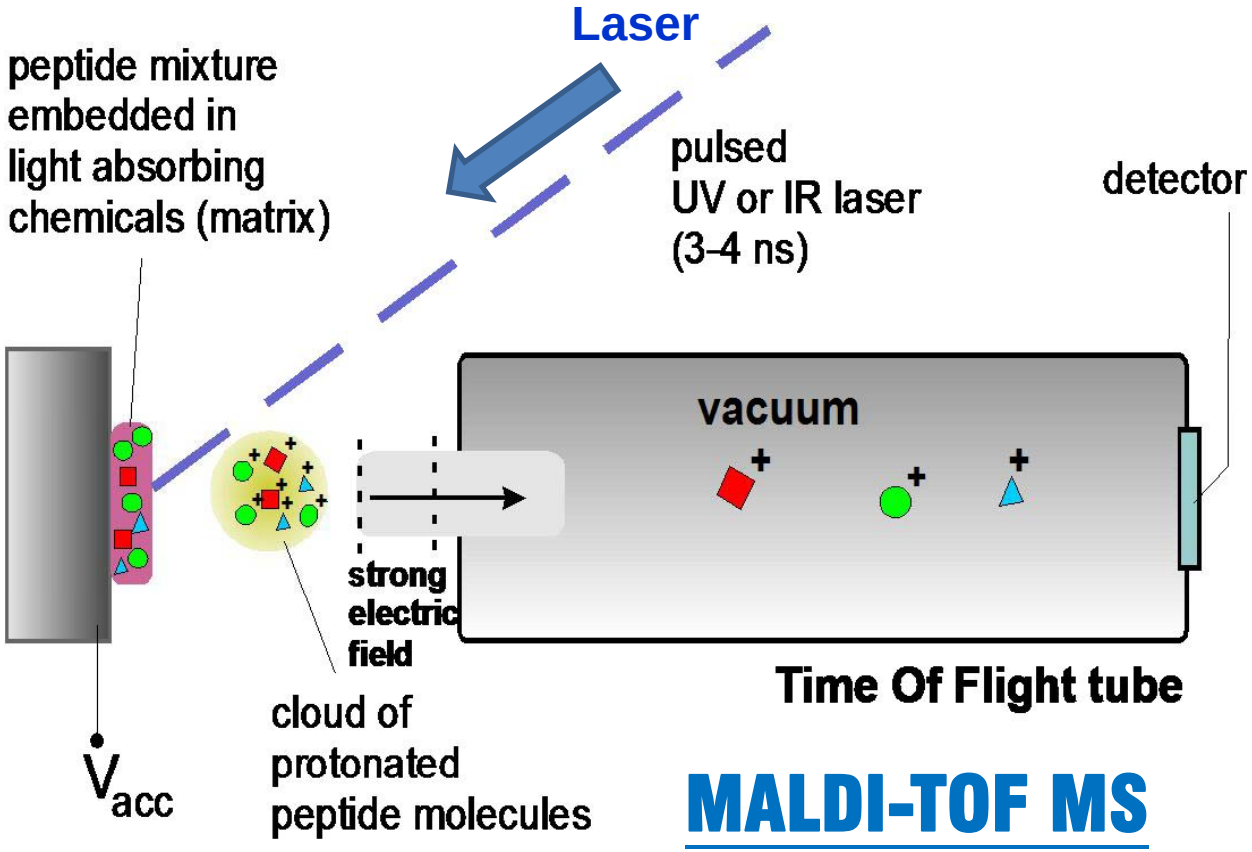
Two <u>MOST</u> related technologies for proteomics study:

2-D electrophoresis & Mass spectrometry

Matrix-assisted laser desorption/ionization (MALDI) mass spectrometry



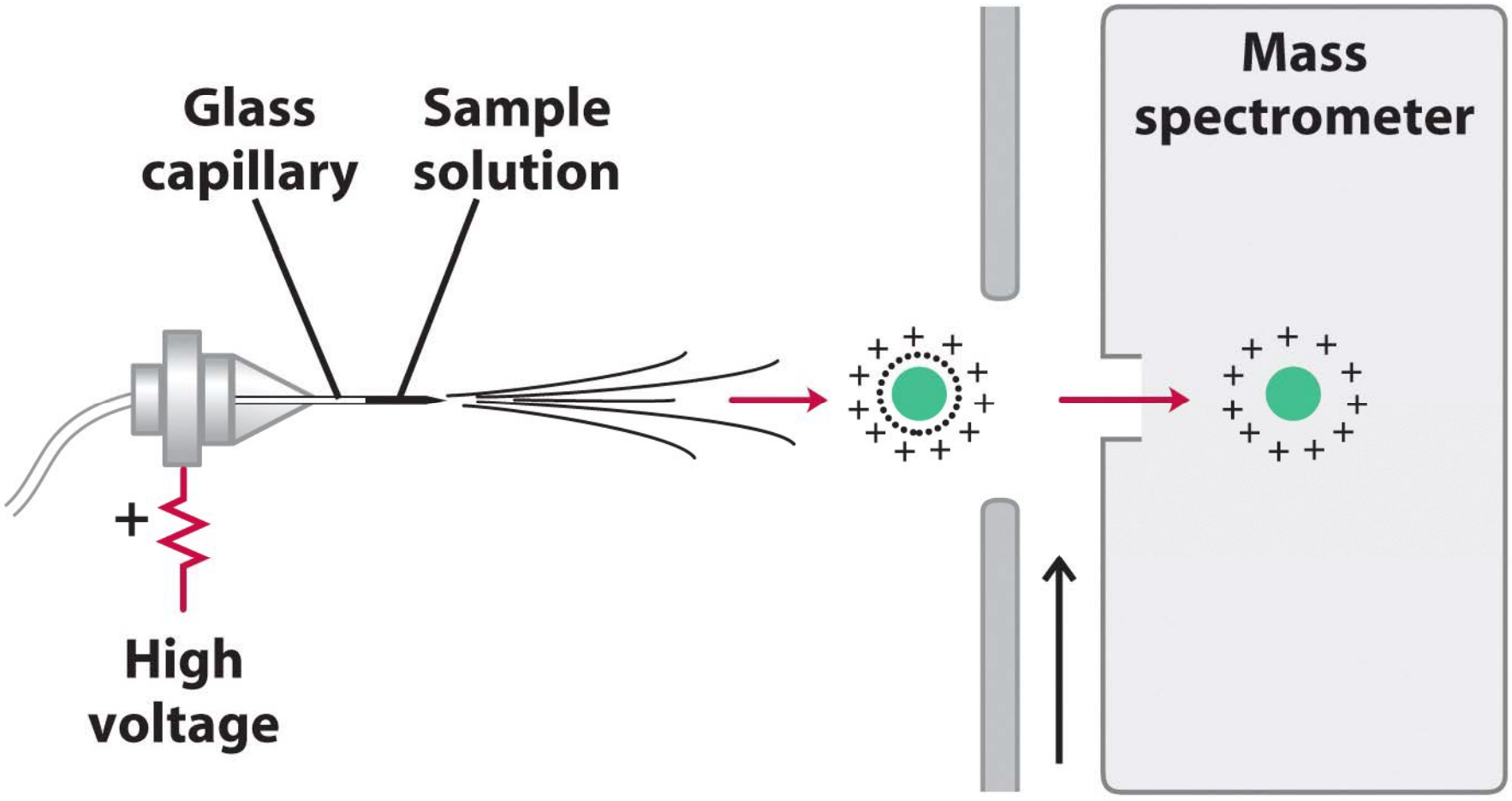
http://qbab.aber.ac.uk/roy/mss/tofspec.htm



MALDI-TOF MS

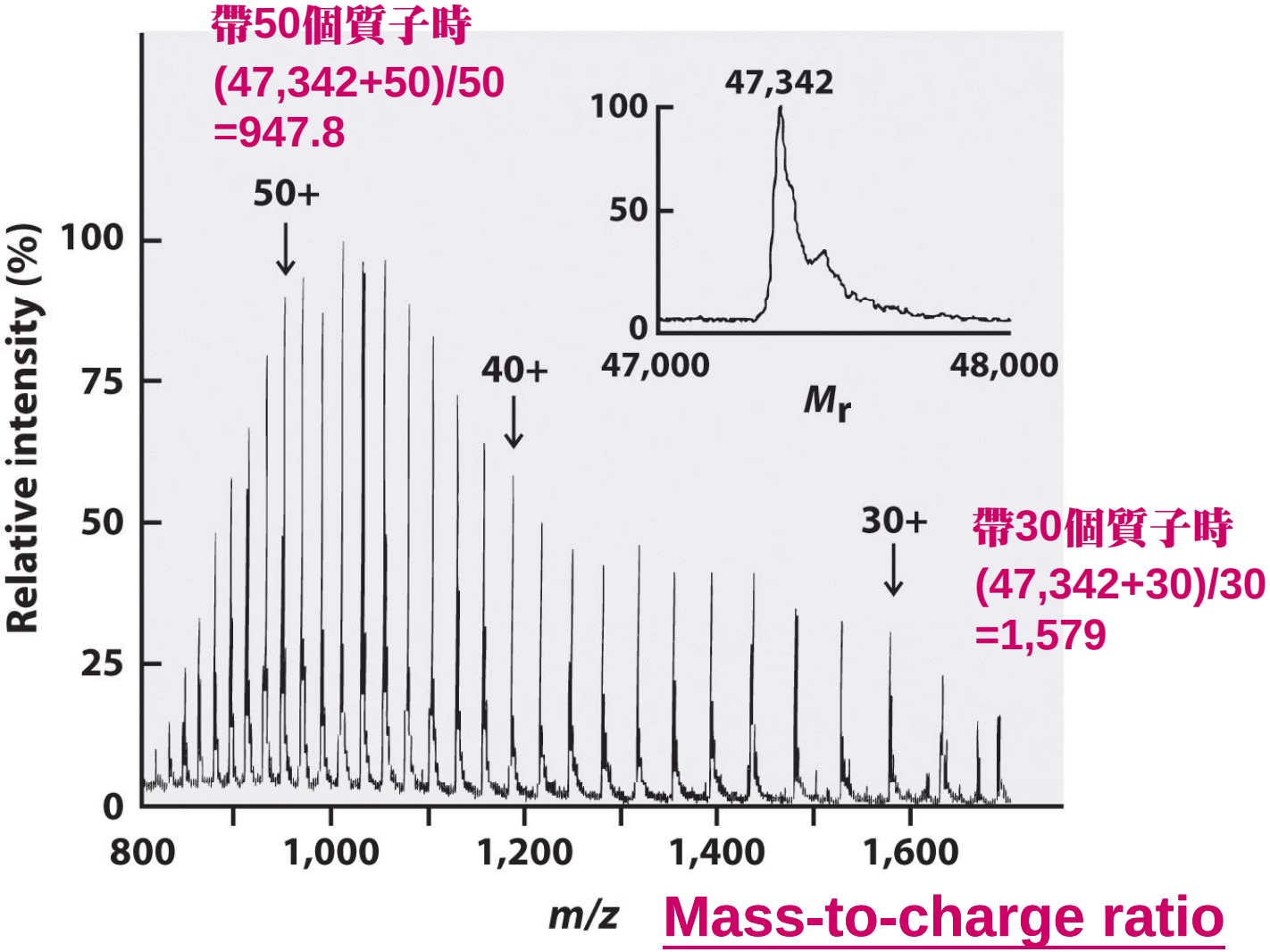
MALDI-TOF MS可以提供正確的質量 (mass) 資訊，卻不能獲得詳細的序列 (sequence data) 結果

Electrospray ionization mass spectrometry



ESI MS

Determining the molecular mass by ESI MS

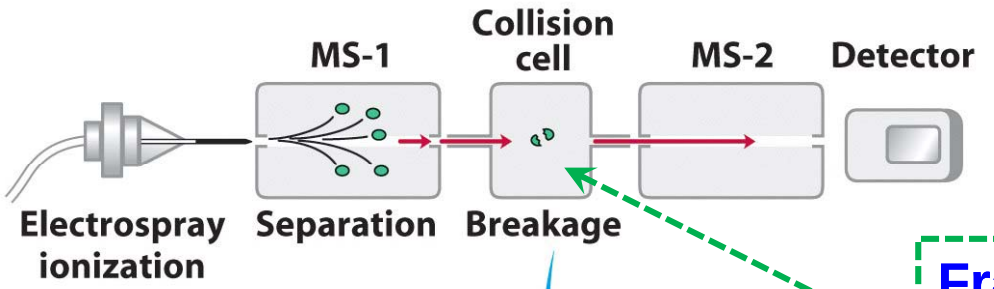


Mass-to-charge ratio

例如：一個 50 kDa 的蛋白質帶有一個質子時的 m/z = 50,001，當其在溶液中接受 20 個質子時的 m/z 值為 50,020/20=2,501

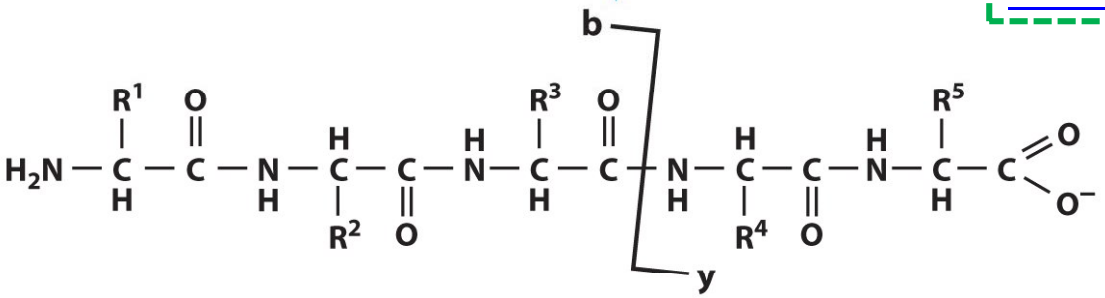
Obtaining peptide/protein sequence information with tandem MS (串聯質譜儀)

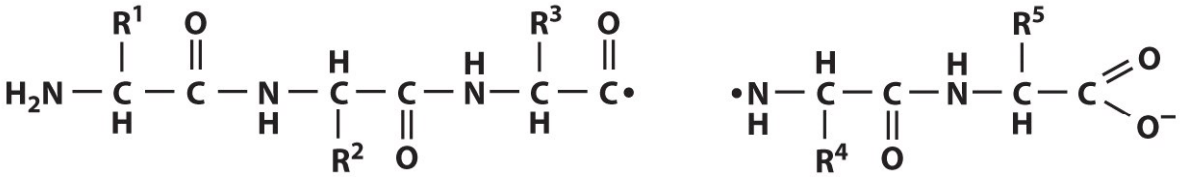
A solution containing the protein is first treated with a protease or chemical reagent to hydrolyze it to a mixture of shorter peptides



ESI tandem MS

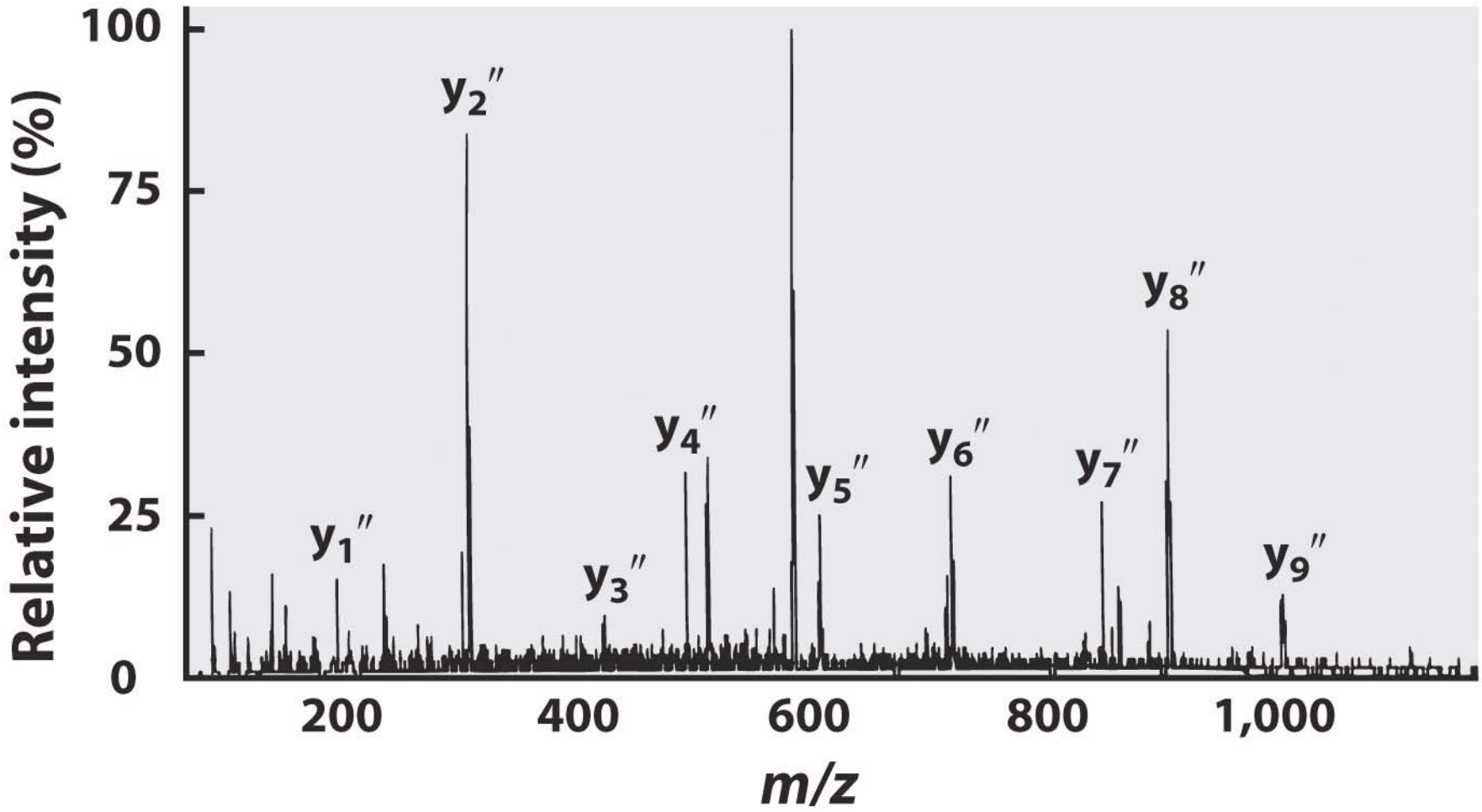
Fragmented by "collision gas" such as helium or argon





Tandem MS with HPLC: LC tandem MS or LC MS/MS) 液相層析串聯質譜儀

The successive peaks differ by the mass of a particular amino acid in the original peptide



Phe-Pro-Gly-Gln-(Ile/Leu)-Asn-Ala-Asp-(Ile/Leu)

Consensus sequence

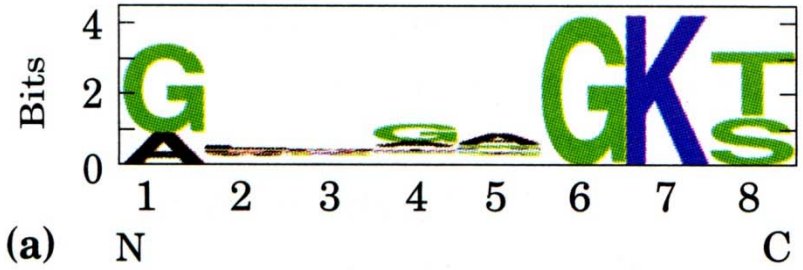
Consensus sequence is applied to such sequences in DNA, RNA, or protein. When a series of related nucleic acid or protein sequences are compared, a consensus sequence is the one that reflects the most common base or amino acid at each position. Parts of the sequence that have particularly good agreement often represent evolutionarily conserved functional domains.

Presentations of two consensus sequence by sequence logos

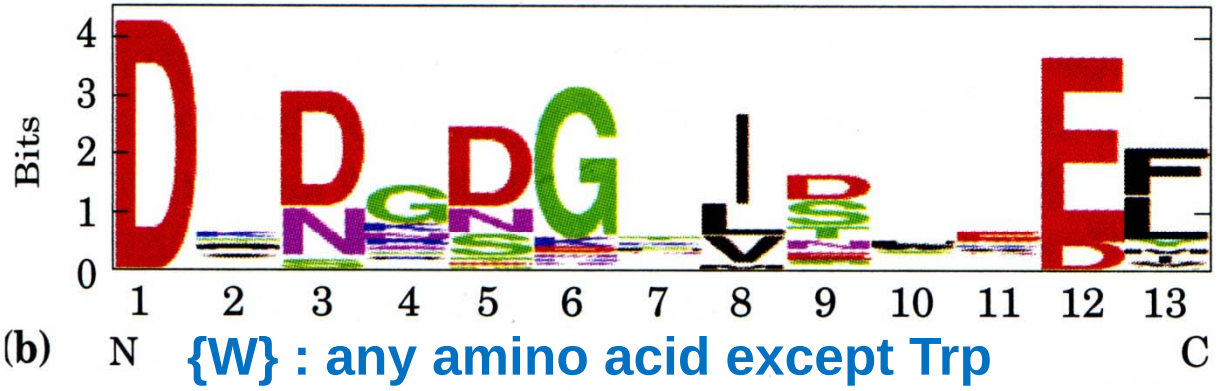
- Relative frequency at that position
- Degree of sequence conservation

[AG] : Ala or Gly

[AG]-x(4)-G-K-[ST].



D-{W}-[DNS]-{ILVFYW}-[DENSTG]-[DNQGHRK]-{GP}-[LIVMC]-[DENQSTAGC]-x(2)-[DE]-[LIVMFYW].



{W} : any amino acid except Trp

FIGURE 1 **Representations of two consensus sequences.** **(a)** P loop, an ATP-binding structure; **(b)** EF hand, a Ca^{2+}-binding structure.

ExPASy Proteomics Server
http://au.expasy.org/

SIB Swiss Institute of Bioinformatics

Search ExPASy web site for Go Clear

ExPASy Proteomics Server

Databases Tools Services Mirrors About Contact

You are here: ExPASy AU

The ExPASy (**Ex**pert **P**rotein **A**nalysis **Sy**stem) proteomics server of the Swiss Institute of Bioinformatics (SIB) is dedicated to the analysis of protein sequences and structures as well as 2-D PAGE (Disclaimer / References / Linking to ExPASy).

Databases

UniProtKB, PROSITE, HAMAP, SwissVar, ViralZone, SWISS-MODEL Repository, SWISS-2DPAGE, World-2DPAGE Repository, MIAPEGelDB, ENZYME, GlycoSuiteDB, UniPathway

[details] [full list]

Tools & Software

Proteomics tools, Blast, ScanProsite, Melanie, MSight, Make2D-DB, SWISS-MODEL, Swiss-PdbViewer, SwissDock, SwissParam

[full list]

Education & services

Downloads, Protein Spotlight, Protéines à la ‹‹Une››, e-proxemis, Bioinformatics core facility for Proteomics, Click2Drug - in silico Drug Design tools

[full list]

Documentation

What's New?, E-mail alerts, UniProtKB documentation, How to link to ExPASy, Advanced search

[full list]

Latest News

New tools for in silico drug design and molecular modeling - September 29, 2010

We are pleased to announce the release of two new tools developed by the Molecular Modeling group of the SIB for computer-aided drug design and molecular modeling: SwissDock, a docking web service to predict the molecular interactions that occurs between a target protein and a small molecule, and SwissParam, a web service to provide topology and parameters for small organic molecules for use with CHARMM and GROMACS. The Molecular Modeling group is also launching a new directory of in silico drug design tools: Click2Drug.

New proteomics data uploaded into the World-2DPAGE Repository - July 28, 2010

Data from recents publications has been added into the World-2DPAGE Repository. Currently, 131 maps for 20 species are available and queriable

Last modified 28/Sep/2010 by GRR

Swiss Institute of Bioinformatics | Disclaimer | Sitemap | Documentation | Contact Us

Primary structure analysis

- ProtParam - Physico-chemical parameters of a protein sequence (amino-acid and atomic compositions, isoelectric point, extinction coefficient, etc.)
- Compute pI/Mw - Compute the theoretical isoelectric point (*pI*) and molecular weight (*Mw*) from a UniProt Knowledgebase entry or for a user sequence
- ScanSite pI/Mw - Compute the theoretical *pI* and *Mw*, and multiple phosphorylation states
- MW, pI, Titration curve - Computes *pI*, composition and allows to see a titration curve
- Scratch Protein Predictor new
- HeliQuest - A web server to screen sequences with specific alpha-helical properties
- Radar - De novo repeat detection in protein sequences
- REP - Searches a protein sequence for repeats
- REPRO - De novo repeat detection in protein sequences
- TRUST - De novo repeat detection in protein sequences
- XSTREAM - De novo tandem repeat detection and architecture modeling in protein sequences
- SAPS - Statistical analysis of protein sequences at EMBnet-CH [Also available at EBI]
- Coils - Prediction of coiled coil regions in proteins (Lupas's method) at EMBnet-CH [Also available at PBIL]
- Paircoil - Prediction of coiled coil regions in proteins (Berger's method)
- Paircoil2 - Prediction of the parallel coiled coil fold from sequence using pairwise residue probabilitis with the Paircoil algorithm.
- Multicoil - Prediction of two- and three-stranded coiled coils
- 2ZIP - Prediction of Leucine Zippers
- ePESTfind - Identification of PEST regions
- HLA_Bind - Prediction of MHC type I (HLA) peptide binding
- PEPVAC - Prediction of supertypic MHC binders
- RANKPEP - Prediction of peptide MHC binding
- SYFPEITHI - Prediction of MHC type I and II peptide binding
- ProtScale - Amino acid scale representation (Hydrophobicity, other conformational parameters, etc.)
- Drawhca - Draw an HCA (Hydrophobic Cluster Analysis) plot of a protein sequence
- Peptide Builder
- Protein Colourer - Tool for coloring your amino acid sequence
- Three To One and One to Three - Tools to convert a three-letter coded amino acid sequence to single letter code and vice versa
- Three-/one-letter amino acid converter - Tool which converts amino acid codes from three-letter to one-letter and vice versa.
- Colorseq - Tool to highlight (in red) a selected set of residues in a protein sequence
- RandSeq - Random protein sequence generator

Secondary structure prediction

- AGADIR - An algorithm to predict the helical content of peptides
- APSSP - Advanced Protein Secondary Structure Prediction Server
- CFSSP - Chou & Fasman Secondary Structure Prediction Server new
- GOR - Garnier et al, 1996
- HNN - Hierarchical Neural Network method (Guermeur, 1997)
- HTMSRAP - Helical TransMembrane Segment Rotational Angle Prediction
- Jpred - A consensus method for protein secondary structure prediction at University of Dundee
- JUFO - Protein secondary structure prediction from sequence (neural network)
- NetSurfP - Protein Surface Accessibility and Secondary Structure Predictions new
- nnPredict - University of California at San Francisco (UCSF)
- Porter - University College Dublin
- PredictProtein - PHDsec, PHDacc, PHDhtm, PHDtopology, PHDthreader, MaxHom, EvalSec from Columbia University
- Prof - Cascaded Multiple Classifiers for Secondary Structure Prediction
- PSA - BioMolecular Engineering Research Center (BMERC) / Boston
- PSIpred - Various protein structure prediction methods at Bloomsbury Centre for Bioinformatics
- SOPMA - Geourjon and Deléage, 1995
- Scratch Protein Predictor new
- DLP-SVM - Domain linker prediction using SVM at Tokyo University of Agriculture and Technology

Tertiary structure

Tertiary structure analysis

- iMolTalk - An Interactive Protein Structure Analysis Server (currently down)
- MolTalk - A computational environment for structural bioinformatics
- COPS - Navigation through fold space and the instantaneous visualization of pairwise structure similarities
- PoPMuSiC - Prediction of thermodynamic stability changes upon point mutations; design of modified proteins new
- Seq2Struct - A web resource for the identification of sequence-structure links
- STRAP - A structural alignment program for proteins
- TLSMD - TLS (Translation/Libration/Screw) Motion Determination
- TopMatch-web - Protein structure comparison

Tertiary structure prediction

Homology modeling

- SWISS-MODEL - An automated knowledge-based protein modelling server
- 3Djigsaw - Three-dimensional models for proteins based on homologues of known structure
- CPHmodels - Automated neural-network based protein modelling server
- ESyPred3D - Automated homology modeling program using neural networks
- Geno3d - Automatic modelling of protein three-dimensional structure

Post-translational modification prediction

- ChloroP - Prediction of chloroplast transit peptides
- LipoP - Prediction of lipoproteins and signal peptides in Gram negative bacteria
- MITOPROT - Prediction of mitochondrial targeting sequences
- PATS - Prediction of apicoplast targeted sequences
- PlasMit - Prediction of mitochondrial transit peptides in Plasmodium falciparum
- Predotar - Prediction of mitochondrial and plastid targeting sequences
- PTS1 - Prediction of peroxisomal targeting signal 1 containing proteins
- SignalP - Prediction of signal peptide cleavage sites

- DictyOGlyc - Prediction of GlcNAc O-glycosylation sites in Dictyostelium
- NetCGlyc - C-mannosylation sites in mammalian proteins
- NetOGlyc - Prediction of O-GalNAc (mucin type) glycosylation sites in mammalian proteins
- NetGlycate - Glycation of epsilon amino groups of lysines in mammalian proteins
- NetNGlyc - Prediction of N-glycosylation sites in human proteins
- OGPET - Prediction of O-GalNAc (mucin-type) glycosylation sites in eukaryotic (non-protozoan) proteins
- YinOYang - O-beta-GlcNAc attachment sites in eukaryotic protein sequences

- big-PI Predictor - GPI Modification Site Prediction
- GPI-SOM - Identification of GPI-anchor signals by a Kohonen Self Organizing Map
- Myristoylator - Prediction of N-terminal myristoylation by neural networks
- NMT - Prediction of N-terminal N-myristoylation
- CSS-Palm - Palmitoylation site prediction with CSS
- PrePS - Prenylation Prediction Suite

- NetAcet - Prediction of N-acetyltransferase A (NatA) substrates (in yeast and mammalian proteins)
- NetPhos - Prediction of Ser, Thr and Tyr phosphorylation sites in eukaryotic proteins
- NetPhosK - Kinase specific phosphorylation sites in eukaryotic proteins
- NetPhosYeast - Serine and threonine phosphorylation sites in yeast proteins
- GPS - Prediction of kinase-specific phosphorylation sites for 408 human protein kinases in hierarchy new
- Sulfinator - Prediction of tyrosine sulfation sites
- SulfoSite - Prediction of tyrosine sulfation sites
- SUMOplot - Prediction of SUMO protein attachment sites
- SUMOsp - Prediction of sumoylation sites
- TermiNator - Prediction of N-terminal modification (version 3)

- NetPicoRNA - Prediction of protease cleavage sites in picornaviral proteins
- NetCorona - Coronavirus 3C-like proteinase cleavage sites in proteins
- ProP - Arginine and lysine propeptide cleavage sites in eukaryotic protein sequences

DNA -> Protein

- Translate - Translates a nucleotide sequence to a protein sequence
- Transeq - Nucleotide to protein translation from the EMBOSS package
- Graphical Codon Usage Analyser - Displays the codon bias in a graphical manner
- BCM search launcher - Six frame translation of nucleotide sequence(s)

- Reverse Translate - Translates a protein sequence back to a nucleotide sequence
- (Reverse)-Transcription and Translation Tool

- Genewise - Compares a protein sequence to a genomic DNA sequence, allowing for introns and frameshifting errors

Similarity searches

- BLAST Network Service on ExPASy
- BLAST at EMBnet-CH/SIB (Switzerland)
- BLAST at NCBI
- WU-BLAST at Bork's group in EMBL (Heidelberg)
- WU-BLAST and BLAST at the EBI (Hinxton)
- BLAST at PBIL (Lyon)
- Fasta3 - FASTA version 3 at the EBI
- MPsrch - Smith/Waterman sequence comparison at EBI
- PropSearch - Structural homolog search using a 'properties' approach at Montpellier
- SAMBA - Systolic Accelerator for Molecular Biological Applications
- SAWTED - Structure Assignment With Text Description
- Scanps - Similarity searches using Barton's algorithm
- SEQUEROME - BLAST similarity search and sequence profiling at Georgetown University
- SHOPS - Analysis of the genomic operon context for any group of proteins

- BLAST2FASTA - Converts NCBI BLAST output into FASTA format

Pattern and profile searches

- InterPro Scan - Integrated search in PROSITE, Pfam, PRINTS and other family and domain databases
- Hits - Relationships between protein sequences and motifs

Homologs

Paralogs **Orthologs**

- The members of protein families are called homologous proteins, or homologs
- If two proteins in a family (that is , two homologs) are present in the same species, they are referred to as paralogs
- Homologs from different species are called orthologs

Comparisons of the primary structures of proteins reveal evolutionary relationships

- **Closely related species contain proteins with very similar amino acid sequences**
- Differences reflect evolutionary change **from a common ancestral protein sequence**
- **Cytochrome *c*** protein sequences from various species can be aligned to show their similarities
- **Phylogenetic tree (系統發生樹/親源關係樹/演化樹)** shows evolutionary differences in amino acid sequences

Cytochrome c is part of the respiratory chain during cellular respiration. Cytochrome c is found in the mitochondria of every aerobic eukaryote — animal, plant, and protist. The amino acid sequences of many of these have been determined, and comparing them shows that they are related.

A bacterial evolutionary tree derived from GroEL family of amino acid sequence comparisons