Summary of the course proteins & lipids FS2018

29.02.2018 – 28.05.2018

Table of content

Abbreviations

AA: amino acid

aaRS/ARS: aminoacyl-tRNA synthetase

ABPP: activity based protein profiling

Acm: acetamidomethyl

Boc: *tert*-butoxycarbonyl

BSA: bovine serum albumin

CAT: chloramphenicol acetyltransferase

CD: circular dichroism

CPP: cell-penetrating peptide

DCC: *N,N'*-dicyclohexylcarbodiimide

DCM: dichloromethane / methylene dichloride

DCU: dicyclohexylurea

DMF: dimethylformamide

ECM: extracellular matrix

EF: elongation factor

EPO: erythropoietin

FA: fatty acid

FACS: fluorescence activated cell sorting

Fmoc: 9-fluorenylmethoxycarbonyl

FRET: fluorescence/Förster energy transfer

GFP: green fluorescent protein

HOBt: hydroxybenzotriazole

HPLC: high-performance liquid chromatography

ICAT: isotope-coded affinity tag

IF: initiation factor

ivTT: in vitro transcription & translation

LC: liquid chromatography

MS: mass spectroscopy

NCL: native chemical ligation

NMR: nuclear magnetic resonance

NRPS: non-ribosomal peptide synthetase

PCR: polymerase chain reaction

PEP : phosphoenolpyruvate

Pi: phosphate

POI: protein of interest

PPi: pyrophosphate

PTM: posttranslational modification

RF: release factor

RMSD: root-mean-square deviation

scFv: single chain variable fragment

SEP: synthetic erythropoiesis protein

SPPS: solid phase peptide synthesis

TFA: trifluoroacetic acid

Tris: tris(hydroxymethyl)aminomethane

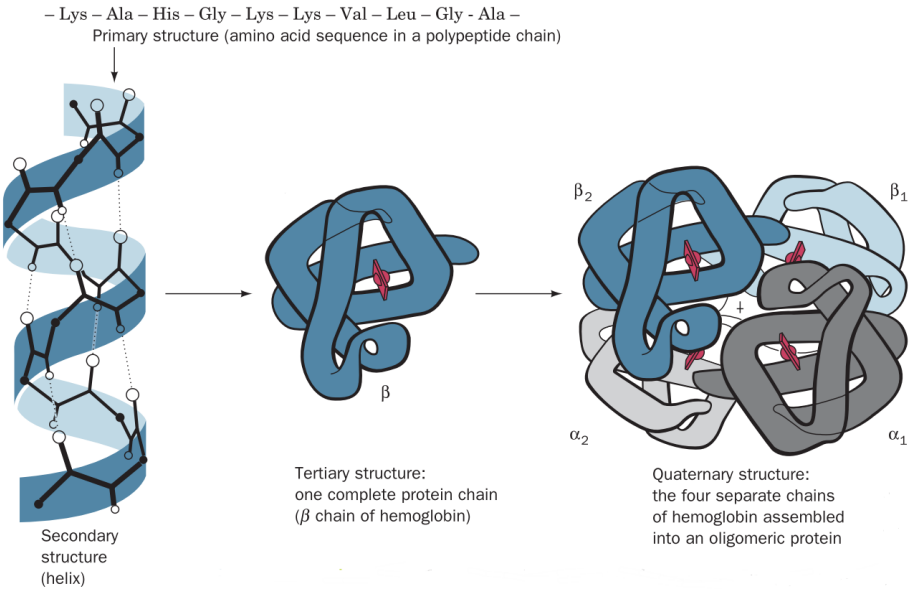
Xaa: any of the 21 basic amino acids

Protein structure and stability 蛋白质的结构与稳定性

(19.02.2018)

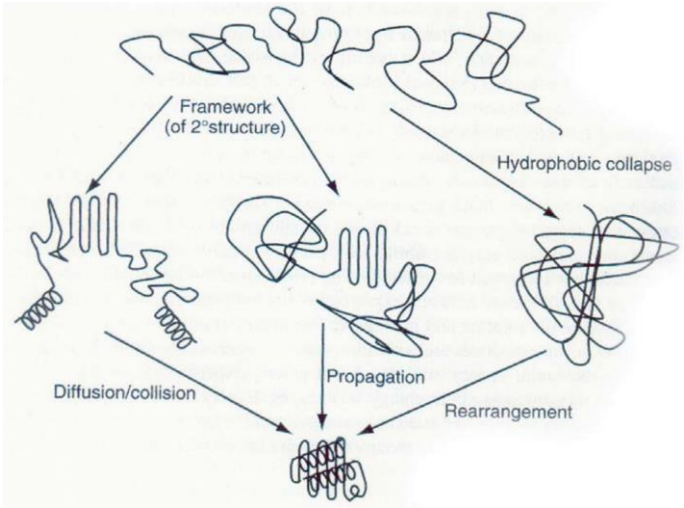
General

- The first X-ray structure of a protein (sperm whale myoglobin), was reported in 1958 by John Kendrew et al. At the time, protein chemists were chagrined by the complexity and apparent lack of regularity in the structure of myoglobin. In retrospect, such irregularity seems essential for proteins to fulfil their diverse biological roles. However, comparisons of the nearly 50,000 protein structures now known have revealed that proteins actually exhibit a remarkable degree of structural regularity.
- Four levels of structural complexity

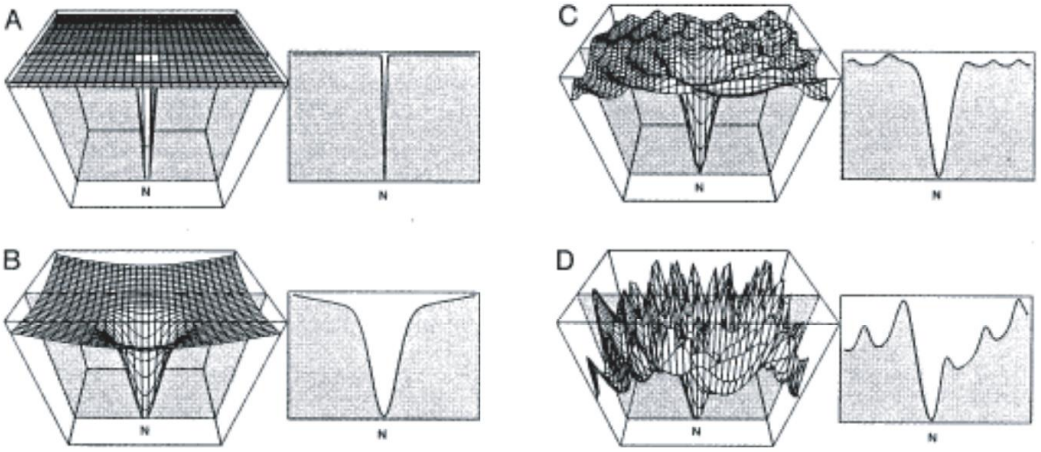


  - Primary structure: linear sequence of amino acids (AAs) in a protein
  - Secondary structure: local spatial arrangement of a polypeptide's backbone atoms without regard to the conformations of its side chains
  - Tertiary structure: three-dimensional structure of an entire polypeptide, including its side chains
  - Quaternary structure: spatial arrangement of the subunits in a protein
- Protein folding

  Three classical mechanisms:

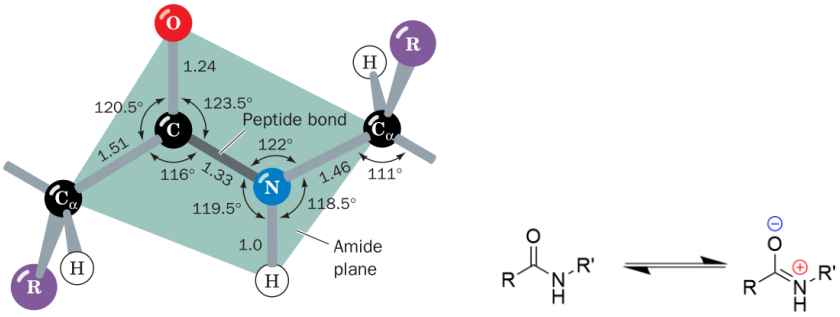


The folding funnel hypothesis assumes that a protein's native state corresponds to its free energy minimum under the solution conditions usually encountered in cells.
(Energy is on the vertical axis and the other axes represent conformational degrees of freedom. N is the native structure. A) "Golf-course" landscape. B) Smooth funnel landscape in which every conformation can reach N without encountering barriers. C) Rough landscape. D) "Rugged" landscape with local minima and barriers.)

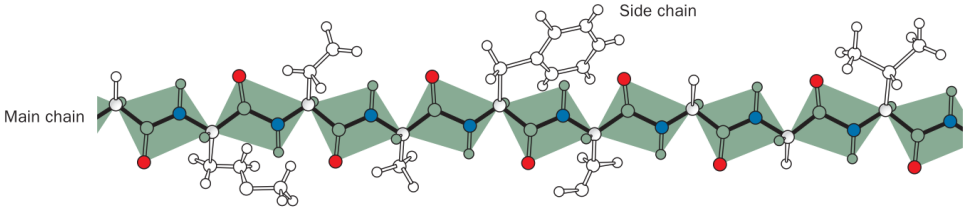


Secondary structure

- The peptide group has a rigid, planar structure as a consequence of resonance interactions that give the peptide bond ~40% double-bond character. This explanation is supported by the observations that a peptide group's C-N bond is 0.13 Å shorter than its N-C single bond and that its C=O double bond is 0.02 Å longer than that of aldehydes and ketones.

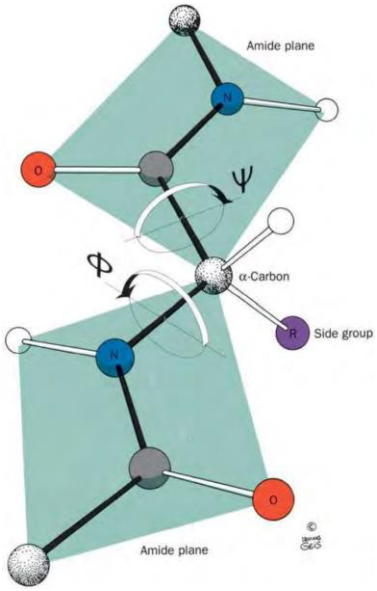


- Peptide groups, with few exceptions, assume the trans conformation, in which successive C_α atoms are on opposite sides of the peptide bond joining them. The cis conformation is ~8 kJ/mol less stable than the trans conformation because of steric interference between neighbouring side chains. This steric interference is reduced in peptide bonds to proline residues, so ~10% of proline residues in proteins follow a cis peptide bond.

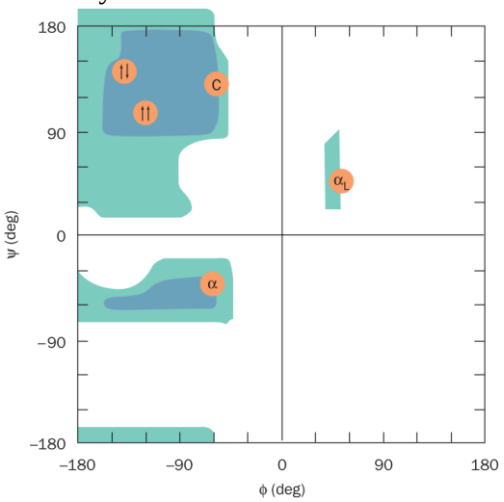


- The backbone or main chain of a protein refers to the atoms that participate in peptide bonds. Conformation of the backbone can be described by the torsion angles (also called dihedral angles or rotation angles) around the C_α-N bond (ϕ) and the C_α-C bond (ψ) of each residue.

Both angles are defined as 180° when the polypeptide chain is in its fully extended conformation and increase clockwise when viewed from C_α.



Sterically allowed and forbidden conformations are indicated by the Ramachandran diagram.



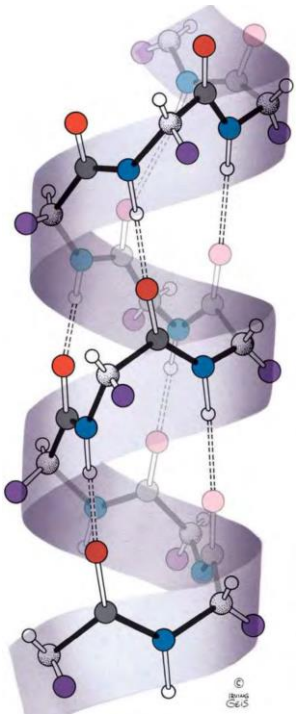
- Regular secondary structures

  A few elements of protein secondary structure are so widespread that they are immediately recognisable in proteins with widely differing sequences. These are called regular secondary structures because they are composed of sequences of residues with repeating ϕ and ψ values.

  - α helix

    In the α helix, the backbone hydrogen bonds are arranged such that the peptide C=O bond of the n-th residue points along the helix axis toward the peptide N-H group of the (n + 4)-th residue. This results in a strong hydrogen bond that has the nearly optimum N-O distance (2.8 Å). Side chains project outward and downward from the helix, thereby avoiding steric interference with the backbone and with each other. The core of the helix is tightly packed, i.e. its atoms are in van der Waals contact.

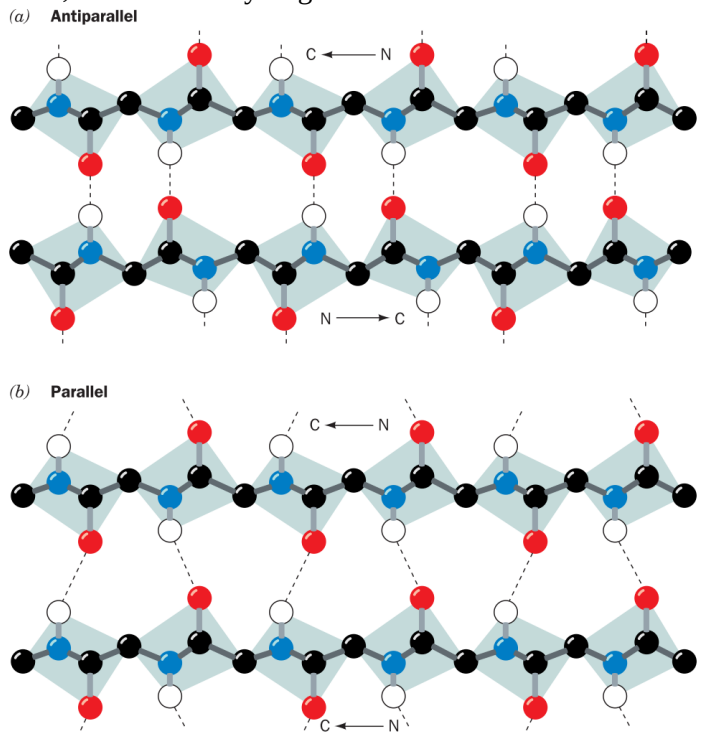
    The α helix is right-handed; that is, it turns in the direction that the fingers of a right hand curl when its thumb points in the direction that the helix rises. The α helix has 3.6 residues per turn and a pitch (the distance the helix rises per turn) of 5.4 Å.



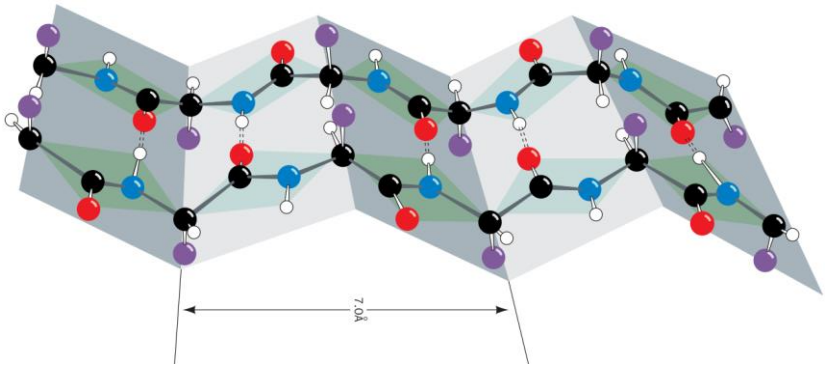
- β (pleated) sheet

  Like the α helix, the β sheet uses the full hydrogen-bonding capacity of the polypeptide backbone. In β sheets, however, hydrogen bonding occurs between neighbouring polypeptide chains rather than within one.

  β sheets come in two varieties: (1) the antiparallel β sheet, in which neighbouring hydrogen-bonded polypeptide chains run in opposite directions, and (2) the parallel β sheet, in which the hydrogen-bonded chains extend in the same direction.

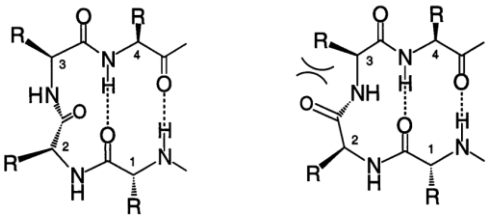


The conformations in which these structures are optimally hydrogen-bonded vary somewhat from that of the fully extended polypeptide. They therefore have a rippled (pleated) edge-on appearance. Successive side chains of a polypeptide chain in a β sheet extend to opposite sides of the sheet with a two-residue repeat distance of 7.0 Å.



- Reverse turn
  Polypeptide segments with regular secondary structures are often joined by stretches of polypeptide that abruptly change direction. Such reverse turns or β bends (so named because they often connect successive strands of antiparallel β sheets) almost always occur at protein surfaces. They usually involve four successive AA residues arranged in one of two ways, Type I and Type II, that differ by a 180° flip of the peptide unit linking residues 2 and 3. Both types are stabilized by a hydrogen bond, although deviations from these ideal conformations often disrupt this hydrogen bond. In Type II turns, the oxygen atom of residue 2 crowds the side chain of residue 3, which is therefore usually Gly. Residue 2 of either type of turn is often Pro since it can assume the required conformation.

Type I β-Turn **Type II β-Turn**

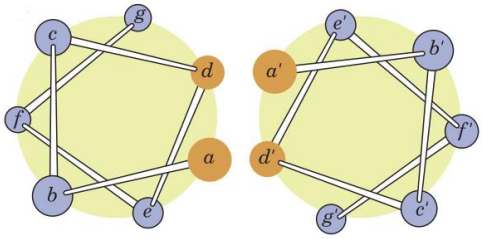


$(\varphi,\psi)_2 = -60°, -30°$ $(\varphi,\psi)_3 = -90°, 0°$

$(\varphi,\psi)_2 = -60°, 120°$ $(\varphi,\psi)_3 = 90°, 0°$

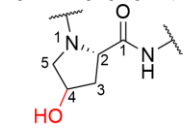
- Keratin and coiled coil
  The X-ray diffraction pattern of α keratin resembles that expected for an α helix. However, α keratin exhibits a 5.1-Å spacing rather than the 5.4-Å distance corresponding to the pitch of the α helix. This discrepancy is the result of two α keratin polypeptides, each of which forms an α helix, twisting around each other to form a left-handed coil. The normal 5.4-Å repeat distance of each α helix in the pair is thereby tilted relative to the axis of this assembly, yielding the observed 5.1-Å spacing. The assembly is said to have a coiled coil structure because each α helix itself follows a helical path.
  Besides keratin, coiled coils also occur in numerous other proteins.

The conformation of α keratin's coiled coil is a consequence of its primary structure: The central 310-residue segment of each polypeptide chain has a 7-residue pseudorepeat, a-b-c-d-e-f-g, with nonpolar residues predominating at positions a and d.

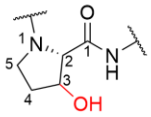


Since an α helix has 3.6 residues per turn, α keratin's a and d residues line up along one side of each α helix. The hydrophobic strip along one helix associates with the hydrophobic strip on another helix. Because the 3.5-residue repeat in α keratin is slightly smaller than the 3.6 residues per turn of a standard α helix, the two keratin helices are inclined about 18° relative to one another, resulting in the coiled coil arrangement.

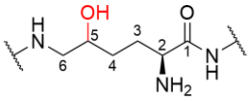
- Collagen and triple helix
  The strong, insoluble collagen fibres are the major stress-bearing components of connective tissues (bone, teeth, cartilage, tendon, etc.). Collagen molecules were found to consist of three polypeptide chains forming a rigid, well-packed triple helix.
  Collagen has a distinctive amino acid composition: Nearly one-third of its residues are Gly; another 15 to 30% of its residues are Pro and 4-hydroxyprolyl (Hyp). 3-Hydroxyprolyl and 5-hydroxylysyl (Hyl) residues also occur in collagen, but in smaller amounts.



4-hydroxyprolyl residue (Hyp)

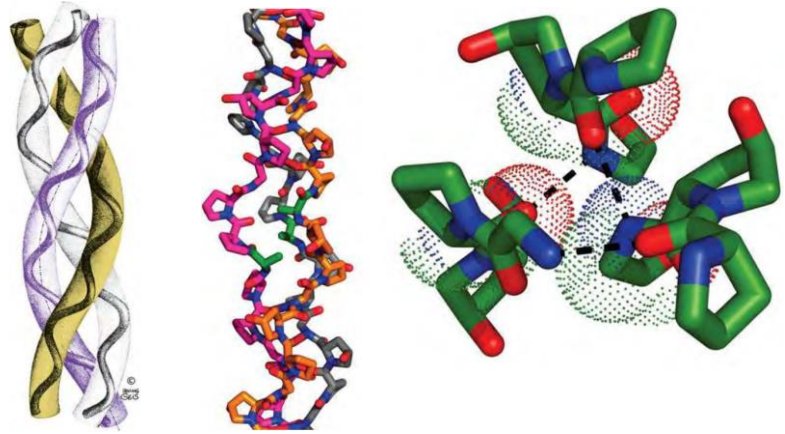


3-hydroxyprolyl residue



5-hydroxylysyl residue (Hyl)

  The AA sequence of a typical collagen polypeptide consists of monotonously repeating triplets of sequence Gly-X-Y over a segment of 1000 residues, where X is often Pro and Y is often Hyp and sometimes Hyl. Collagen's Pro residues prevent it from forming an α helix. Instead, the collagen polypeptide assumes a left-handed helical conformation with about three residues per turn. Three parallel chains wind around each other with a gentle, right-handed, rope-like twist to form the triple-helical structure of a collagen molecule.



- Nonrepetitive structures
  A significant portion of a protein's structure may also be irregular or unique. These nonrepetitive structures are no less ordered than are α helices or β sheets; they are simply irregular and hence more difficult to describe. Segments of polypeptide chains whose successive residues do not have similar ϕ and ψ values are sometimes called coils.

Secondary structure prediction

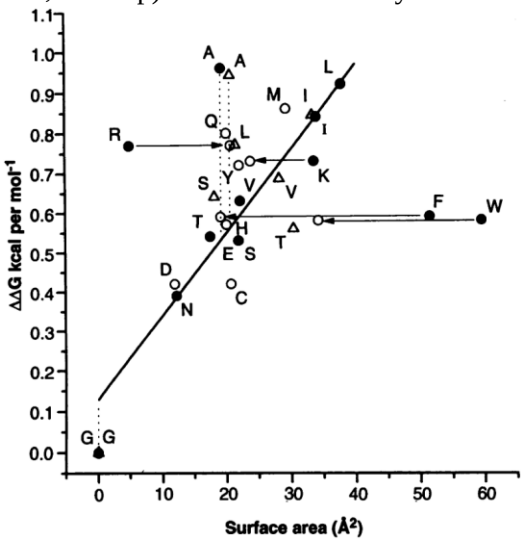
- Analysis of known protein structures has revealed different propensities of different AAs (and AA sequences) to occur in an α helix or β sheet, described either as P_α/P_β (the probability of finding an AA in α helix/β sheet divided by the probability of finding an average AA in that structure) or as ΔΔG (the free energy change during formation of the helix/sheet relative to that of the Gly-substituted version of the protein).

Residue	P_α	P_β
Ala	1.42	0.83
Arg	0.98	0.93
Asn	0.67	0.89
Asp	1.01	0.54
Cys	0.70	1.19
Gln	1.11	1.10
Glu	1.51	0.37
Gly	0.57	0.75
His	1.00	0.87
Ile	1.08	1.60
Leu	1.21	1.30
Lys	1.16	0.74
Met	1.45	1.05
Phe	1.13	1.38
Pro	0.57	0.55
Ser	0.77	0.75
Thr	0.83	1.19
Trp	1.08	1.37
Tyr	0.69	1.47
Val	1.06	1.70

(Source: Chou, P.Y. and Fasman, G.D., Annu. Rev. Biochem. 47, 258 (1978))

Protein stability

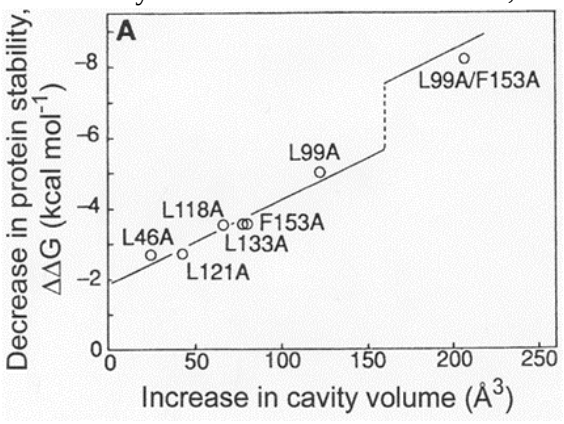
- Hydrophobic stabilisation
  It is generally accepted that hydrophobic effect is the major factor in stabilising the folded structures of globular proteins. Within an α helix, the side chain atoms of one residue can contact the backbone or other side chains and may thereby be (partly) removed from contact with the solvent.
  To test whether hydrophobic effects are related to α helix propensity, the ΔΔG values (relative to Gly) of a series of T4 lysozymes with different AAs substituted at site 44 were plotted against the surface area of the side chain of residue 44 that is buried during helix formation. The values for some of the 20 AAs (Asn, Glu, Ser, Thr, Val, Lys, Ile, Leu) fall approximately on a straight line, with a slope 19 kcal mol^{-1} $Å^{-2}$. The correspondence between this value and the commonly accepted range of 20 to 30 kcal mol^{-1} $Å^{-2}$ for hydrophobic stabilization suggests that hydrophobic effect is a major factor in the determination of the helix propensity of these amino acids. There are structural reasons why the values for other amino acids (Gly, Ala, Arg, Phe, and Trp) deviate substantially from the straight line.



(Source: Blaber, M., Zhang, X.-J. & Matthews, B.W., 1993. Structural Basis of Amino Acid α Helix Propensity. Science, 260(5114), pp.1637–1640.)

In another experiment, six "cavity-creating" mutants (L46A, L99A, L118A, L121A, L133A, and F153A) were constructed within the hydrophobic core of phage T4 lysozyme. Removal of the wild-type side chain allowed some of the surrounding atoms to move toward the vacated space but a cavity always remained, which ranged in volume from 24 $Å^3$ for L46A to 150 $Å^3$ for L99A.

Change in the free energy of unfolding (ΔΔG) of mutant lysozymes relative to wild type was plotted as a function of the cavity volume created by the amino acid substitution(s). The result suggests that the decrease in protein stability associated with a Leu→Ala replacement consists of a constant energy term of 1.9 kcal mol^{-1} plus a second energy term that depends on the size of the cavity created by the substitution. The magnitude of the constant energy term agrees with values of 1.7 to 1.9 kcal mol^{-1} for the difference in hydrophobicity of Leu and Ala estimated by transfer from water to ethanol, octanol, or N-methylacetamide.



(Source: Eriksson, A. et al., 1992. Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect. Science, 255(5041), pp.178–183.)

- Packing

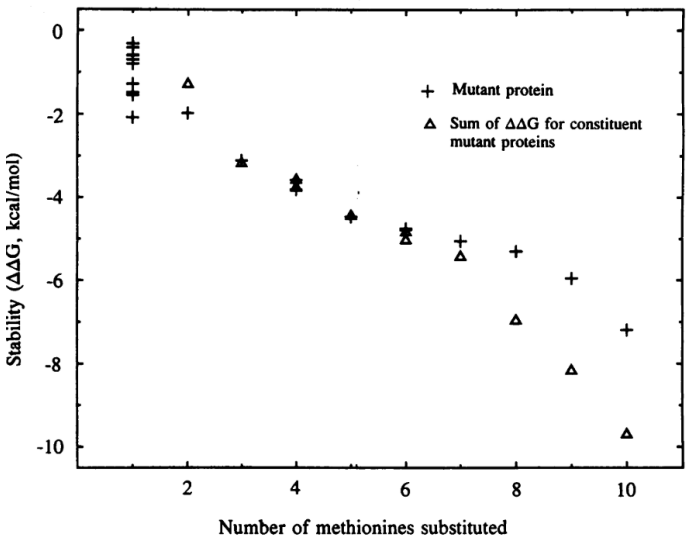
The core of a globular protein consists of buried, primarily hydrophobic AAs, whose side chains are tightly packed. This has led to the "jigsaw puzzle" model of protein folding, which says that the size and shape of the nonpolar amino acids within the core may constrain or define the overall protein fold.

To test this model, up to 10 adjacent residues within the core of T4 lysozyme were replaced by Met. All the variants were found to possess native-like properties and to fold cooperatively with progressively reduced stability.

Table 1. Activity and stability of methionine-substituted lysozymes

Mutant	Activity (%)	ΔT_m (°C)	$\Delta H(T_m)$ (kcal/mol)	ΔH(ref) (kcal/mol)	$\Delta\Delta G$ (kcal/mol)
WT*	100		130	115	—
I78M	70	−3.7	117	111	−1.5
L84M	104	−4.9	110	108	−1.9
L91M	96	−2.0	125	115	−0.8
L99M†	90	−1.3	134	122	−0.4
I100M	105	−4.5	125	121	−1.6
V103M	70	−3.1	117	109	−1.2
L118M	98	−1.8	130	119	−0.7
L121M	87	−2.1	129	119	−0.8
L133M	106	−1.0	128	115	−0.4
F153M†	87	−1.6	128	116	−0.6
7-Met‡	43	−14.5	96	117	−5.0
10-Met‡	≃20	−25	42	88	−7.3

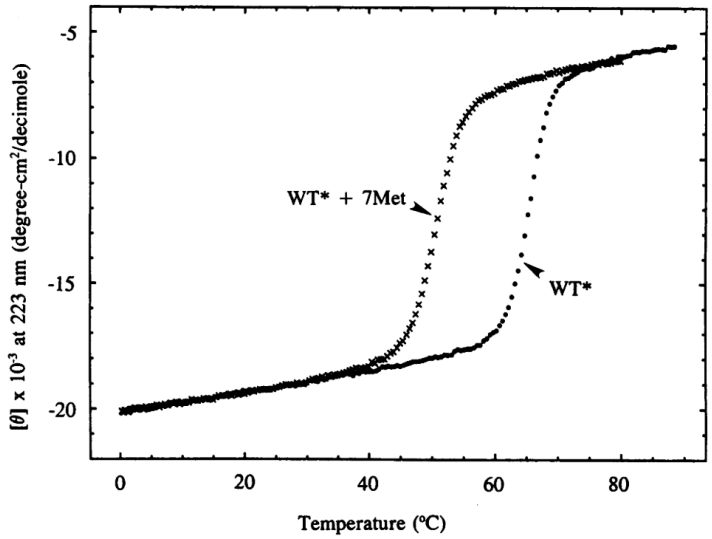
As more and more methionines are introduced into the core of the protein, the overall stability decreases. However, when six or more methionines are substituted, the loss of stability is somewhat less than the sum of the constituent single replacements with the discrepancy increasing to a maximum of 2.5 kcal/mol for the 10-methionine construct.



This indicates that there is some relaxation in the polymethionine protein that either introduces new, favourable, interactions or relieves some of the strain associated with single substitutions. The loss in protein stability is understandable. For each methionine replacement there is a reduction in the solvent transfer free energy (about 0.6 kcal/mol for Leu to Met). Also the side chain of methionine has more degrees of freedom than do other hydrophobic core amino acids. Each Met-to-Leu replacement at a restricted, internal site is predicted to have an entropy cost of about 0.8 kcal/mol.

Taken together, these two factors are expected to reduce the stability of the 7-Met mutant by about 10 kcal/mol relative to wild type, but the actual loss in stability for the 7-Met mutant is only 5.0 kcal/mol.

Also, the structure of the 7-Met variant has been shown, crystallographically, to be similar to wildtype and to maintain a well ordered core, which is also indicated by comparison of the thermal unfolding transition of the 7-Met variant with that of wildtype T4 lysozyme.

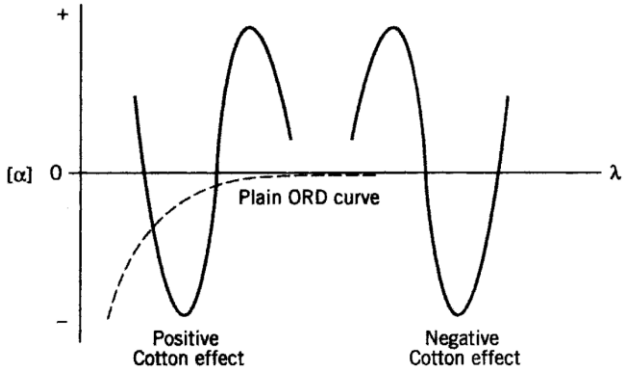


The interaction between the core residues is, therefore, not strictly comparable with the precise spatial complementarity of the pieces of a jigsaw puzzle. Rather, a certain amount of “give and take” in forming the core structure is permitted.

(Source: N C Gassner, W A Baase & B W Matthews, 1996. A test of the "jigsaw puzzle" model for protein folding by multiple methionine substitutions within the core of T4 lysozyme. Proceedings of the National Academy of Sciences of the United States of America, 93(22), pp.12155–8.)

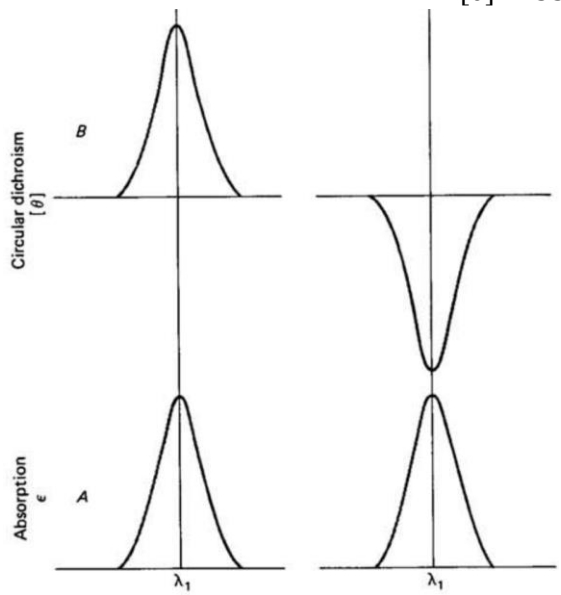
Optical rotatory dispersion (ORD) and circular dichroism (CD)

- The intrinsic spectral properties of AA side chains are barely affected by the conformation of the backbone.
  On the other hand, the polypeptide backbone absorbs light at wavelengths of less than 240 nm and has numerous groups that contribute to vibrational spectra, both are affected by the conformation of the backbone.
- Optical rotatory dispersion (ORD) is the wavelength-dependent rotation of plane-polarised light going through chiral materials.
  In a wavelength region where the molecule does not absorb light, the rotation varies gradually with wavelength. In a wavelength region where the light is absorbed, the absolute magnitude of rotation ([α], in degrees) varies rapidly with wavelength (λ) and crosses zero at absorption maximum. This phenomenon is called Cotton effect.



- Circular dichroism (CD) describes the differential absorption of left- and right-circularly polarised light by chiral materials.
  CD is usually measured as molar ellipticity ([θ]) or difference in molar absorptivity (Δε). The relationship between them is given by:

$$[\theta] = 3300° \cdot \Delta\varepsilon$$



- Both ORD and CD are sensitive to conformational changes and chemical transformation. ORD has the following advantages over CD: (1) It is easier to visualize the Cotton effect with ORD because of the three distinct points in the ORD curve: the peak, the crossover, and the trough. (2) An optically active compound that does not show the band in the wavelength range of interest in the absorption spectrum will not show a CD curve but will show a plain ORD curve.

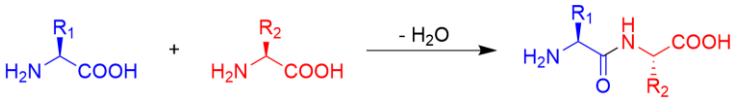
CD, on the other hand, possesses an intrinsic discreteness and is a more sensitive tool in examining the environmental effect on the conformation of macromolecules.

Chemical synthesis of peptides 多肽的化学合成

(26.02.2018)

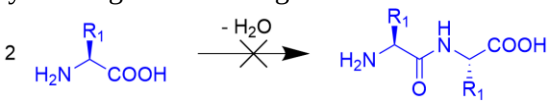
Production of peptides/proteins

- Why would one want to produce peptides/proteins?
  → Prepare the native variant for study and characterisation (e.g. crystal structure)
  → Applications in medicine etc.
- Standard reaction: formation of amide bond between two AAs through condensation



There are several chemical challenges:

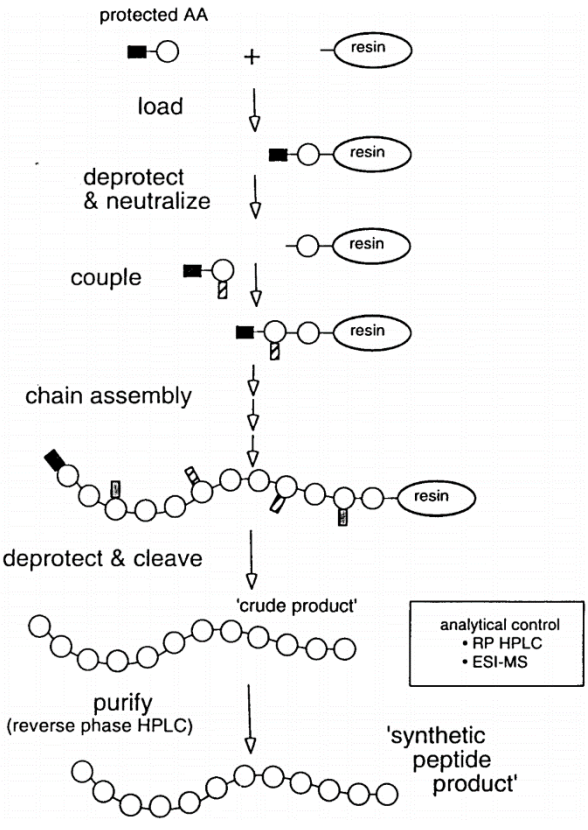
  - Chemoselectivity: prevent unwanted reactions, such as condensation of an AA with itself or reactions involving side chains (e.g. the amino group in the side chain of lysine might also undergo condensation with a carboxyl group).



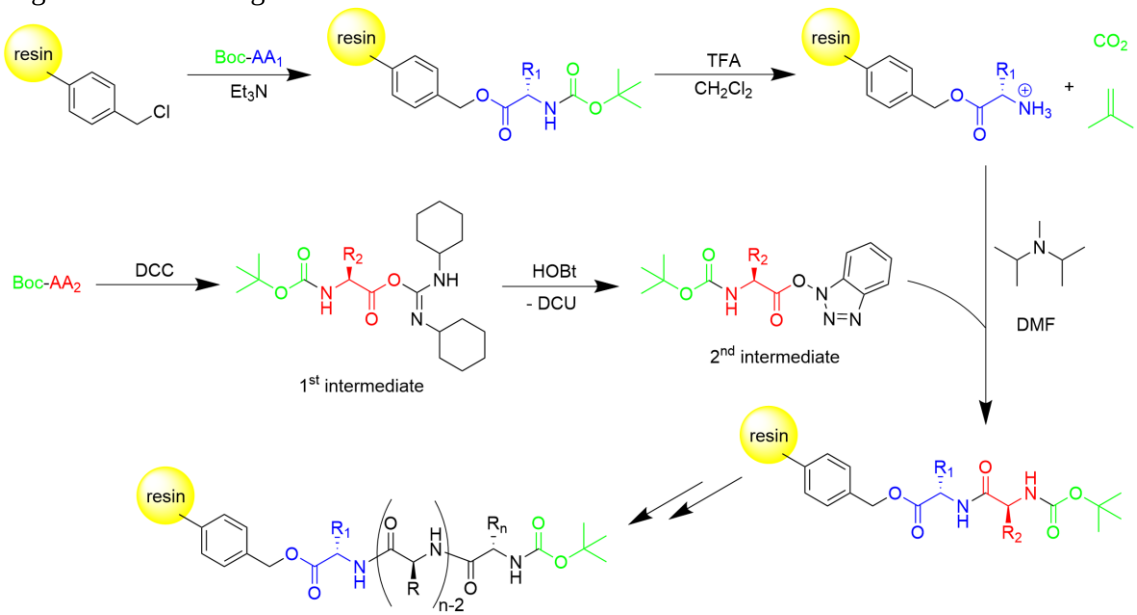
    Solution: use "temporary" protecting groups (e.g. Boc, Fmoc) for the N-terminus and "permanent" protecting groups for side chains.
  - Activation: make amino group and carboxyl group reactive (since both are charged under biological pH and thus do not readily undergo condensation reactions).
    Solution: use coupling and co-coupling reagents (most widely used are DCC and HOBt).
  - Assembly: synthesise long peptide chains with high fidelity.
    Solutions: solution phase (isolation after each step, arduous and limited to short chains) or solid phase (applicable to long chains, revolutionary, see below).

Solid phase peptide synthesis (SPPS)

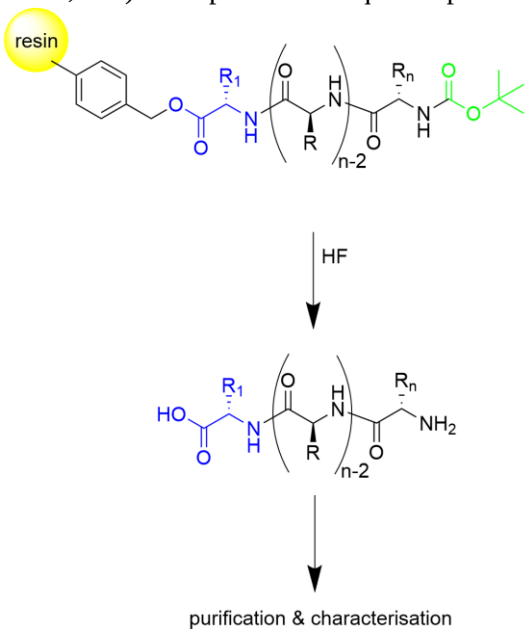
- General procedure
  The first step in SPPS is the attachment of the first AA (whose amino group is protected) to a solid support (resin). The amino group is deprotected to allow coupling with the carboxyl group of the second AA (which is also protected on its amino group). After that, excess reagents are washed away and the procedure is repeated.
- How does it work?
  The first protected AA is added to a chloromethyl group on resin under basic condition (typically Et_3N). After loading, the Boc protecting group is removed under mild acidic condition (typically TFA in DCM).
  The second AA is activated by DCC and HOBt. In principle, one can use only DCC, but the high reactivity of the first intermediate leads to racemisation of the stereocentre at the α-position. Also, it turns out that the acyl group can migrate to one of the nitrogens in DCC, resulting in a very



stable amide. These problems are solved by adding the co-coupling reagent HOBt, which reacts rapidly with the first intermediate and forms a less reactive second intermediate. The second intermediate of the second AA is added in large excess to the resin that has been loaded with the first AA. The target dipeptide is then formed under basic condition (typically diisopropylmethylamine in DMF). The solid support is then collected and washed extensively to get rid of remaining second AA.

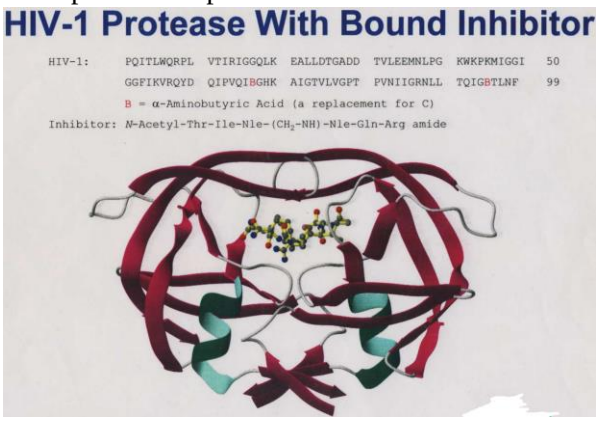


All new AAs can be added in the same manner as the second AA, resulting in a polypeptide chain with defined length (i.e. number of AAs) and sequence (i.e. type and order of AAs). By adding HF, the bond to resin is broken and all side chains are deprotected, releasing the target polypeptide, which can then be purified by HPLC and characterised by various methods (e.g. NMR, MS). This procedure requires special machinery and trained personnel.



- SPPS is one of the most highly optimised chemical methods for peptide synthesis. Advantages of SPPS:
  - High yield and quality due to large excess of the AA to be added in each step
  - Simple purification
  - Routine synthesis of long (⩾30 AAs) polypeptides
  - Synthesis of unnatural proteins (e.g. D-proteins) that cannot be produced biologically

- Example: HIV-1 protease



HIV-1 protease (PR) is a retroviral aspartyl protease (retropepsin) that cleaves newly synthesised polyproteins (namely Gag and Gag-Pol) at nine cleavage sites to create the mature protein components of an infectious HIV virion. Without effective HIV protease, HIV virions remain uninfectious.

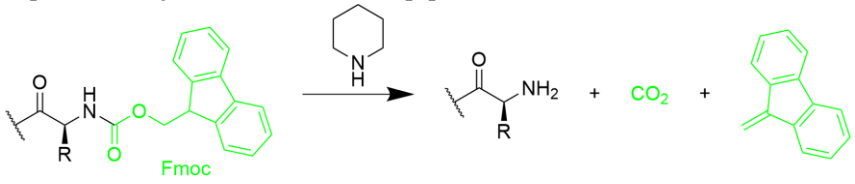
The enzyme contains two identical subunits, each consisting of 99 AAs.

Challenges:

- Low yield
  The synthesis requires 196 steps (coupling and deprotection for each AA). Even if the yield for each step was 98%, the overall yield would be only around 2%.
- Difficult purification
  Peptides that contain errors might be very similar to the correct one (e.g. it would be hard to distinguish between the target peptide and a truncation product that misses only one AA).

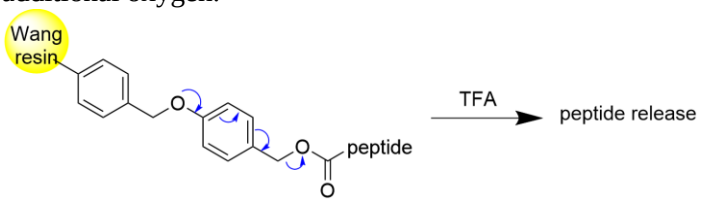
Solutions:

- Capping (permanent protection of unreacted amino groups) with acetic anhydride
  Add large excess of acetic anhydride after each coupling step to make sure that all amine left in the solution becomes stable amide.
- Multiple couplings
  Carry out the coupling step multiple times to bring the yield close to 100%.
- Alternative chemistry
  Example: use Fmoc instead of Boc as the temporary protecting group. Fmoc can be deprotected by a mild base such a piperidine.



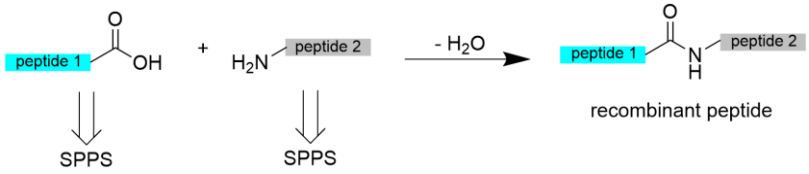
The Fmoc approach is a bit more expensive and less optimised compared to the Boc approach. However, it allows deprotection under mild conditions and easy release of the polypeptide chain from resin.

One of the commonly used resin is the so-called Wang resin, which contains an additional oxygen.

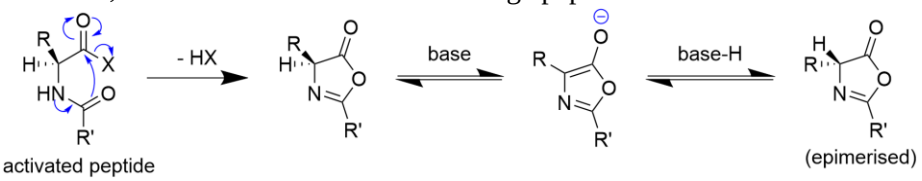


Towards larger proteins: convergent strategy / fragment coupling

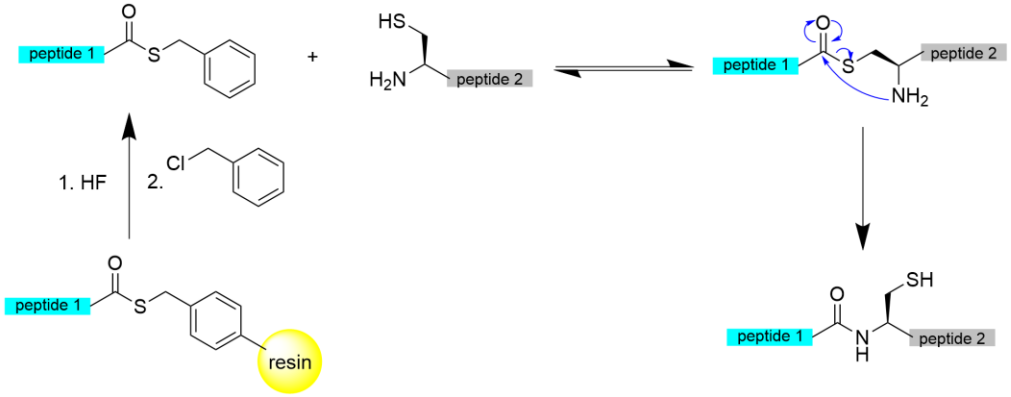
- Idea
  Synthesise different parts of the peptide separately and then condense them.



- Challenges:
  - Protected polypeptides tend to be poorly soluble
  - Chemoselective activation of the C-terminus is hard
  - Epimerisation of the activated peptide
    In SPPS, we can mitigate this problem by quickly trapping the DCC ester with HOBt.
    However, this is much harder in case of large peptides.



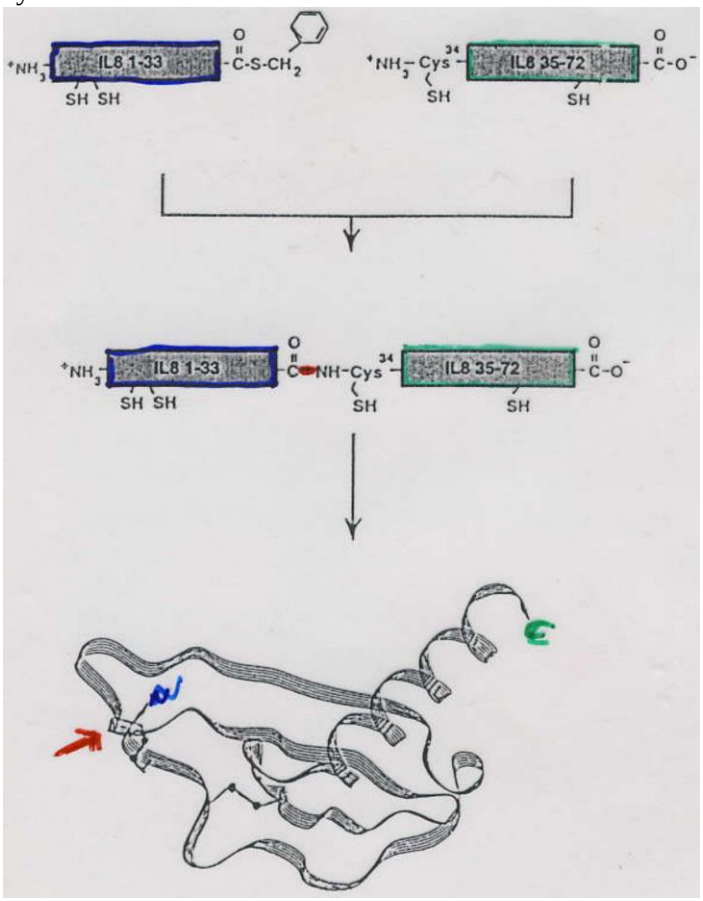
- Solutions:
  - Use an enzyme for ligation (not a topic of this course)
  - Native chemical ligation (NCL)
    In NCL, a peptide synthesised by SPPS is activated chemoselectively on its C-terminus as a thioester.
    The thioester is prepared on a resin by cleavage with HF and alkylation with chlorotoluene. The second peptide has a cysteine at its N-terminus, which reacts efficiently and chemoselectively with the thioester (transthioesterification). The resulted thioester undergoes an intramolecular transfer of the acyl group from sulphur to nitrogen to give the native chemical amide bond.
    Note that other cysteines present in the peptide need not to be protected because the reaction is reversible and only the terminal cysteine (which is adjacent to a free amino group) can be converted into a final product.



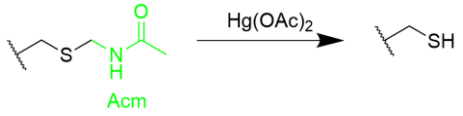
- Advantages of convergent strategy:
  - Deprotected peptides can be used (meaning that the reaction can be carried out in aqueous solution)
  - Thioesters are activated but not overactivated, which minimises epimerisation
  - Acceleration possible with benzenethiol (which replaces the poor leaving group with a good leaving group) and guanidinium chloride (which solubilises the peptide and exposes the cysteine residue by denaturation of the peptide).
  - High chemoselectivity

- Easy purification since fragments and full-length peptides have very different properties

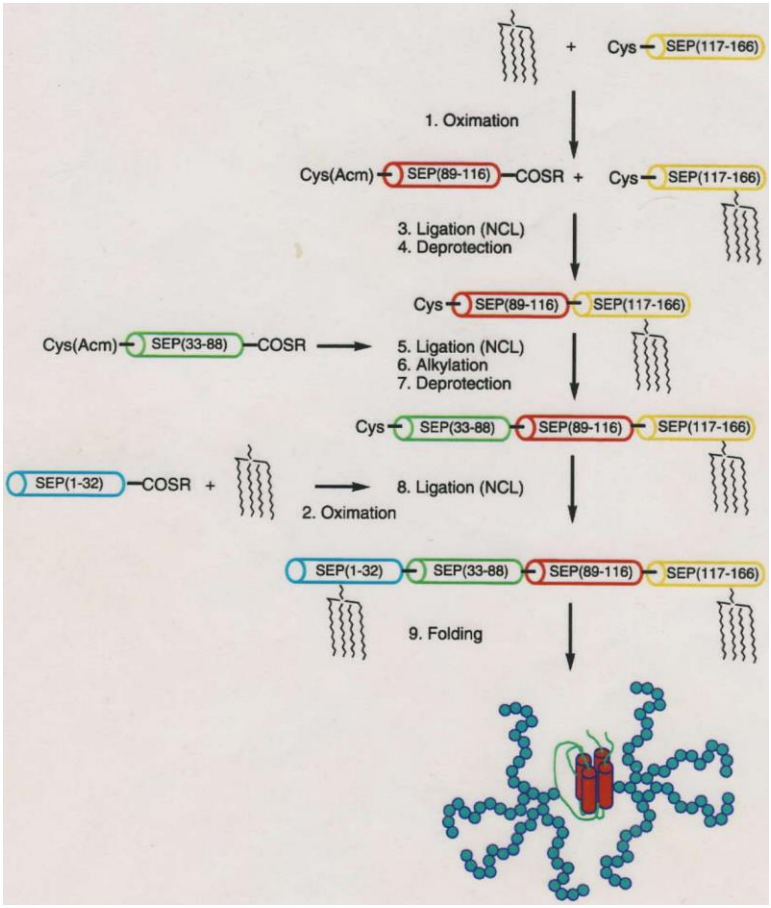
- Example 1: synthesis of interleukin 8 (IL-8, also known as neutrophil chemotactic factor)
  Interleukin 8 is a chemokine produced by macrophages and other cell types such as epithelial cells, airway smooth muscle cells and endothelial cells. It induces chemotaxis in target cells (primarily neutrophils but also other granulocytes), causing them to migrate toward the site of infection, and induces phagocytosis once they have arrived. IL-8 is also known to be a potent promoter of angiogenesis.
  This 72-AA peptide has a cysteine at position 34. Thus it is made by assembling a 33-AA peptide activated as benzyl thioester and a 38-AA long peptide carrying an N-terminal cysteine.



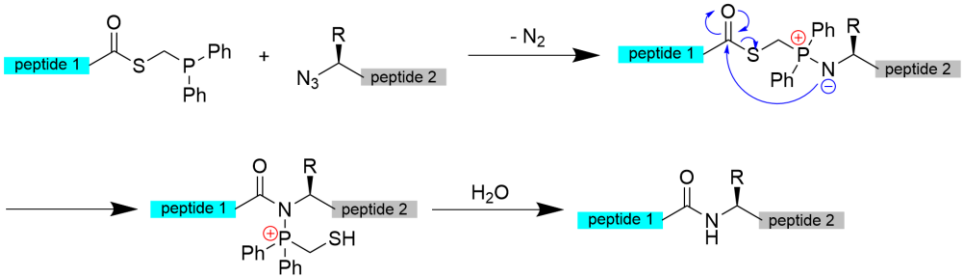
- Example 2: synthesis of EPO (erythropoietin)
  EPO is a hormone that induces red blood cell production (and is sometimes used for doping in endurance sports). It is a four-helix bundle protein containing 166 AAs and heavy glycosylation, a PTM that increases the half-life of the molecule in serum.
  Chemical synthesis of the protein is achieved by assembling four fragments through three ligation steps, starting from the C-terminus. Each new peptide fragment added is selectively activated at its C-terminus and each N-terminal cysteine is protected with an Acm group, which is later removed using mercury acetate.



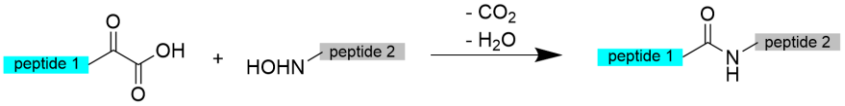
  Instead of glycosylation, a synthetic precision dendrimer is added to protect the protein from immune attack and to extend its lifespan, making a completely unnatural construct. This synthetic version of EPO is called synthetic erythropoiesis protein (SEP). The whole molecule weighs 50825 Da.



- Accessing ligation sites other than Xaa-Cys
  - Staudinger ligation
    The C-terminus is activated as a thioester with a phosphine, which reacts with another polypeptide that has an azide on its N-terminus, resulting in an ylide. This ylide then undergoes acyl transfer and hydrolysis to give the target peptide chain.



  - Ketoacid-hydroxylamine ligation (KH ligation)
    In this approach, the C-terminus is activated as α-ketoacid and the N-terminus as hydroxylamine. The native bond is generated with elimination of CO_2 and water.



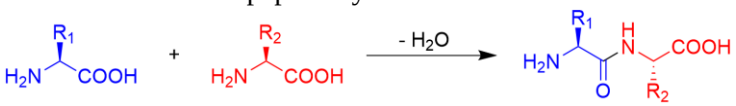
- Conclusion: fragment coupling
  - Allows us to synthesise large proteins using chemical methods
  - Competitive with biosynthetic approaches, especially when non-canonical AAs are needed
  - Same approach applicable to fully unnatural polymers such as β-peptides & -proteins

Microbial non-ribosomal peptide synthesis 微生物中的非核糖体多肽合成

(05.03.2017)

Peptides in microorganisms

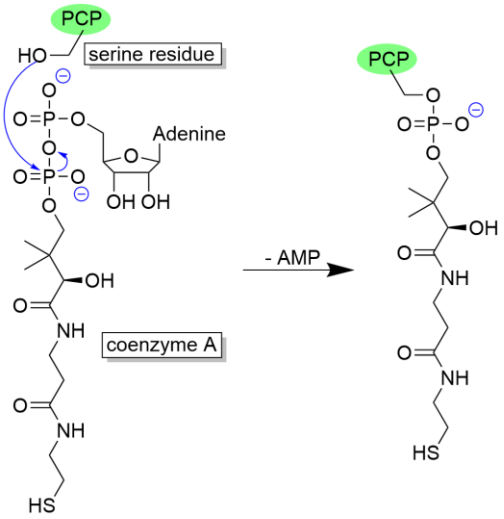
- Peptides are a diverse class of secondary metabolites (organic compounds that are not directly involved in the normal growth, development, or reproduction, the absence of which does not result in immediate death).
- In Microorganisms, peptides have many functions such as defence, signalling. They also serve as antibiotics, immunosuppressants, pigments, toxins etc.
- Microbial peptides are typically:
  - 3-25 AAs long (can be either linear or cyclic)
  - can be made of more than 200 different building blocks
  - possess often extensive modifications (e.g. cyclisation, oxidation, reduction, N-methylation)
- Standard reaction in peptide synthesis:



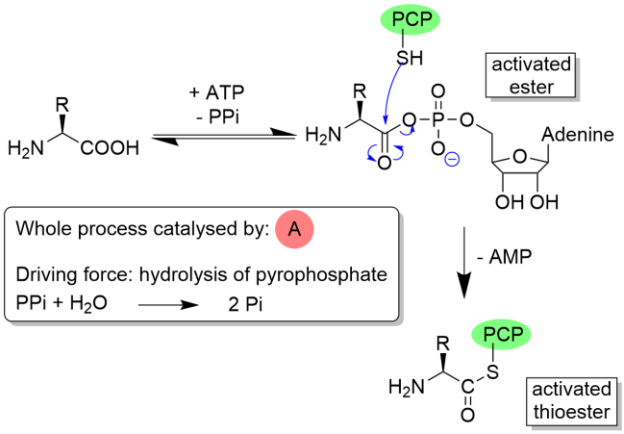
  Cells have to deal with several challenges:
  - activation (how to trigger the reaction)
  - elongation (how to make the reaction proceed)
  - direction (to which end should a new AA be added)
  - sequence (which AA should be added)
  - length (where to stop the reaction)

Assembly-line biosynthesis

- Assembly line biosynthesis is a peptide synthesis pathway found in microorganisms. Instead of ribosomes, this process is catalysed by very large, multifunctional mega-enzymes called non-ribosomal peptide synthetase (NRPS).
- The chemical "logic" of NRPS:
  - An NRPS consists of so-called modules, each being able to covalently attach a specific AA to the polypeptide chain.
  - Growing polypeptide is transferred from one module to the next.
  - Length & sequence are determined by the number & order of NRPS modules.
- Domains
  Each module of NRPS consists of several domains, which carry out different functions.
  The most important domains are:
  - Peptide carrier protein (PCP), ~80 AAs
    PCP is the place where the polypeptide chain under construction is attached to. Before being linked to the polypeptide, PCP has to be activated through pantetheinylation on a serine residue. Coenzyme A is needed for this reaction as pantetheine donor.

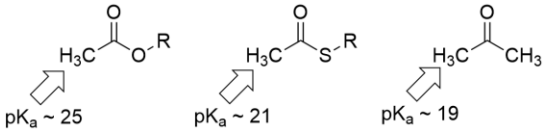


- Adenylation domain (A), ~500 AAs
  The A domain serves as a "gate keeper". An AA is activated through the following reaction with activated PCP, under catalysis of adenylation domain (A).

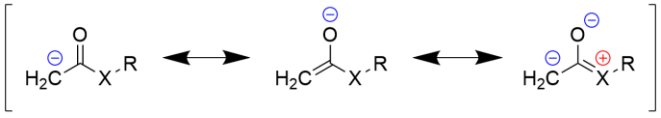


ATP can be seen as a biological equivalent to DCC.
The first reaction is reversible. However, hydrolysis of pyrophosphate (PPi) brings the balance to the right side. The second reaction essentially converts an activated ester into a "less activated" thioester.
Compared to ester, thioester has several advantages:

- Thioester is activated, but still reasonably stable compared to the anhydride with phosphate.
  The reactivity of thioester is reflected by the pK_a of alcohol (typically 15-16) and thiol (typically 9-10), which makes thiol a much better leaving group than alcohol.
  The stability of thioester, on the other hand, extends its lifespan and thus allows enzyme-catalysed reactions to take place.
- Thioester is activated toward aminolysis
  For hydrolysis, thioester reacts roughly twice as fast as ester. However, when it comes to aminolysis (attack of the ester/thioester by amine), thioester can react 1000 times faster. This means the material can be generated and kept relatively stable in aqueous buffer but react very fast when an amine (on an amino acid) is present at a proper position.
- Thioester stabilises adjacent carbanions

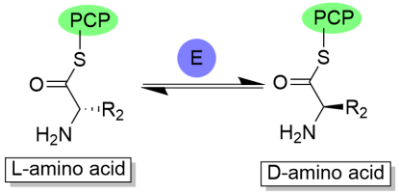


This can be explained by the ability of carbonyl groups to stabilise adjacent carbanions through resonance. In ester, however, the oxygen next to the carbonyl group is electron-donating, thus compensating the electron-withdrawing property of the carbonyl group. This effect is much weaker in thioester (since sulphur has a larger radius and doesn't readily form π bonds).

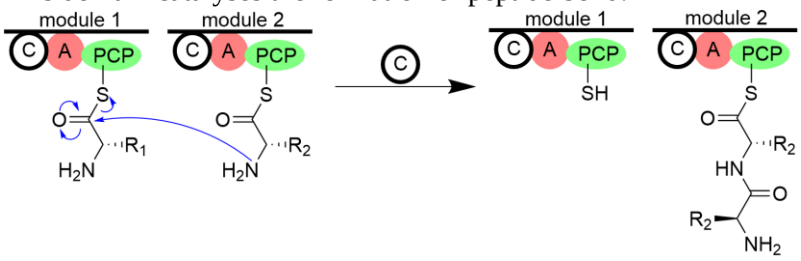


This resonance structure leads to loss of chirality on the α carbon, which is utilised by the epimerase domain to convert the amino acid between L- and D-forms (see below).

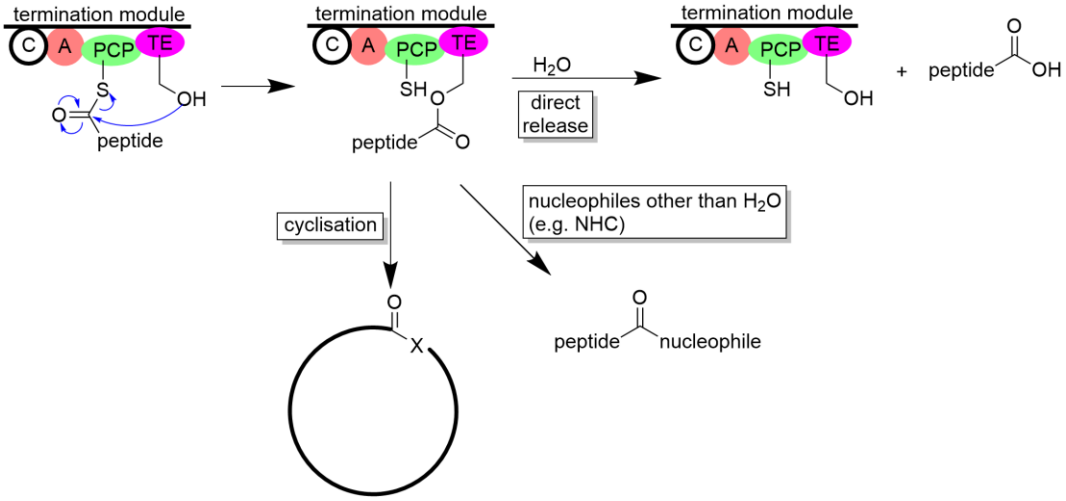
- Epimerase (E), ~250 AAs
  Epimerase changes an AA from L-form to D-form.



  This reaction is reversible. However, the downstream C domain is selective for D-form, thus pushing the equilibrium to the right side.
- Condensation domain (C), ~450 AAs
  This domain catalyses the formation of peptide bond.



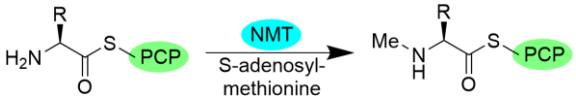
- Termination domain (thioesterase, TE)
  Completed peptides are transferred to a serine residue on the TE domain, where it is later released either by nucleophilic attack.
  Nucleophiles that can be used for the nucleophilic attack are not limited to H_2O. Many different nucleophiles have been observed in this process. Cyclisation occurs when the nucleophile comes from a side chain inside the peptide.



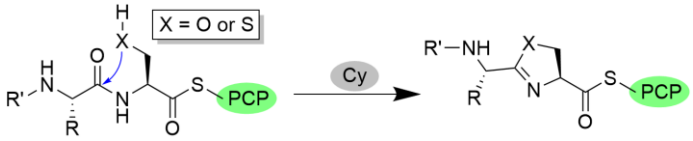
There are also domains that are responsible for additional modifications:

- Methylation domain (N-methyltransferase, NMT)

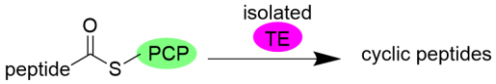
N-methylation is a valuable strategy to control protein conformation and stability.
NMT transfers a methyl group to the amine of an AA attached to PCP, with the cofactor (methyl group donor) S-adenosylmethionine.
The methylated amine serves as a nucleophile in the following condensation reaction, generating an N-methylated peptide.



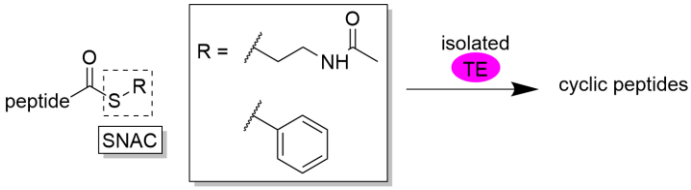
- Cyclisation domain (cyclase, Cy)
  In case the side chain contains a hydroxyl or thiol group (such as in Ser, Thr and Cys), it can be positioned by the cyclase to attack the upstream carbonyl group to give an acetal. This acetal then loses water and forms a heterocycle, which can undergo further modifications such as oxidation.



- Utilisation of the TE domain in the lab
  TE domains are typically promiscuous, meaning that they only recognise a few residues of the whole peptide and the rest is quite flexible. As a result, TE can be used outside the NRPS as very versatile catalysts for making cyclic peptides.



The PCP domain in the reaction above can be replaced by some easier-to-prepare groups that mimic the pantetheine arm of PCP, such as SNAC.



It is also possible to prepare the peptide directly on a resin (as normal esters), such as PEGA resin (polyacrylamide PEG copolymer).

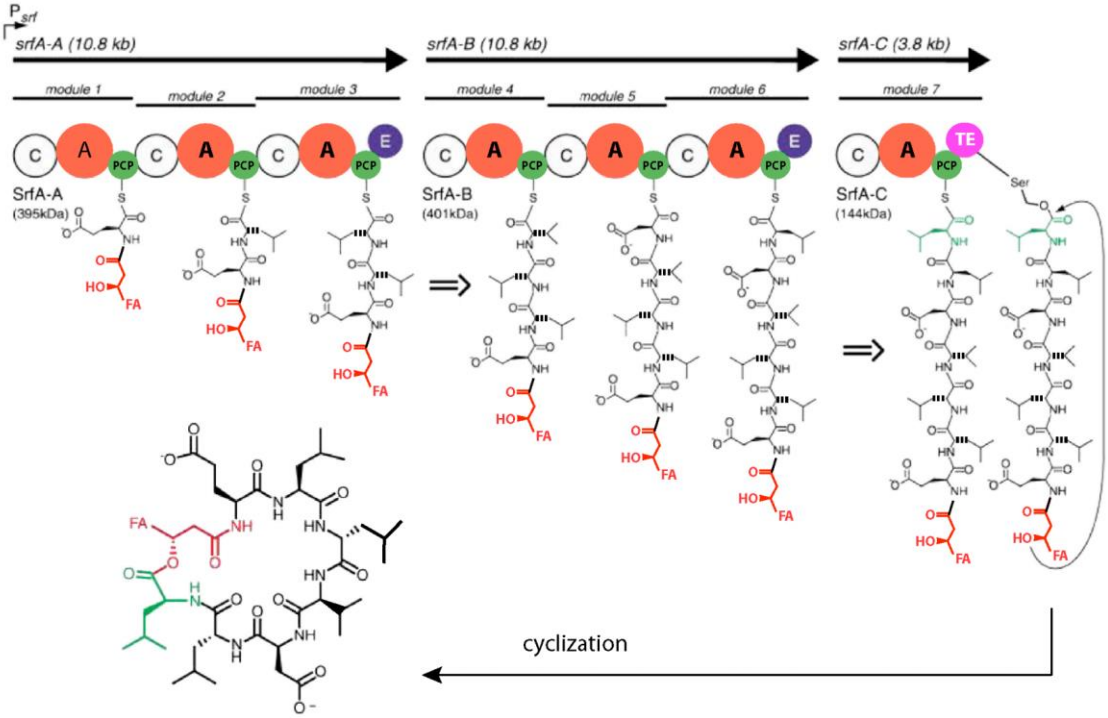
This technique can be used to prepare peptidomimetics and to construct peptide libraries.

- Modules
  As mentioned above, a module is needed for each building block (i.e. each amino acid).
  Each module comprises several domains:
  - Initiation module
    (C if needed +) A + PCP
  - Elongation module
    C + A + PCP (+ E if needed)
  - Termination module
    C + A + PCP + TE
- What is the evolutionary advantage of having such complicated assembly-lines?
  These assembly-lines are able to create new bioactive substances by mixing & matching different domains and modules in a combinatorial fashion, without being limited by genetic code.
  People can also mimic this in laboratory.
- What remains unknown: structural and dynamic properties of NRPS

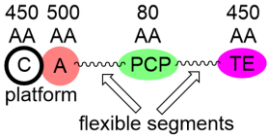
Based on X-ray and NMR data, we already have the crystal structure for many individual domains. However, we still don't know how these domain work together and structural data for modules are emerging slowly.
The termination module (module 7) of surfactin A synthetase is the first one to be structurally characterised (see next section).

Example of NRPS: Surfactin A synthetase

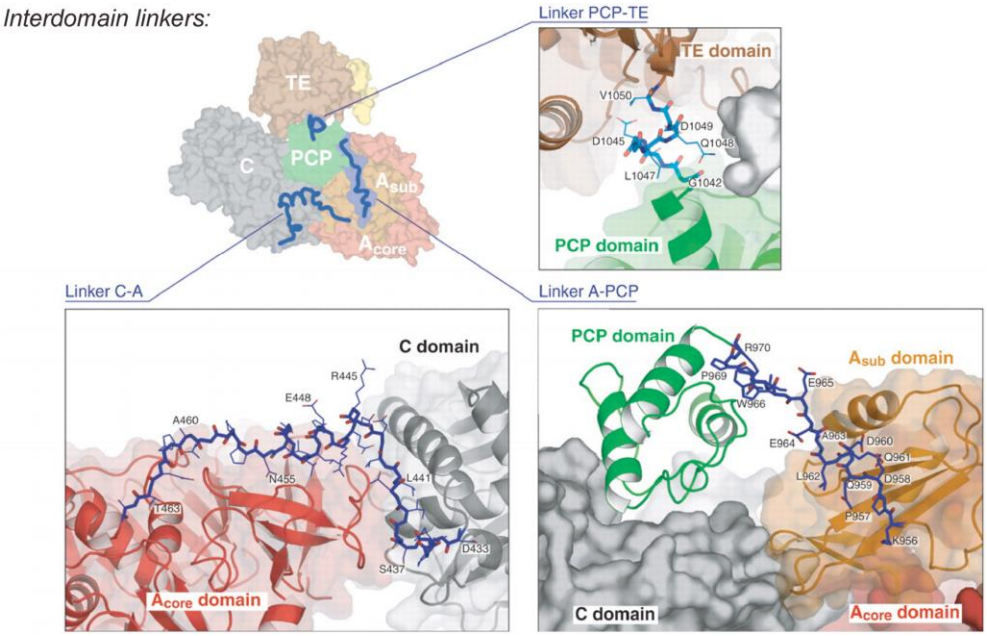
- Surfactin is a very powerful surfactant (compound that lowers surface tension between two liquids) commonly used as an antibiotic. Its structure consists of a peptide loop of seven AAs and a 13-15 carbons long fatty acid chain (FA-Glu-Leu-D-Leu-Val-Asp-D-Leu-Leu).
- The NRPS responsible for the synthesis of surfactin A has 7 modules, each adding one AA to the structure.
  Noteworthy:
  - The starting point is an FA-acylated glutamate.
  - The 3rd and 6th module have E domains to make D-leucine.
  - Upon termination, a hydroxyl group in the FA part serves as nucleophile and attacks the C-terminus to form a lactone.



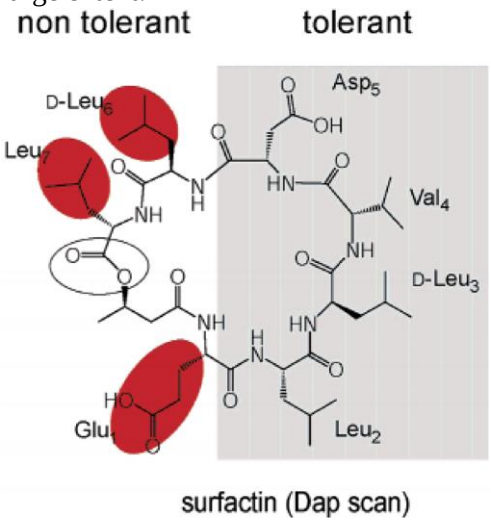
- Structure of the termination module (module 7)
  The whole module contains 1274 AAs, including four domains (PCP, A, C, TE).
  The A and C domains have a very extensive and intimate interface (more than 1600 Å^2). They form a platform to which the PCP and TE domains are attached. The link between A and PCP and the link between PCP and TE are flexible.



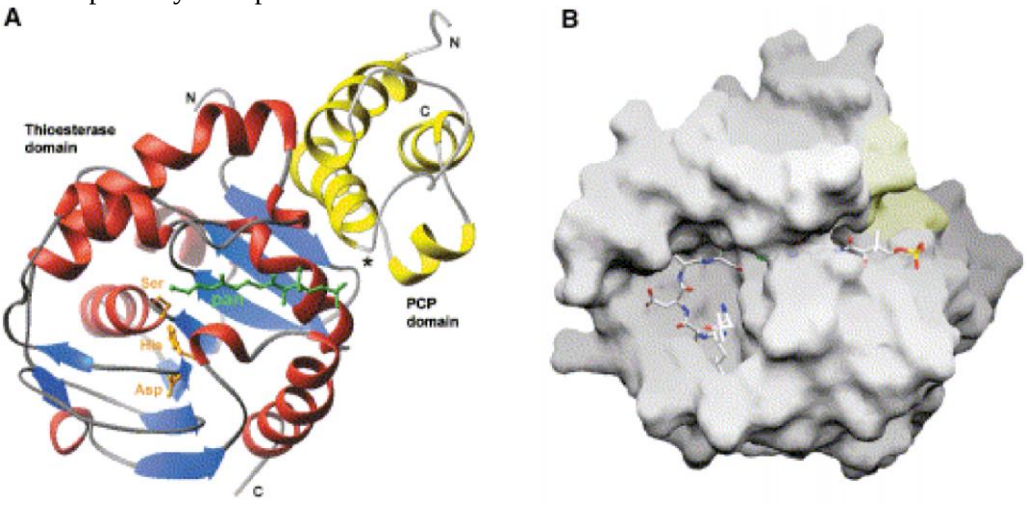
The active sites of C, A and TE domains are quite far away from each other. However, the flexible segments allows the PCP domain to interact with all other domains, so that it can capture the activated AA from A domain, move it to the C domain for condensation and finally transfer it to the TE domain.



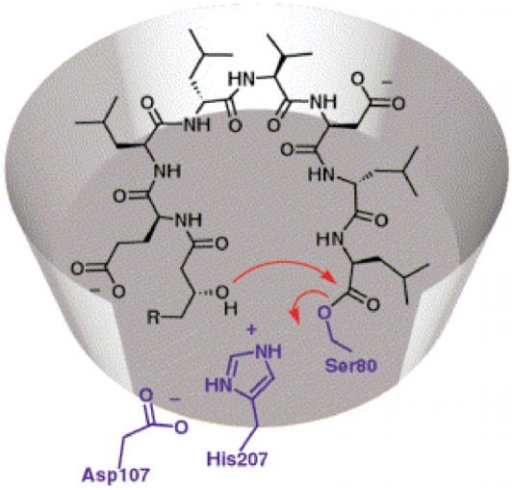
- Three residues out of the seven (Glu1, D-Leu6 and Leu7), flanking the site of reaction, are recognised by the TE domain. Other residues, as well as the FA, can be changed to a quite large extent.



- Crystal structure of the TE domain resembles that of a serine protease.
  The catalytic site consists of a nucleophilic serine, which is activated by a histidine (base) and held in place by an aspartate.



The entire active site is shaped like a bowl. It binds the substrate and allows it to fold around such that only the hydroxyl group in FA part comes into the proximity of the C-terminus, rendering the ring-closure process highly selective.

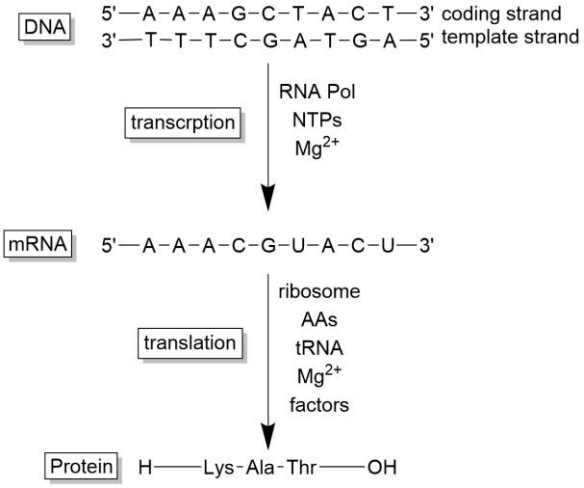


Protein biosynthesis 蛋白质的生物合成

(12.03.2018)

General

- Protein biosynthesis produces biopolymers of defined sequence & length (20 – 34000 AAs). This process is very fast (4 – 20 AA/s, 10^6 fold over spontaneous reaction) and nearly error-free.
- In contrast to SPPS and NRPS, protein biosynthesis is template-directed.
  The process involves transcription of DNA into mRNA using RNA polymerase, various NTPs and magnesium cofactor, followed by translation of the mRNA using ribosome, AAs, tRNAs, magnesium cofactor and various protein factors.



Genetic code

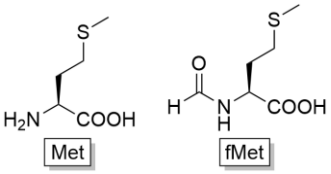
- Genetic code is based on triplets called codons, where three bases together determine the AA that is to be added into the growing polypeptide chain.
- In total, there are $4^3 = 64$ codons, which encode start, finish, and all 20 standard AAs. In fact, the vast majority (61/64) codons are used to encode AAs, which means that the genetic code is highly degenerate (i.e. many "synonyms").

Second Position

First Position (5′ End)	U	C	A	G	Third Position (3′ End)
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	C
	UUA Leu	UCA Ser	UAA* *Stop*	UGA* *Stop*	A
	UUG Leu	UCG Ser	UAG* *Stop*	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC Leu	CCC Pro	CAC His	CGC Arg	C
	CUA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CUG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC Ile	ACC Thr	AAC Asn	AGC Ser	C
	AUA Ile	ACA Thr	AAA Lys	AGA Arg	A
	AUG† Met	ACG Thr	AAG Lys	AGG Arg	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC Val	GCC Ala	GAC Asp	GGC Gly	C
	GUA Val	GCA Ala	GAA Glu	GGA Gly	A
	GUG† Val	GCG Ala	GAG Glu	GGG Gly	G

- There are some interesting patterns in genetic code:
    - NNC is always the same as NNU, whereas NNG is usually the same as NNA.
    - NPyN is typically a hydrophobic AA, whereas NPuN is rather a hydrophilic AA. (Pu = G, A; Py = C, U).

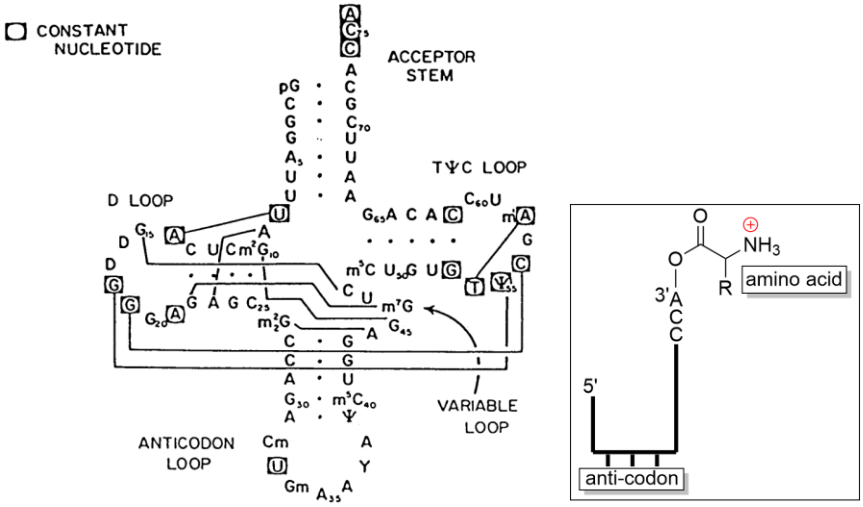
  These patterns give us some insight into the evolution of genetic code.
  We can see that mutation of the third base often has no effect on the sequence of the protein, and mutation of the second base often doesn't lead to change in hydrophilicity. In other words, organisms have a certain resistance to spontaneous mutations.
- Start codon
  The start codon is AUG, which encodes methionine in eukaryotes and N-formylmethionine (fMet) in prokaryotes. The modification of Met to fMet makes the N-terminus less reactive, thus preventing the peptide chain from cyclisation.



  Not all AUGs serve as starting point. In fact, cells also use AUG for the insertion of normal methionine. In prokaryotes, the AUG serving as start point is recognised by ribosome through a purine rich region (Shine-Dalgarno sequence) around its -10 position, which binds to ribosome to initiate translation.
- Stop codon
  There are three stop codons, namely UAA, UAG and UGA, which do not encode any AA. These codons are not recognised by tRNA but by protein release factors (RF).

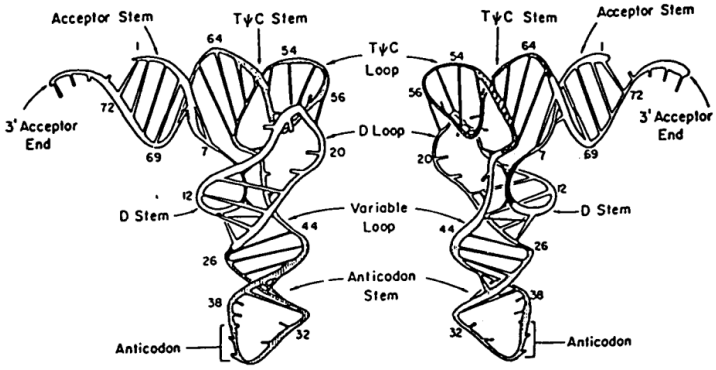
AA decoding

- tRNA
  tRNA is the adaptor molecule that mediates translation of mRNA into protein. The molecule contains ~80 nucleotides, forming a highly conserved 2D structure (clover leaf).
  The anti-codon contained in tRNA recognises codon in template (mRNA) by forming Watson-Crick minihelix, where the third base-pair is a wobble (i.e. there are some distortion that allows flexibility).
  tRNA carries the AA corresponding to the recognised codon on its 3' end (acceptor end).

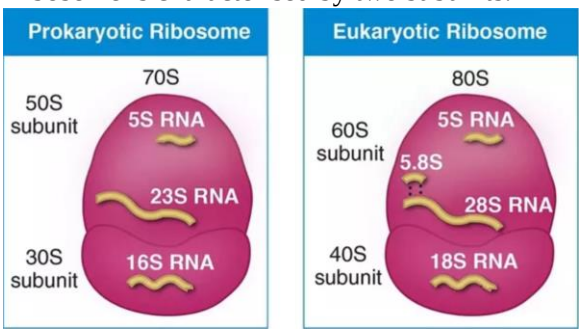


  It turns out that most organisms have fewer than 45 types of tRNA. In other words there aren't enough tRNAs to match all 61 AA-encoding codons. However, all these codons can actually be recognised since the third base-pair is flexible.

- In 3D, tRNA actually folds into an L-shape.

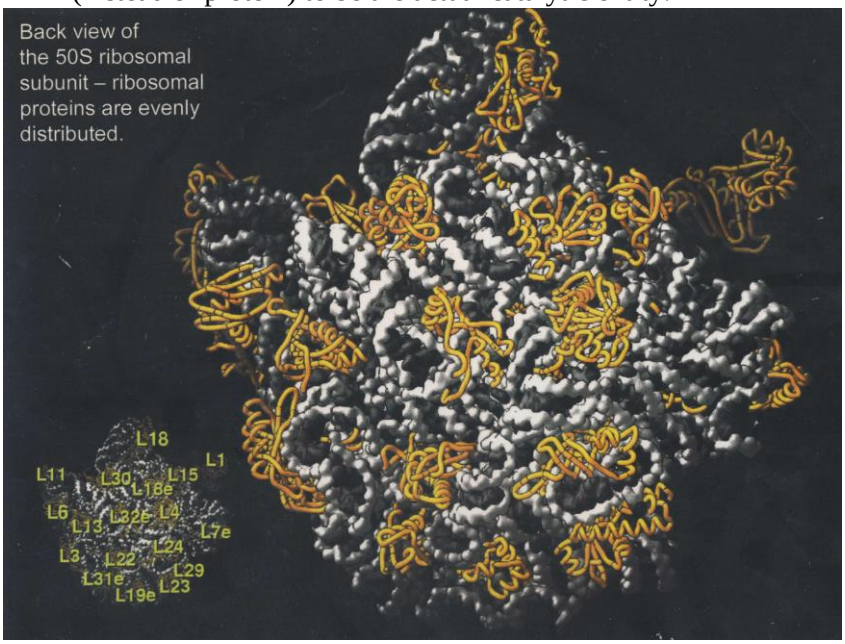


- Ribosome
  Ribosome is the catalyst for translation. This is an ancient molecular machine typically present in ~30000 copies per cell. It contains more than 200000 atoms, of which 2/3 are RNA and 1/3 are protein.
- Ribosome is characterised by two subunits.

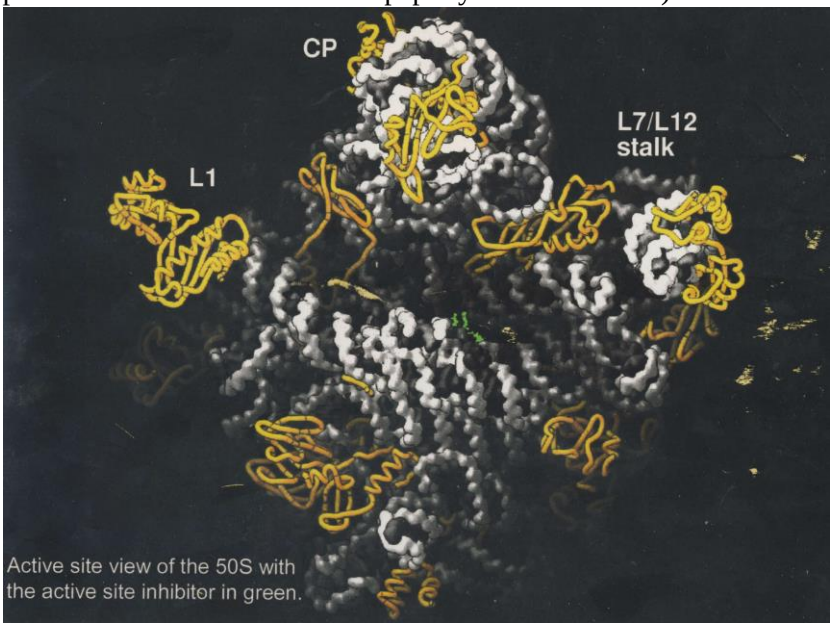


  For prokaryotic ribosome (70S, S = Svedberg), the two subunits are:
  - 50S (large subunit): consists of ~30 proteins, the 23S RNA (~2900 nt) and the 5S RNA (~120 nt), serves peptidyl transfer (formation of peptide bond).
  - 30S (small subunit): consists of ~20 proteins and the 16S RNA (~1540 nt), recognises Shine-Dalgarno sequence, helps decoding (i.e. codon recognition by tRNA).

  Eukaryotic ribosome (80S) is a bit different but basically follows the same blueprint.
- Ribosome is a ribozyme in which the RNA scaffold is supported by structural proteins. X-ray structure showed that the nearest protein to the active site is 18 Å away from it, suggesting RNA (instead of protein) to be the actual catalytic entity.



(The green molecule is the transition state analogue CC-dA-P-puromycin, which indicates the position of the active site where peptidyl transfer occurs)



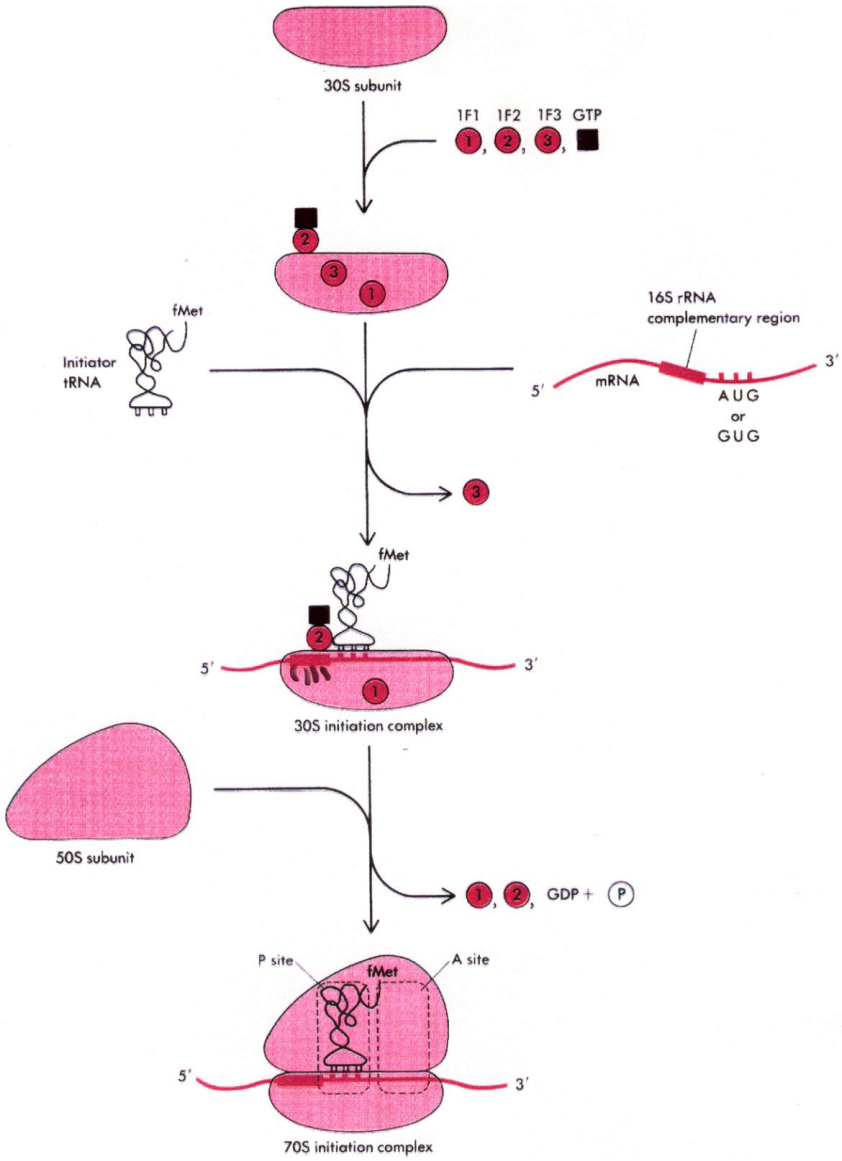
- In contrast to most proteins in the cell, proteins in ribosome are "intrinsically disordered". That is, they contain long, floppy extensions that are presumed to be important for connecting distant parts of rRNA and bringing them together. In addition, these proteins typically carry many positive charges in order to compensate the negative charges carried by RNA.
- The growing polypeptide chain is extruded through a tunnel in the backside of the big subunit (50S), where other molecules can bind and decide the fate of the polypeptide chain.

Translation

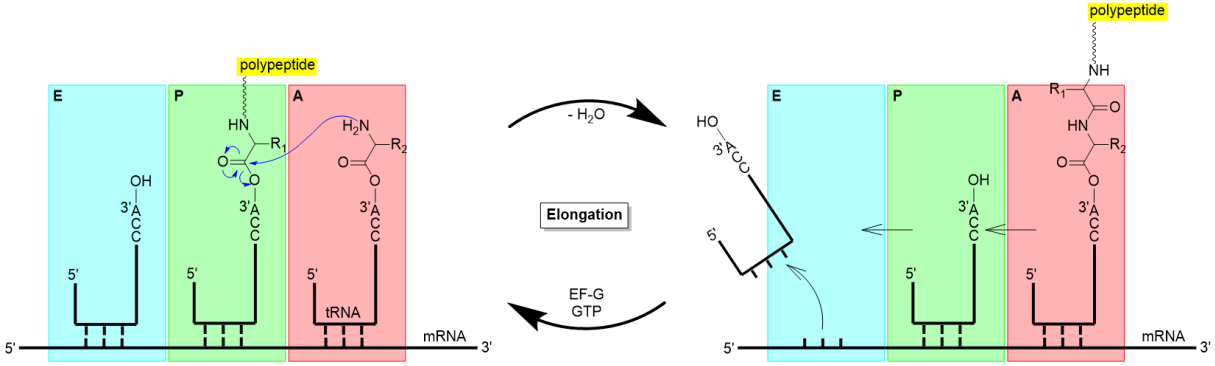
- Direction
  mRNA reading: 5' → 3'
  peptide synthesis: N → C (opposite to SPPS)
- Steps

Step	**Accessory factors**	**Energy consumption**
Initiation	IF-1, IF-2, IF-3	1 GTP (per polypeptide)
Elongation	EF-Tu, EF-Ts, EF-G	2 GTP (per AA)
Termination	RF-1, RF-2, RF-3	1 GTP (per polypeptide)

- Initiation
  Translation begins with binding of the three initiations factors (IF-1, IF-2, IF-3) and a GTP to the small (30S) subunit, forming a complex that binds the initiator tRNA (fMet-$tRNA_F^{Met}$) and an mRNA. mRNA is recognised by a pyrimidine-rich region (anti-Shine-Dalgarno sequence) at the 3' end of 16S rRNA.
  The resulted 30S initiation complex is then able to bind to the big (50S) subunit to give the complete ribosome (70S). This process consumes energy that is generated by hydrolysis of the GTP.
  Once the 70S initiation complex is formed, the initiator tRNA is clearly bound in the P site. However, it is not known whether it initially contacts the A site.

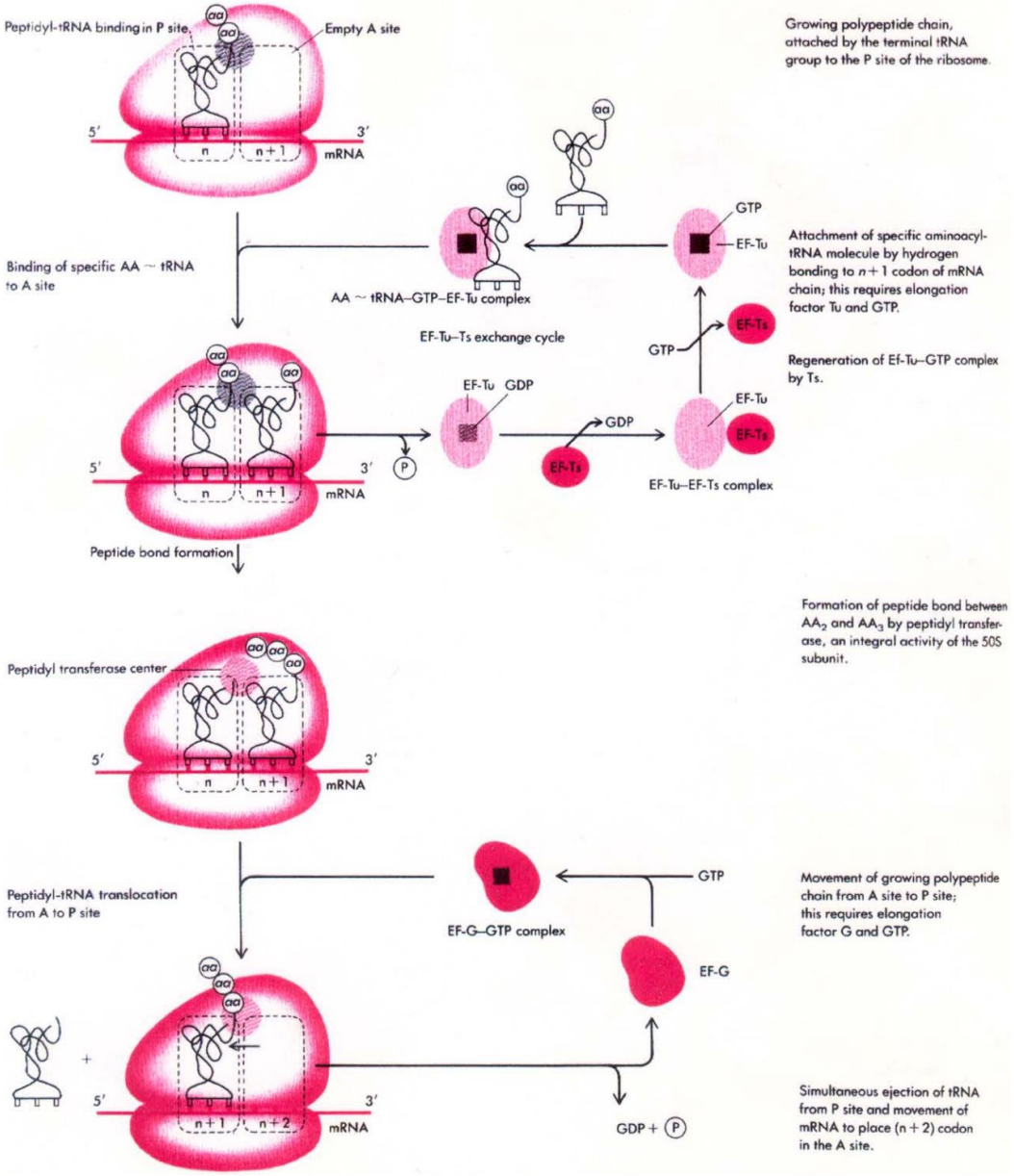


- Elongation
  Crystallography has identified three distinct binding sites in the complete ribosome-mRNA complex:
  - Peptidyl site (P site): the site to which the growing polypeptide chain (in form of peptidyl-tRNA) is attached.
  - Aminoacyl site (A site): the site to which the new amino acid (in form of aminoacyl-tRNA) is recognised (by codon-anticodon interaction) and attached.
  - Exit site (E site): the site at which unloaded tRNA leaves ribosome.



With respect to the mRNA, the three sites are oriented 5’ to 3’ E-P-A (recall that ribosomes move toward the 3' end of mRNA).

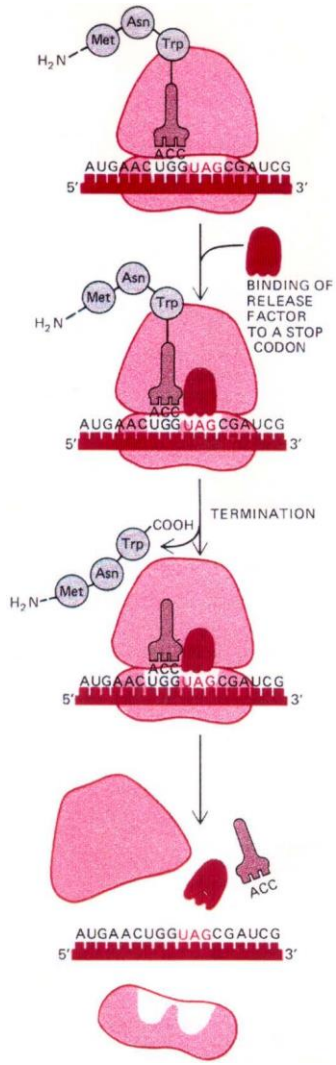
The elongation factor EF-Tu, activated by GTP with the help of EF-Ts, binds to the required aminoacyl-tRNA and brings it into the A site. In the meantime, EF-Tu protects the AA (which is an activated ester) from spontaneous hydrolysis. Once aminoacyl-tRNA enters the A site, EF-Tu is released under GTP hydrolysis.
New peptide bond is formed between the AA in the A site and the C-terminal AA in the P site, during which the growing polypeptide chain is transferred to the A site. Then, the tRNA in the P site (now without AA) is moved to the E site and released from ribosome, and the tRNA in the A site (now charged with the polypeptide chain) takes over the P site.
After the new amino acid is added to the chain, the energy provided by the hydrolysis of a GTP bound to the translocase EF-G moves the ribosome down one codon towards the 3' end.



- Termination

  Once a stop codon is present in the A site, it can be recognised by a release factor, which then enters the A site and catalyses the hydrolysis of the full-length polypeptide from tRNA. After releasing the completed polypeptide, the ribosome falls apart.

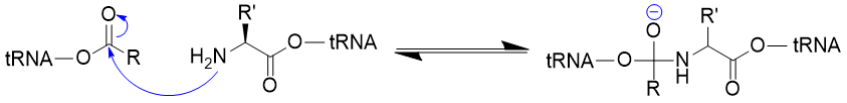
  NB:
  - Release factors are also GTPases. Termination consumes one GTP.
  - The shape of release factor resembles that of tRNA, so that release factors can fit into the binding pocket of tRNA.



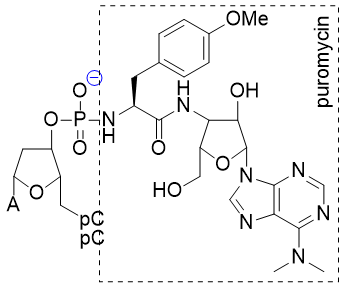
- Proofreading is essential to get high-quality (nearly error-free) peptides. This is achieved in three different ways:
    - Watson-Crick pairing between codon and anticodon
    - Accommodation of aminoacyl-tRNA into the A site (correct interaction between AA and tRNA promotes GTP hydrolysis in EF-Tu)
    - Premature truncation in case an incorrect AA is added (induced by imperfect Watson-Crick pairing between codon and anticodon)

Chemical mechanism of peptidyl transferase

- The acyl transfer reaction catalysed by ribosome occurs via a tetrahedral, anionic intermediate.

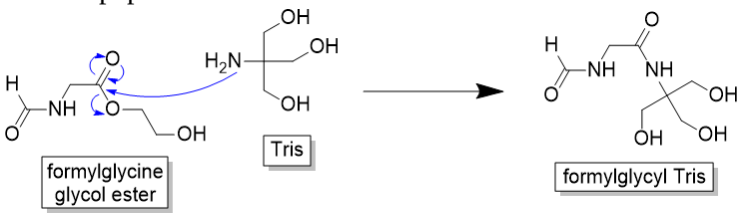


- To understand the kinds of reactions that are important in the binding pocket, a transition state analogue was prepared using a phosphorus derivative called CC-dA-P-puromycin (cytidine-cytidine-deoxyadenosine-phosphate-puromycin). Puromycin is an antibiotic that inhibits translation by binding to the A site of ribosome.
(CC-dA-P-puromycin looks similar to the intermediate but is structurally closer to the transition state, because P-O bond is longer than C-O bond.)

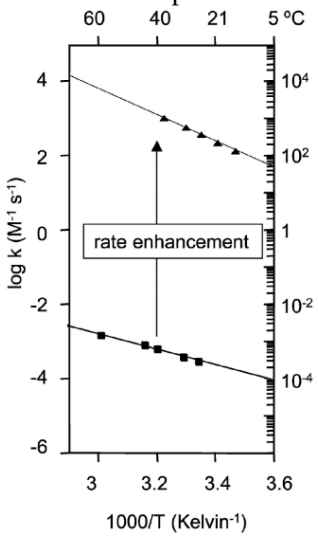


CC-dA-P-puromycin acts as an inhibitor of ribosome and can be used to indicate the position of the active site (see above).

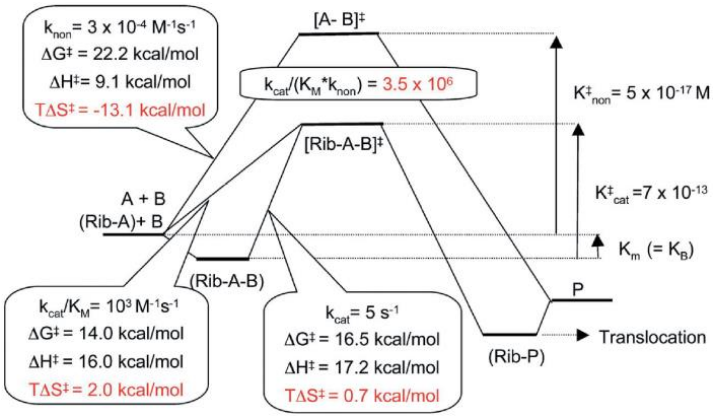
- People have measured the (second-order) rate constant of the peptidyl transfer reaction at different temperatures in order to investigate its temperature dependency. For contrast, a background reaction was also set up using formylglycine activated as a glycol ester, which forms a peptide bond with Tris.



The results showed that ribosome catalyses the peptidyl transfer reaction by more than 10^6 fold over the spontaneous reaction.



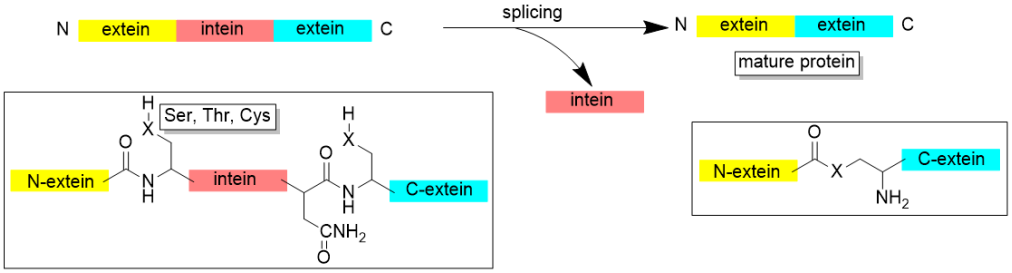
Based on this information, activation parameters of the reaction can be determined, which suggest ribosome to be an entropy trap, i.e. the peptidyl transfer reaction becomes enthalpically less favourable but entropically more favourable, with $T\Delta\Delta S^{\ddagger}$ = 15 kcal/mol.



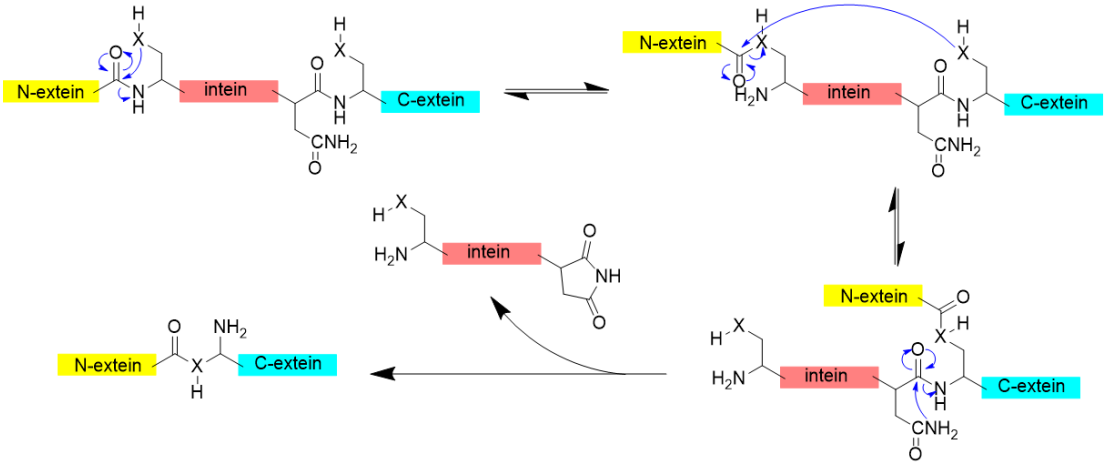
As a classical example of entropy trap, ribosome uses the binding energy to take together two molecules that exist in low concentration, so that the chance for reaction becomes much higher. In other words, ribosome pre-organises reactants for productive reaction.
(Annette Sievers et al., 2004. The ribosome as an entropy trap. Proceedings of the National Academy of Sciences of the United States of America, 101(21), pp.7897–7901.)

Protein splicing

- Similar to pre-mRNA, newly synthesised proteins can also undergo splicing, during which an internal protein segment (called intein) is removed from a precursor protein with a ligation of C- and N-terminal external proteins (called exteins) flanking it.
  The N-terminal extein is connected to the intein by a nucleophilic residue X (X = Ser, Thr or Cys). The C-terminal extein is characterised by an Asn followed by X (X = Ser, Thr or Cys).

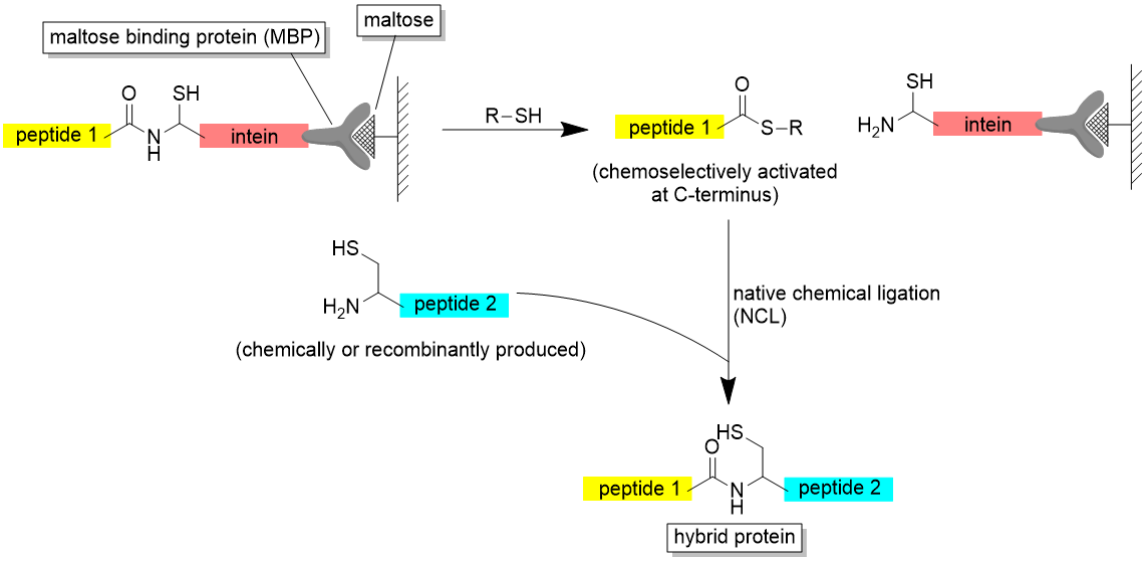


- Mechanism of protein splicing
  In the first step, the nucleophilic residue on the N-side attacks internally, resulting in an ester intermediate. This reaction is reversible and is favoured toward the left side, but once the product is formed, the structure of the intein is such that it brings the C-terminal nucleophile X into proximity of the N-terminal extein, allowing acyl transfer from the N-terminal X to the C-terminal X. Then, the side chain of the C-terminal Asn orients itself and attack the peptide bond next to it, thus cleaving the intein from the protein.



- The evolutionary advantage of producing such self-splicing proteins is still unclear.
- Biotechnological application of protein splicing
  Protein splicing is used in the laboratory to produce unusual and complex proteins, or more specifically, to selectively activate a long peptide chain on its C-terminus.
  For example, the C-terminal Asn of intein can be replaced by a maltose binding protein (MBP), which allows the whole construct to be produced in a cell and selectively enriched using affinity column chromatography.
  With the help of the intein, the peptide on its N-side can be transferred to a thiol species added externally and get eluted from the column. Meanwhile, this process activates the peptide

selectively at its C-terminus (which is hard to achieve chemically, especially for long peptides containing numerous carboxyl groups).
This activated peptide can now be combined with another peptide that has an N-terminal Cys (produced either chemically by SPPS or recombinantly) through NCL.



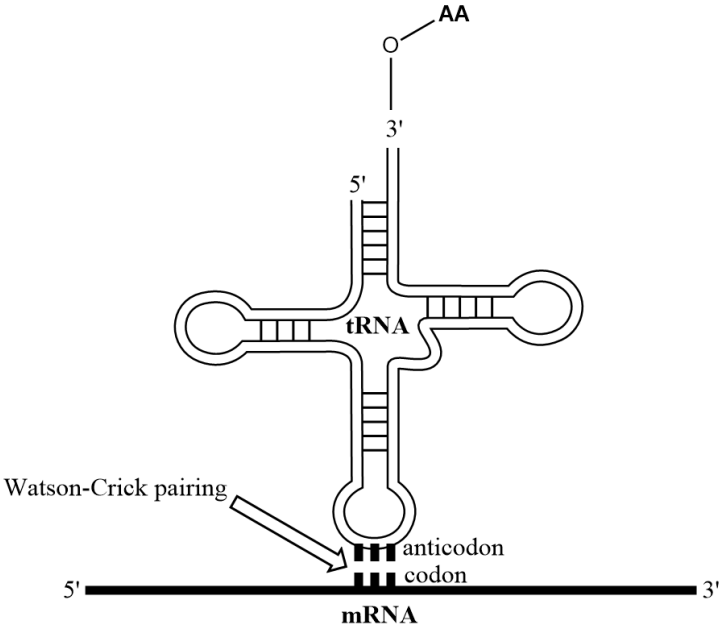
This is a very useful technique for NMR investigation of protein structure (particularly for small proteins), because both peptides (or part of them) can be selectively labelled with ^{13}C and/or ^{15}N, making only the remaining part of the hybrid protein visible in the NMR spectrum. Other uses of this technique include biosensors, mechanistic investigations etc.

Genetic code & translation fidelity 遗传密码与翻译保真度

(19.03.2018)

Genetic code

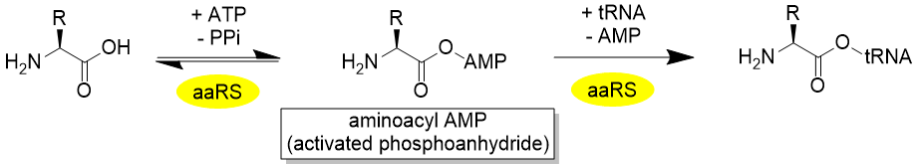
- The genetic code is an algorithm that links information (DNA) with phenotype (amino acids). It is ancient (~$3*10^9$ years old) and nearly universal, indicating that it originated very early during evolution.
- The algorithm works in a fairly simple way. The message, which consists of a series of triplet codons, is read by the adaptor molecule tRNA, which carries amino acids (and the synthesised peptide chain) on its 3' end.
  The recognition occurs through Watson-Crick pairing between the triplet codon on mRNA and the triplet anticodon on tRNA.



- In mitochondria and some lower organisms, the codes are slightly different.
  These variations mostly represent simplifications of the "standard" genetic code.
  Examples:
  AUA → Met (instead of Ile)
  UGA → Trp (instead of STOP)
  AGA, AGG → STOP (instead of Arg)

aaRS: Link between AA and tRNA

- Appropriate amino acids are attached to corresponding tRNAs by the enzyme aminoacyl-tRNA synthetase (aaRS or ARS). It does so by catalysing the esterification of a specific cognate amino acid (or its precursor) to one of its compatible cognate tRNAs.
  aaRS is the only molecule in the cell that "knows" the genetic code (i.e. responsible for linking the correct amino acid and tRNA).



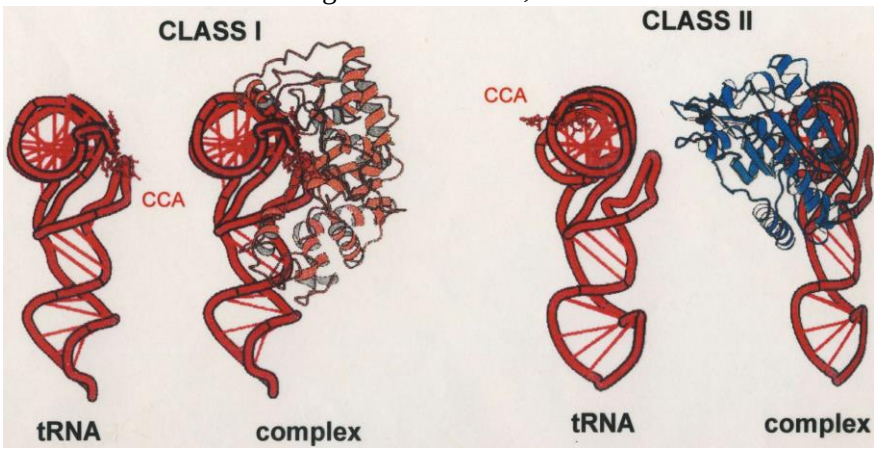
Formation of each aminoacyl-tRNA consumes one molecule ATP. In the first step, an aminoacyl AMP is formed out of an amino acid and an ATP, with loss of pyrophosphate.

Pyrophosphate is hydrolysed into two phosphates, rendering this process essentially irreversible. The aminoacyl AMP is very active (half-life minutes to seconds in aqueous solution) and is held tightly inside the active site before it undergoes another reaction with tRNA, in which an aminoacyl-tRNA is yielded with loss of AMP.
This is an expensive process but is worthy for the cell because the fidelity of translation has to be ensured.

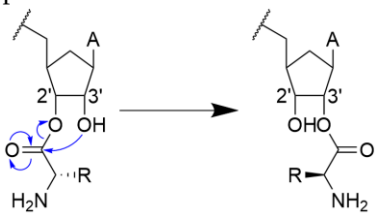
- At least one aaRS is found for each of the 20 basic amino acids. An aaRS recognises only one particular amino acid but may recognise more than one tRNA.
- Like the genetic code, aaRS is also phylogenetically ancient.
  Two classes of aaRS are found in nature, indicating that this enzyme might have evolved twice during evolution. The two classes are responsible for different amino acids. The only amino acid that gets recognised by both classes is Lys.
  The two classes show completely different structures. Various quaternary states are found within both classes (e.g. monomers [α], dimers [α_2], tetramers [α_4], heterotetramers [$\alpha_2\beta_2$]).

	Class I	**Class II**
Quaternary state	Usually monomeric or dimeric	Usually dimeric or tetrameric
Catalytic site	ubiquitous Rossmann fold & parallel beta-strands	antiparallel β-sheet
AAs recognised	Mostly large, hydrophobic AAs (V, L, M etc.)	Mostly small, polar AAs (G, S, A etc.)
Number of AAs recognised	11	10
Chemoselectivity (which hydroxyl group of tRNA is acylated)	2'-OH	3'-OH

From structural studies we also know that the two classes bind to different faces of tRNA. Class I aaRS binds to the right side of tRNA, while class II binds to the left side.



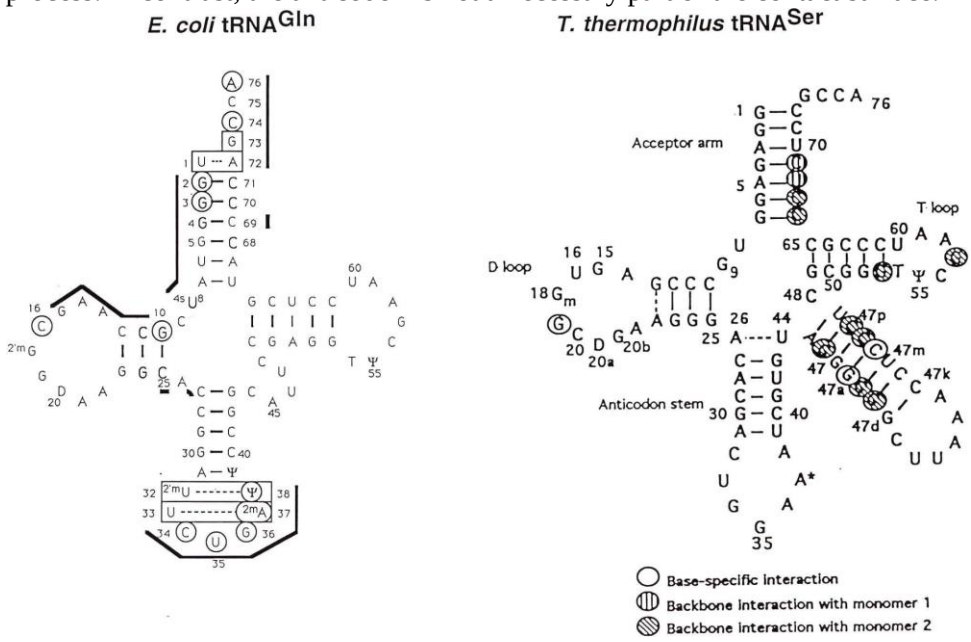
- Chemoselectivity of aaRS
  Although the two classes of aaRS have different chemoselectivity, the end result is the same since the two products (2' ester and 3' ester) undergo rapid interconversion and only the 3' product will be delivered to ribosome.



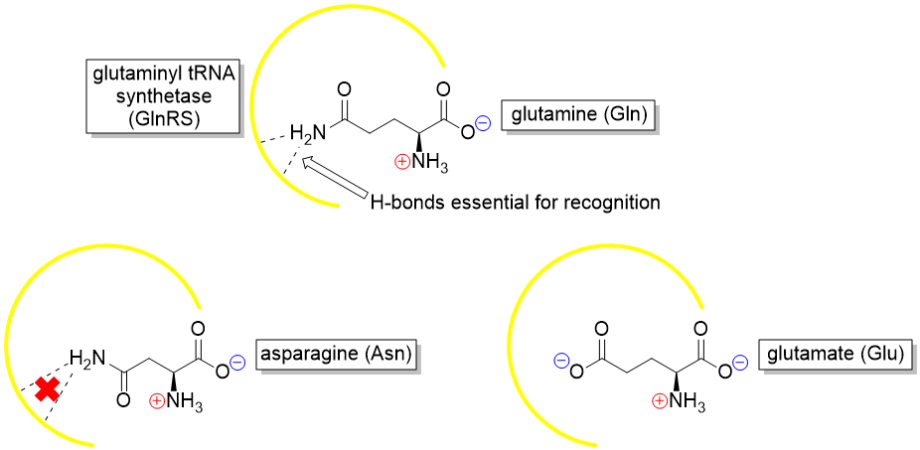
- Exact molecular recognition is essential for protein synthesis.
  The average error rate (Σ) of aaRS is roughly 10^{-6} for tRNA and 10^{-4}-10^{-5} for amino acids. For a protein consisting of 1000 AAs, this means more than 90% of the copies will be error-free.

The higher error rate for amino acids can be explained by the fact that amino acids are much smaller than tRNAs and can be distinguished only by their even smaller side chains.

- tRNA recognition
  The contact surface between aaRS and tRNA is typically very large, but localised to the inner surface of the "L" shape. The enzyme interacts with both the backbone and individual bases (distribution varies from enzyme to enzyme).
  The acceptor arm (i.e. the 7- to 9-bp stem made by the base pairing of the 5'-terminal nucleotides with the 3'-terminal nucleotides of tRNA) is of particular importance for this process. In contrast, the anticodon is not a necessary part of the contact surface.

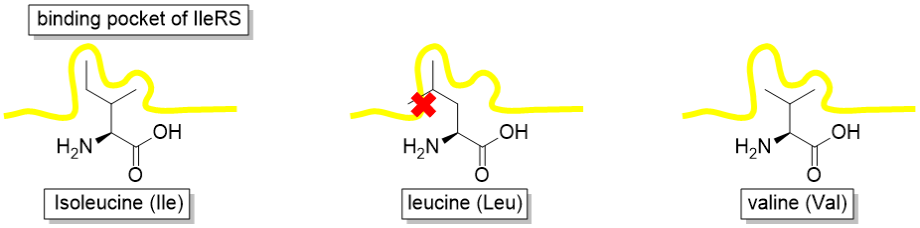


- AA recognition
  Since amino acids are much smaller than tRNA, the recognition is also more difficult (which explains why the error rate for AA recognition by aaRS is higher than that of tRNA).
  The chemical property of the side chains plays an important role in this recognition process.
  - Example 1: How does GlnRS distinguish between Gln, Asn and Glu?
    GlnRS forms two hydrogen bonds with the target amino acid (glutamine). The side chain of Asn is too short to make this contact. As for Glu, its side chain is deprotonated at physiological pH and thus cannot form any hydrogen bond.



The shape of the binding pocket is another determining factor in the recognition.
In some cases, two amino acids can be so similar that it is hard to distinguish them effectively by their chemical or physical properties. The aaRS then uses its "proof-reading" mechanism, which is carried out by a so-called editing domain, to remove noncognate AAs that are smaller (than the cognate AA) by selective hydrolysis.

- Example 2: How does IleRS distinguish between Ile, Leu and Val?
  The binding pocket of IleRS fits perfectly to the shape of isoleucine.



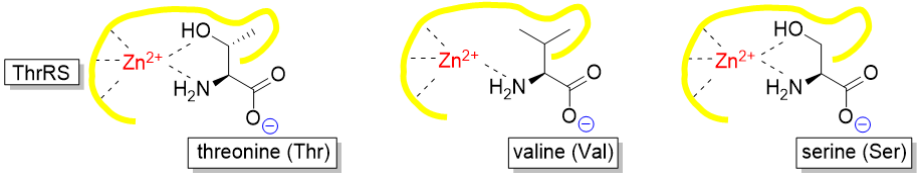
  The structure of leucine prevents it from fitting into this pocket.
  For valine, the situation is more complicated, since lack of a single methyl group results in a decrease of merely <3 kcal/mol in binding energy (ΔG). Given the fact that [Ile]:[Val] = 1:5 in cells, we can now estimate the error rate:

$$\Sigma = \frac{k_{\mathrm{Val}} \cdot [\mathrm{Val}]}{k_{\mathrm{Ile}} \cdot [\mathrm{Ile}]} = e^{-\frac{\Delta\Delta G}{RT}} \cdot \frac{[\mathrm{Val}]}{[\mathrm{Ile}]} \approx 0.05$$

  which means that 5% of all isoleucine will be mistaken by valine.
  In this situation, valines that are incorrectly loaded to the tRNA for isoleucine can be removed selectively by an editing domain in IleRS, which matches valine very well and excludes isoleucine sterically.
  This mechanism, through which larger noncognate AAs are excluded sterically and smaller ones removed by proof-reading, is called the "double-sieve" mechanism.

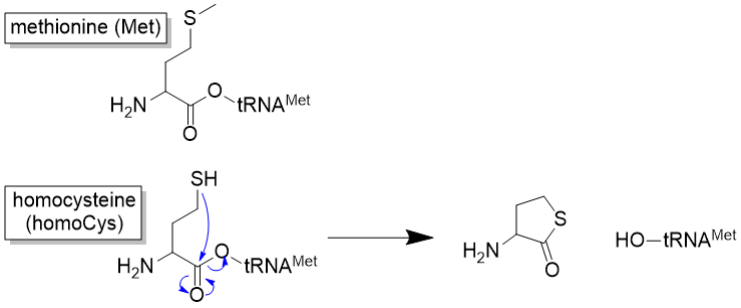
Editing domains are present also in many other aaRS, such as ThrRS.

- Example 3: How does ThrRS distinguish between Thr, Val and Ser?
  Recognition of Thr by ThrRS is mediated by a Zn^{2+} ion, which is held in place by three different side chains of ThrRS. Recognition of Thr occurs through two hydrogen-bonds between Thr and Zn^{2+}.
  Valine is excluded because it forms only one hydrogen bond.
  Ser, in contrast, cannot be distinguished efficiently from Thr. Therefore, the same proof-reading mechanism as in IleRS is needed to selectively hydrolyse incorrectly loaded Ser-tRNA^{Thr}.



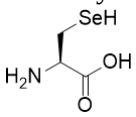
Not all aaRS have editing domains. Sometimes, exclusion of an incorrectly loaded AA can occur spontaneously. This is called chemical editing.

- Example 4: How does MetRS distinguish between Met and homoCys (a non-proteinogenic AA serving as synthetic precursor of Met)?
  When homoCys is loaded to tRNA^{Met} , it undergoes an intracellular cyclisation and gets cleaved (self-correction).

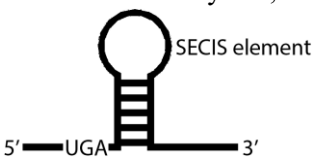


Expansion of genetic code

- Selenocysteine (Sec/U)

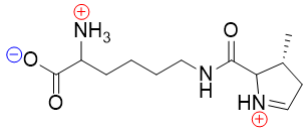


Sec was found to be the "21st proteinogenic AA".
In human proteome, 25 essential selenium proteins have been identified, such as glutathione peroxidase (GPx), which has a Sec at its active site to detoxify alkyl hydroperoxides, thus protecting the cell against oxidative stress. Another good example of active site Sec is iodothyronine deiodinase, which selectively removes an iodine from thyroxine (T_4), a thyroid hormone that has 4 iodine atoms.
Sec is encoded by UGA (which normally encodes STOP). In prokaryotes, UGA encoding Sec is distinguished from normal UGA by an RNA hairpin structure following immediately the UGA codon. This hairpin structure is called the selenocysteine insertion sequence (SECIS) element. In eukaryotes, the SECIS element is located in the 3' non-coding region.



Effective decoding of Sec requires four genes: $tRNA^{Sec}$, SelA, SelB, SelD
$tRNA^{Sec}$ is very similar to $tRNA^{Ser}$. Therefore, it is recognised by SerRS and charged with a Ser, which then gets converted into Sec with SelA and SelD serving as selenium source. Finally, SelB, the Sec-specific EF-Tu analogue, recognises and binds to Sec-$tRNA^{Sec}$ as well as the SECIS element.
Note that there is no specific aaRS for Sec. This AA is encoded by the so-called "over-coding" mechanism.

- Pyrrolysine (Pyl/O)



Expansion of genetic code seems to have happened more than once. After the discovery of Sec, Pyl was found to be the "22nd proteinogenic AA".
Pyl is used by methanogenic bacteria in an enzyme that catalyses the conversion of methylamine (CH_3NH_2) into $CH_3S(CH_2)_2SO_3H$, which serves as a precursor of methane.
Unlike Sec, Pyl does have a dedicated aaRS. $tRNA^{Pyl}$ recognises UAG (which normally encodes STOP).
Since this is a free-standing system, it is easily transferred to other organisms.

Beyond the 20 proteinogenic amino acids 基本氨基酸以外的其它氨基酸

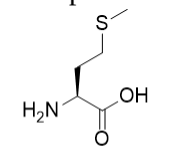
(26.03.2018)

General

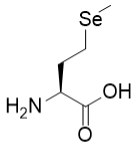
- Nature has expanded the standard genetic code by STOP codon reassignment. Examples include Sec (UGA) and Pyl (UAG). (See the last lecture)
- Further expansion of the genetic code is possible and can be done in laboratory.

Strategy 1: Take advantage of structural similarities to "fool" aaRS

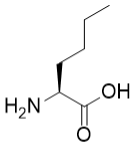
- aaRS is the only enzyme that is responsible for combining tRNA with the correct AA. If there exists a non-canonical that is structurally similar enough to a canonical AA, then it might get recognised by an aaRS and loaded onto a tRNA.
- Example: incorporation of SeMet and nLeu



methionine (Met)



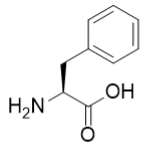
selenomethionine (SeMet)



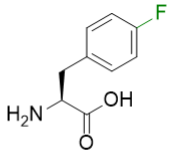
n-leucine (nLeu)

Selenomethionine (SeMet) and n-leucine (nLeu) are very similar to Met. Both are accepted by MetRS to charge the corresponding $tRNA^{Met}$.
SeMet-containing proteins are very useful for determination of protein structure by X-ray and NMR, because (1) selenium shows distinctive X-ray scattering and large NMR dispersion, and (2) SeMet and Met are isosteric, i.e. replacing a Met with SeMet will not lead to significant change of protein folding and its overall structure.
To produce proteins with SeMet instead of Met, a plasmid encoding the protein of interest (POI) is introduced into Met-auxotroph *E. coli* cells. These cells are grown on minimal medium with Met supplement but without inducer of the introduced plasmid (so that the POI is not expressed). After the required cell number is reached, the cells are washed extensively to remove excessive Met. The inducer is now added together with SeMet, allowing the production of the POI whose Met residues are replaced by SeMet.

- Other examples following the same principle:
  - Replacement of Phe with fluorophenylalanine

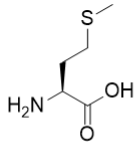


phenylalanine (Phe)

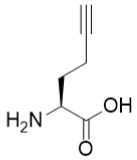


fluorophenylalanine (F-Phe)

  - Replacement of Met with a terminal alkyne
    Since the first (N-terminal) AA of a protein is always Met, this replacement can be used to label a protein at its N-terminus through Click reaction.

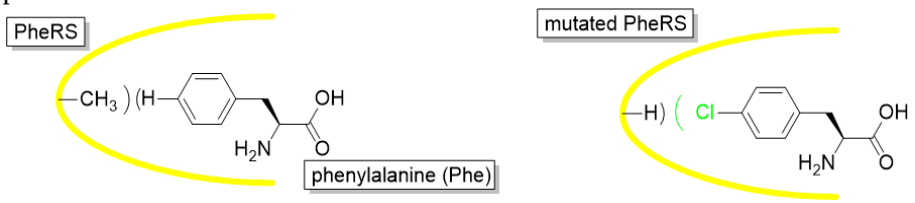


methionine (Met)

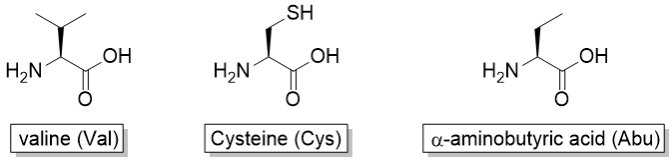


Strategy 2: Relax the substrate specificity of aaRS

- Specificity of aaRS can be reduced by mutating a site chain in its binding pocket.
  Example: PheRS
  The binding pocket of PheRS contains an Ala residue that interacts with the aryl side chain of Phe. Replacement of this Ala by Gly allows recognition of larger atoms or groups on the para position of the Phe side chain.



- Another way to reduce specificity of aaRS is to impede proof-reading by damaging the editing domain.
  Example: ValRS
  ValRS excludes incorrectly loaded Cys and aminobutyric acid (Abu) by selectively cleaving them through its editing domain. If the editing domain is mutated and loses its function, Cys and Abu will not longer be excluded.



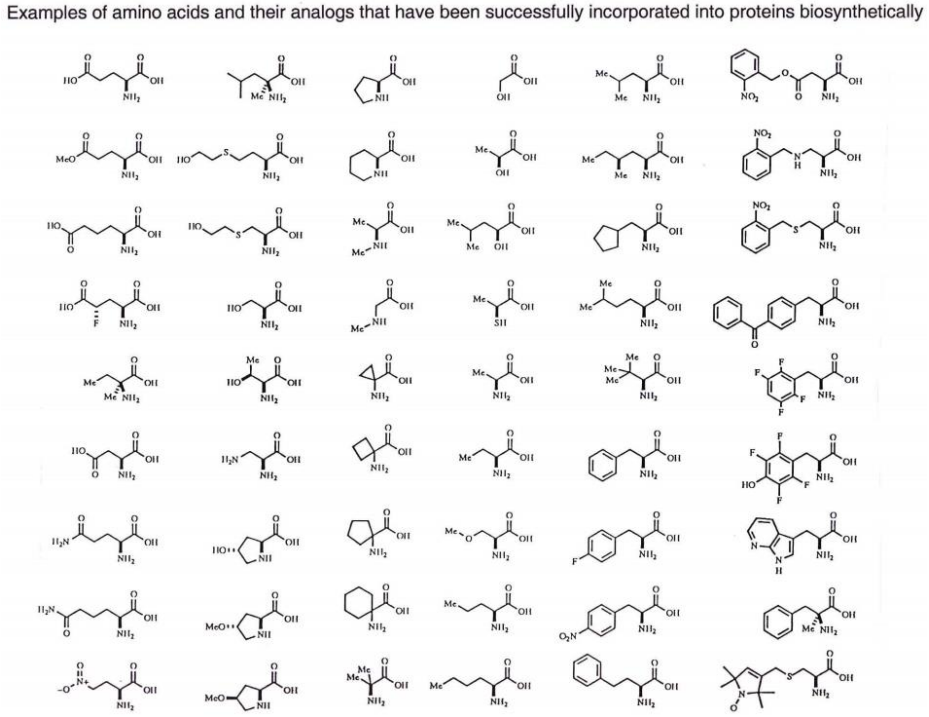
  In one of the experiments to invade the Val coding pathway, an essential Cys in the enzyme thymidylate synthase was replaced (genetically) with Val. Thymidylate synthase is an enzyme involved in the biosynthesis of thymidine and is essential for making new DNA.
  The genome was then mutated and variants that could replace Val with Cys (rescuing the thymidylate deficiency and allowing the cell to grow) were selected. Among these variants, a mutant ValRS* was found with the mutation T222P, which is located in the editing site and inactivates proof-reading.
  This mutant was introduced into *E. coli* cells. After growing the cells in the presence of Abu, more than 20% of all Val in the proteome were found to be replaced by Abu.
- Producing noncanonical proteins with AA surrogates turns out to be relatively easy, but this strategy has some limitations:
  - The noncanonical AA has to resemble a natural counterpart.
  - Incorporation of the noncanonical AA is global (i.e. nonspecific)
  - Suppression efficiency (proportion of AA that gets replaced by the noncanonical one) is not 100% and depends on position in protein.

Strategy 3: STOP codon reassignment in vitro

- This approach aims to mimic the way nature expands the genetic code (see the last lecture).
- STOP codon suppression can be done with a suppressor aminoacyl-tRNA that recognises the STOP codon.
- The most frequently used STOP codon for this strategy is UAG because it is the rarest one among all three STOP codons.
  The corresponding suppressor tRNA should thus have the anticodon CUA. To avoid being recognised by an *E. coli* aaRS (since otherwise the tRNA will be loaded with a canonical AA), a tRNA of archaea with engineered anticodon is used. Incorporation of the noncanonical AA with the suppressor tRNA can be achieved by engineered aaRS (not always available) or by chemical synthesis.

- The protein containing noncanonical AA is usually obtained by in-vitro transcription & translation (ivTT).
  Recipe:
  - cell extract (enzymes, ribosome, accessory factors for translation, etc.
  - the gene that has the STOP codon at desired position(s)
  - NTPs
  - standard and noncanonical AAs
  - standard tRNAs and the suppressor tRNA loaded with noncanonical AA
  - PEP (source of energy)
  - salts (e.g. Mg^{2+})

  Until now, more than 150 noncanonical AAs have been incorporated site-specifically into a huge number of proteins (fluorophores, probes, cross-linkers etc.) with this procedure, enabling the investigation of physical organic chemistry on large macromolecules.



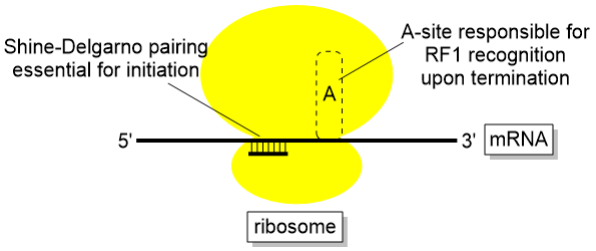
- Limitations:
  - low yields (often < 100 mg)
  - suppression efficiency varies (10% - 100%) and is context-dependent, meaning that many truncated proteins are produced

Strategy 4: Codon reassignment in vivo

- The foundation of codon reassignment in living cell is a novel codon.
  - Solution 1: simply use UAG because it is rare and there is little competition. (Sometimes it is also possible to delete all UAG codons from an organism.)
  - Solution 2: use base quartets instead of base triplets.
  - Solution 3: introduce unnatural bases into the genome

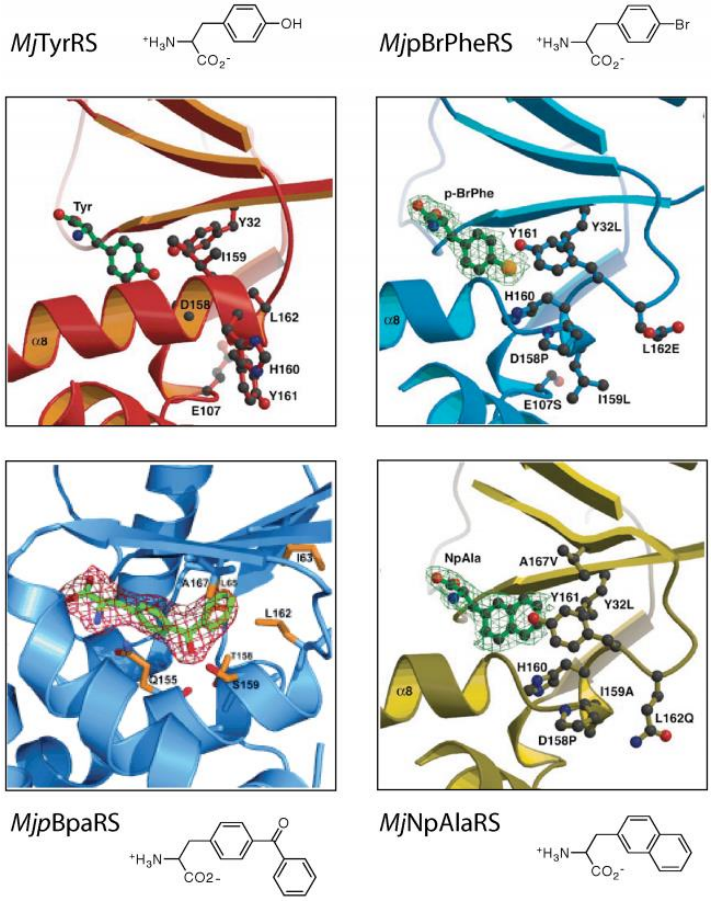
  In all cases, a suppressor tRNA that has an appropriate anticodon needs to be prepared and it should not be recognised by any aaRS of the host cell. In addition, a corresponding aaRS that selectively loads the noncanonical AA onto the new tRNA is required.
- Source of noncanonical AA can be either uptake or biosynthesis.
- The new tRNA can be obtained by taking a tRNA from another organism and engineering its anticodon.

- Evolution of a novel aaRS specific for a tRNA and an AA is often the key problem in the procedure.
  One approach is to conduct iterative cycles of positive & negative selection.
  - Starting material:
    TyrRS or PylRS from the archaeon *Methanococcus jannaschii*.
  - Positive selection:
    E. coli cells are transformed with two plasmids, one encoding a library of aaRS (each cell gets one variant), the other encoding the suppressor tRNA and an enzyme called chloramphenicol acetyltransferase (CAT) which detoxifies the antibiotic chloramphenicol. However, the gene for CAT has a premature STOP codon (nonsense mutation) that leads to a truncated version of the protein. This STOP codon is recognised by the suppressor tRNA. If the suppressor tRNA is loaded with an AA, truncation of CAT will be rescued.
    The cells are then grown on a medium containing chloramphenicol and the noncanonical AA. In order to survive the antibiotic, a cell has to be able to produce full-length, functional CAT by aminoacylating the suppressor tRNA.
  - Negative selection:
    Plasmids encoding the novel aaRS are extracted from the survivor cells of positive selection and introduced into another population of *E. coli* cells, together with a second plasmid which encodes the suppressor tRNA and a toxic RNase called barnase. The gene for barnase has two premature STOP codons (nonsense mutations), making it extremely hard for the cell to express it, unless the cell can load AAs onto the suppressor tRNA.
    The cells are then grown in the absence of the noncanonical AA. Cells that can load AAs to the suppressor tRNA will not survive due to production of active barnase. Therefore, aaRS variants that load the suppressor tRNA with an endogenous AA will be eliminated.
    Finally, "winners" of this negative selection goes into a new cycle of selection. The cycle is repeated until a required specificity is achieved.
- To evolve the aaRS, one can randomly mutate 5-6 AAs in the binding pocket of the enzyme. This method has a very high throughput (typically 10^8 variants in each experiment) and has been used to generate more than 70 mutated TyrRS as well as many mutated PylRS that specifically recognise and load a noncanonical AA to a suppressor tRNA.
- In vivo production of non-standard proteins with engineered aaRS turned out to be much more efficient (i.e. have a better yield) than in vitro transcription & translation (ivTT), probably because the aaRS has a very high catalytic efficiency and because the required aminoacyl-tRNA is constantly regenerated.
  Nevertheless, the suppression efficiency of this strategy is still low (ca. 20% - 30%), since the suppressor tRNA has to compete with a RF1 (in case of UAG).
  This problem can be solved either by deleting the gene for RF1 from the genome or by evolving a mutant ribosome that does not bind RF1. However, since RF1 is needed to terminate the translation of some proteins in the cell, the inability to produce or recognise RF1 can lead to unexpected side effects.
  The alternative method is to evolve an orthogonal ribosome (O-ribosome) with corresponding orthogonal mRNA (O-mRNA), which form an isolated system parallel to the normal translation system of the cell.
  Crosstalk between the two systems can be prevented because mRNA is recognised by the ribosome through its Shine-Delgarno sequence, which pairs with the anti-Schine-Delgarno sequence of 16S rRNA (in prokaryotes). Therefore, one can make the O-ribosome and O-mRNA simply by mutating the anti-Shine-Delgarno in 16S rRNA and the Shine-Delgarno in mRNA. After that, the A-site can be mutated to select for variants that do not bind RF1.



This leads to improved UAG decoding. For mRNAs containing one UAG, the suppression efficiency is improved to 60%. For mRNAs containing two UAGs, an efficiency of 20% can still be reached.
(More details: Oliver Rackham & Jason W Chin, 2005. A network of orthogonal ribosome·mRNA pairs. Nature Chemical Biology, 1(3), pp.159–166.)

- aaRS seems to be incredibly malleable (structurally plastic). X-ray studies have shown that they can undergo both side chain and backbone modifications, which allow optimisation of H-bonding and packing. Their binding pockets can be very distinctive from each other.



Future goals

- Expand the strategies to more AAs
- Increase the efficiency of introducing multiple noncanonical AAs
- Apply to higher organisms (yeast, drosophila, humans)
- Permanent addition of a noncanonical AA to an organism (+ ethical problems?)

Posttranslational modification 翻译后修饰

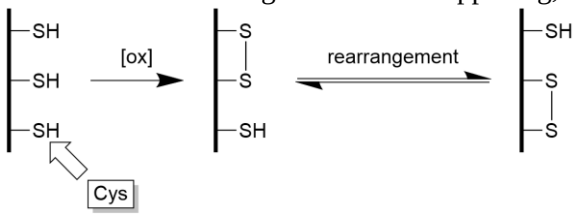
(09.04.2018)

PTM in general

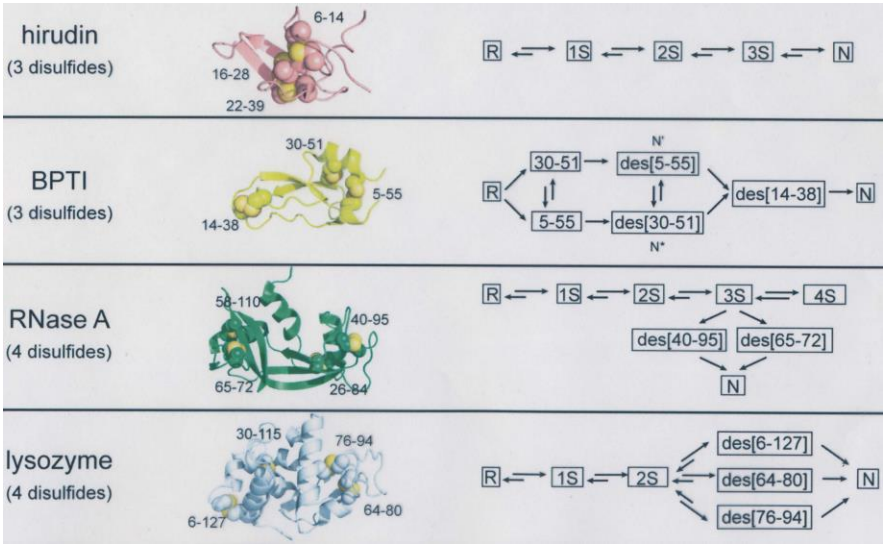
- Natural proteins are synthesised from 22 basic amino acids. However, more than 200 noncanonical amino acids have been identified in isolated proteins. Most of these noncanonical amino acids are introduced co- and posttranslationally.
- Why does the cell need PTM?
    - Nature exploits co- and posttranslational modifications for the fine-tuning of proteins in the cell.
    - PTM controls structure , conformation, localisation and lifetime of proteins.
    - PTM introduces unique recognition handles, which can be exploited to bring two molecules in a cell together in various fashions.
    - PTM can be used as "on/off" switches (e.g. phosphorylation).
    - Sometimes PTM gives rise to new cofactors that can be used for catalysis and generate spectroscopic signatures and new chemistry.
- Major types of modifications are:
    - Oxidation
    - Addition of specific groups (phosphate, acetyl, saccharide etc.)
    - Hydrolysis
    - Rearrangement (both backbone and side chain)

Oxidation

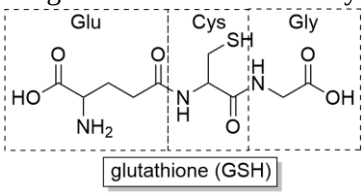
- Oxidative protein folding
  In the presence of an oxidant (e.g. molecular oxygen), cysteine residues in a peptide can undergo oxidation and form a cross-link in the polypeptide, restricting the conformational probabilities of the polypeptide chain and providing an element of stability.
  A challenge for the cell now is to find out the native (correct) cross-links and prevent unwanted cross-linking events from happening, and the solution here is a rearrangement step.



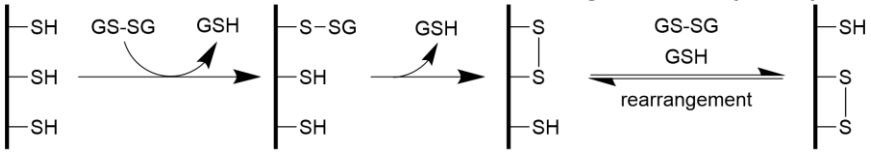
  The reactions are catalysed by enzymes containing a conserved -CXXC- motif. These enzymes act as molecular rheostats, because they have reduction potentials ranging from -120 to -300 mV. They are generally specific for the oxidation or the rearrangement step.
  For proteins that contain more than one disulphide bonds, there are several different strategies to achieve the native (fully and correctly oxidised) state, depending on the molecular context. Some proteins fold up in a linear fashion (e.g. hirudin), some undergo bifurcated (e.g. BPTI, RNase A) or trifurcated (e.g. lysozyme) pathways.



- Oxidative protein re-folding in vitro
  If a protein containing multiple disulphides is to be produced recombinantly in vitro.
  A possible solution is to use a redox buffer that mimics the activities of the -CXXC- redox enzymes. Glutathione (GSH), for example, is a very typical redox buffer used in this approach. It is a special tripeptide consisting of a glutamate, a cysteine and a glycine, in which the glutamate is attached to cysteine via its side chain.



  This molecule can be converted to its oxidised form, glutathione disulphide (GS-SG), in the presence of an oxidant (E_0' = - 275 mV).
  During in vitro re-folding, GS-SG and GSH are introduced into the solution containing the protein to be folded. GS-SG is an oxidising agent that drives formation of disulphide bonds, while GSH does the opposite. Oxidative folding is then driven by constant loops of oxidation and reduction, a mechanism similar to the rearrangement catalysed by the -CXXC- enzymes.

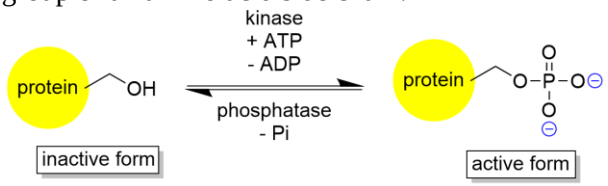


  A typical setting for this approach is 0.2 mM GS-SG + 1.0 mM GSH (close to the natural concentration of GSH in living cells).

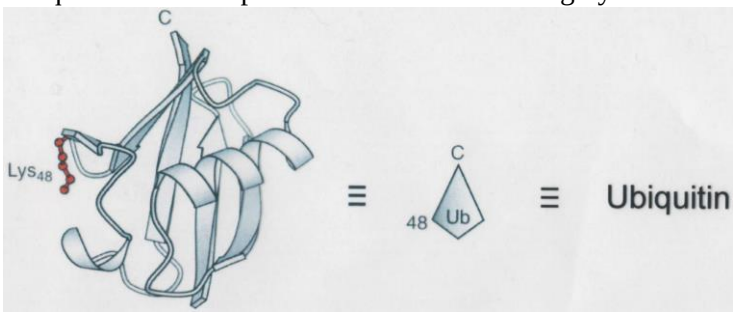
Group addition

- In a group addition modification, a nucleophilic side chain of an amino acid reacts with an electrophilic reagent (donor of the group to be added).
  Some examples of electrophilic reagent:
  ATP → phosphorylation
  3'-Phosphoadenosine-5'-phosphosulfate (PAPS) → sulfation
  Acetyl CoA → acetylation
  UDP-glucose → glycosylation
  S-adenosylmethionine (SAM) → methylation
  Farnesyl pyrophosphate (FPP) → farnesylation

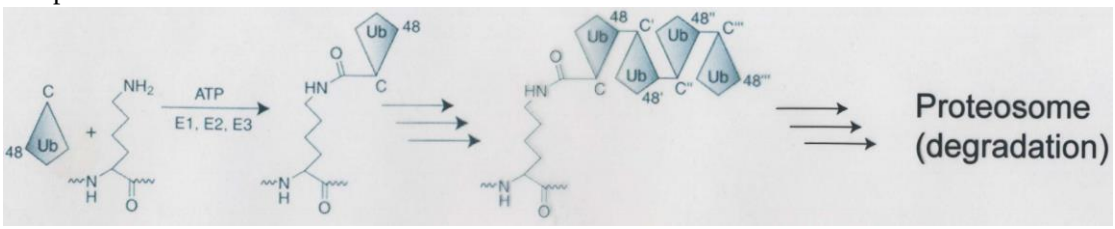
- Phosphorylation
  Phosphorylation is important for signalling in higher cells. It usually occurs on the hydroxyl group of an amino acid side chain.



  Phosphorylation results in change in charge and bulk of the residue. This process is reversible and can serve as a "molecular switch".
  In humans, ~500 kinases (mostly for Ser, Thr, Tyr) and ~150 phosphatases have been identified, which together work on roughly 10000 target molecules.
  The "kinome" (the set of molecules in the cell that can be phosphorylated and their degree of phosphorylation) is highly dynamic.
- Protein tags
  A good example for protein tags is ubiquitination.
  Ubiquitin is a small protein of 76 AAs and is highly conserved from bacteria to eukaryotes.



  In the process of ubiquitination, a ubiquitin is attached to lysine residue of a client protein through its C-terminus (This is called an isopeptide bond since it links the C-terminus not to the N-terminus but to a side chain). Another ubiquitin can then be attached to the lysine residue on position 48 of the first ubiquitin. This process is repeated to form a long chain of ubiquitin (polyubiquitination), which marks the protein for transport into various cell compartments.



  The most common target of ubiquitin tagging is proteosome, where the protein gets degraded and recycled. This is a mechanism to control the lifetime of proteins in cells.
- "Second genetic code"
  The site of modification is dictated by sequence and structural motifs. For example, glycosylation of secreted proteins occurs on the **Asn**-X-Ser/Thr motif, while farnesylation on the **Cys**-Ala-Ala-X-COO^- motif.
  However, these motifs are often difficult to recognise and exploit, which is further complicated by combinatorial diversity.
  Example: histone code hypothesis
  Modification of the flexible tail of histone is used by the cell to control transcription.
  For example, on the tail of histone H3, there are 4 serines that can be phosphorylated, 2 lysines that can be acylated, and 4 lysines and 2 arginines that can be methylated.
  These modifications change the charge and bulk of histone tails, modifying their interactions with DNA and thus regulating transcription.

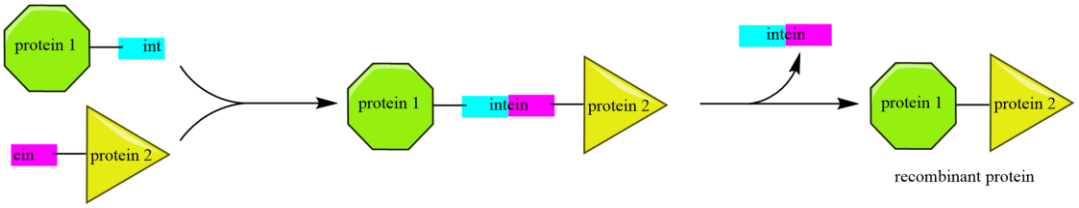
While phosphorylation and acylation are reversible, methylation is mostly irreversible.

Hydrolysis

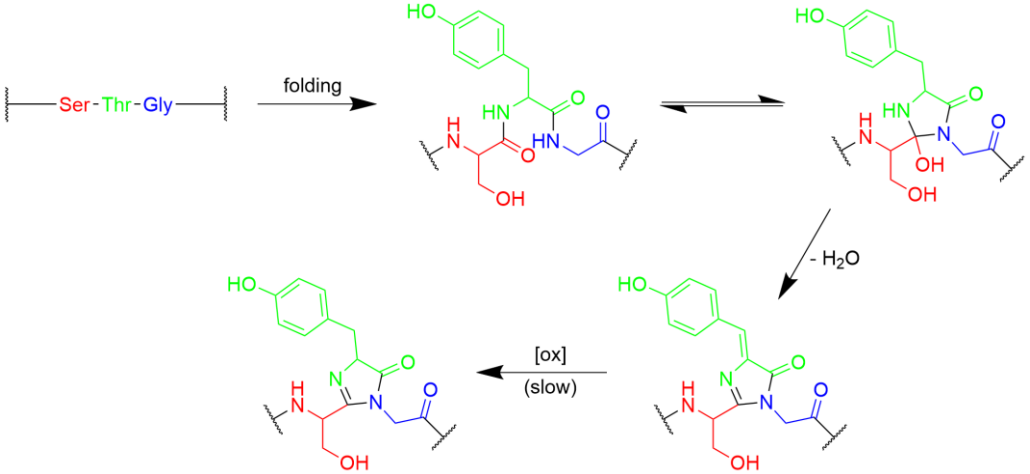
- One example of hydrolysis is the maturation of the big polyprotein that constitutes the building blocks of a virus. These building blocks are produced together as a very long polypeptide, which needs to be cleaved proteolytically to generate the small fragments that undergo self-assembly.
- Hydrolysis also plays an important role in the blood coagulation cascade, in which hydrolysis of a series of zymogens (preenzymes, an inactive precursor of an enzyme) is needed. A zymogen is activated by hydrolysis. Then it acts on the next zymogen in the cascade, leading to its hydrolysis and activation.
  Such a catalytic cascade enables enormous amplification of signal in response to insult.

Rearrangement

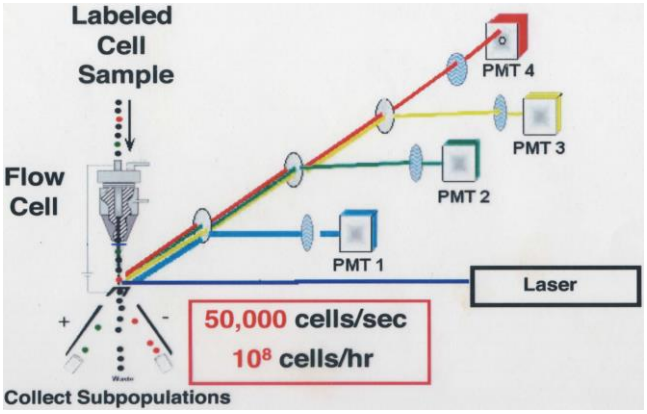
- Backbone rearrangement
  Inteins are a good example for backbone rearrangement.
  Inteins can be used to produce recombinant proteins (e.g. attaching a reporter molecule to the protein of interest). To do this, an intein is split into two fragments. Each fragment is attached to one of the two proteins to be combined. Once the two parts comes together, the intein gets activated and cut out of the protein, leaving the recombinant protein.



- Side chain rearrangement
  The chromophore in GFP (green fluorescent protein) is generated by PTM.



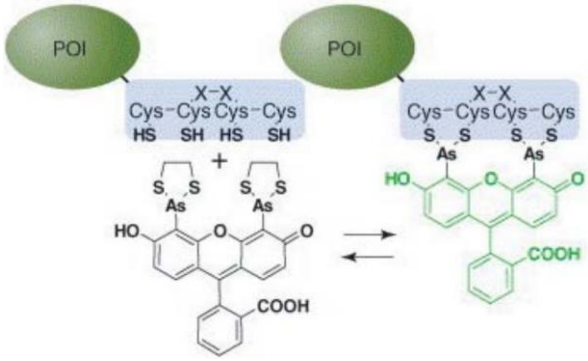
GFP is a commonly used, genetically encoded protein tag (reporter). It can be introduced into various cell types through transgenic technology and be combined with nearly any target protein through backbone rearrangement (see above).
There are also tags of many other colours (cyan, yellow, red etc.).
Once proteins tagged with GFP is produced, they can be localised using microscopy and quantified by FACS (fluorescence-activated cell sorting) in flow cytometry.



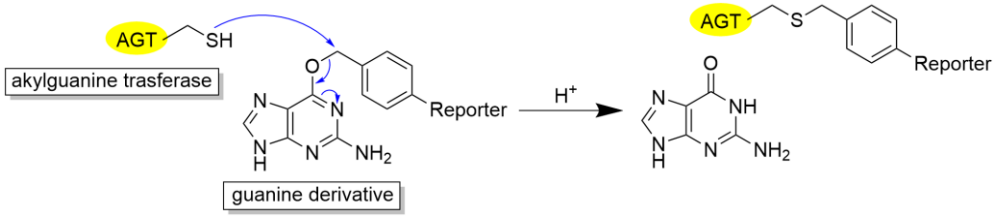
In FACS, cells that contain the fluorophore(s) are sent one by one through a very small pore, where the fluorophore gets activated by laser. Dichroic mirrors then direct emitted light of different wavelengths to the corresponding detectors, generating a fluorophore profile for each cell. Cells are sorted according to the fluorophores they contain and are collected separately.

Diversification & generalisation of PTM labelling using chemical methods

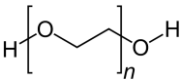
- In general, a tag molecule is bound to the protein of interest and a ligand specific for the tag is attached to some reporter molecule. The tag and the ligand can form a complex (either covalently or noncovalently), allowing the detection of the protein of interest in a complex biochemical mixture.
- Some examples of (POI-)tag → ligand(-Reporter) combinations:
  Streptavidin → biotin ($k_d = 10^{-14}$)
  DHFR → methotrexade (an antibiotic) ($k_d = 10^{-12}$)
  CCXXCC → biarsenical dye ($k_d = 10^{-9}$)



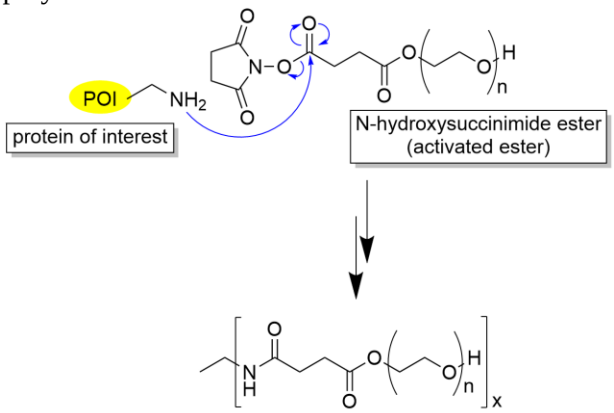
Akylguanine transferase (a DNA repair enzyme) → akylated guanine (covalent)



- Genetically encoded tags are very big, which can potentially perturb the interaction between the natural protein and its ligand. An alternative to genetic modification is the direct chemical modification of the protein of interest. Here we distinguish between non-specific and specific approaches.
- Non-specific reaction: PEGylation
  PEG (polyethylene glycol) is a long polymer.



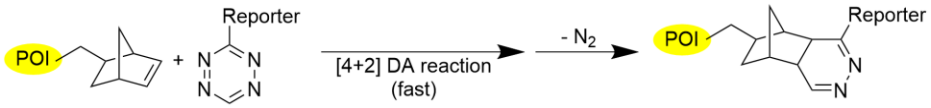
The polymer can be converted into an activated ester (e.g. N-hydroxysuccinimide ester), which can then react with the lysine residues on the surface of a protein, and attach the polymer to these residues.



The PEG tags generated on the surface of the POI protect it from proteolysis and are typically non-immunogenic, thus increasing the serum half-life of the POI. This is useful for various purposes, including enzyme therapy (e.g. adenosine deaminase, an enzyme that SCID [severe combined immunodeficiency] patients lack).

- Specific reaction: bio-orthogonal "click" reaction with non-standard AAs
  Nowadays it is possible to produce proteins with non-standard amino acids (and in some cases very efficiently). Introduction of such AAs into the POI allows specific modification on selected positions.
  For example, if the POI contains an azide, then it can be modified by an alkyne in the presence of a copper catalyst.

  Here, the linker between the reporter and the tag is very small.
  The POI can be produced not only genetically but also recombinantly, meaning that in principle this approach can be applied to any protein.
  In another example, a more complex side chain containing a bicyclic group is used to react with a tetrazine through Diels-Alder reaction, which is known to be very fast.



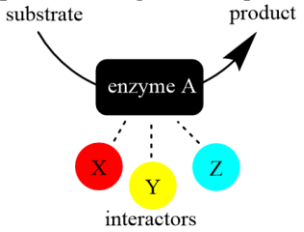
  Beside [4+2], other kinds of cycloadditions such as [1+3] can also be used in a similar manner.
- All the examples above show that we can also do reactions on proteins and get defined products, and there are many ways to artificially expand the diversity of protein structures.
- Ideal tag: genetically encodable, small
  Ideal reporter: rapid and bio-orthogonal reaction with the tag
  This is now an active area of research. There is a need of better interactions as well as temporal & spatial resolutions.

Proteomics & protein-protein interactions 蛋白组学与蛋白-蛋白相互作用

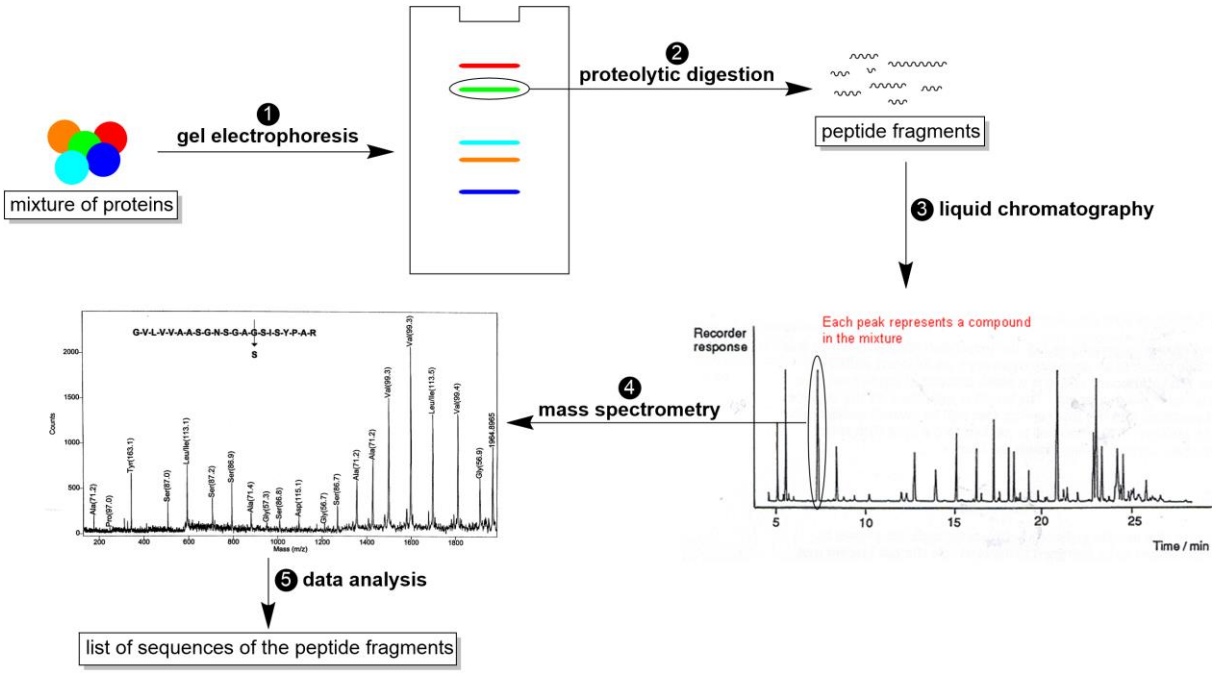
(16.04.2018)

Proteomics

- The term "proteome" is coined after the term "genome".
  A genome is by definition the total genetic material of an organism, consisting of genes and non-coding areas.
  A gene can be of either known function or unknown function. Two genes that are highly homologous to each other (i.e. sharing a similar sequence and a common ancestor) are called related genes. Note that related genes can have different functions.
  Homologous sequences are called orthologous if they are descended from the same ancestral sequence separated by a speciation event. Homologous sequences are called paralogous if they were created by a duplication event within the genome.
- Proteome is defined as the total repertoire of proteins in a cell.
  Proteome is usually significantly larger than the genome because of PTM, alternative splicing, alternative cleavage etc. In contrast to genome, proteome is a dynamic system and its composition depends strongly on temporal regulation.
- The relationship between protein and phenotype/function is complex.
  In the classical view, a phenotype (e.g. catalysis of a reaction) is caused by one or several defined proteins (e.g. enzymes). However, people have now realised that most functions in the cell are carried out by a complex network of proteins (For example, multiple interactor proteins might be required for an enzyme to function properly).

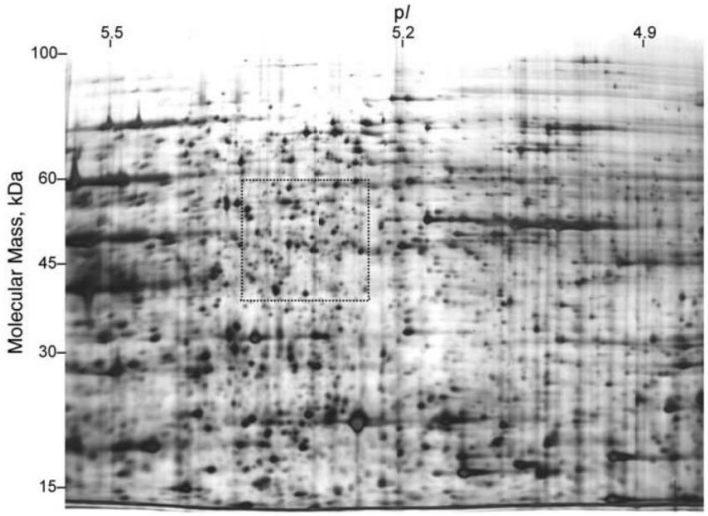


- Basic procedure of proteome analysis
  The typical way to analyse a protein mixture (e.g. proteome separated from a cell culture) compromises five steps:



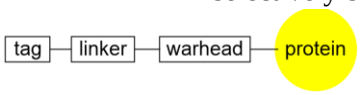
Step 1: separation of individual protein species, often carried out by gel electrophoresis.
Step 2: proteolytic digestion of the individual protein species obtained in step 1, e.g. by the protease trypsin (which cleaves after arginine and lysine). The result is a large pool of small peptide fragments for each protein.
Step 3: separation of the peptide fragments of the digested protein, usually with liquid chromatography.
Step 4: identify (i.e. determine the sequence of) all fragments, e.g. by mass spectroscopy. (no detailed introduction here)
Step 5: Information of all sequences are used to reconstruct the sequence of the protein and sequences of all proteins forms the proteome.

- Separation of different proteins is the key step in proteome analysis.
The most commonly used method is two-dimensional gel electrophoresis, with SDS-PAGE as the first dimension (separation by charge) and IEF (isoelectric focusing) as the second dimension (separation by isoelectric pH).

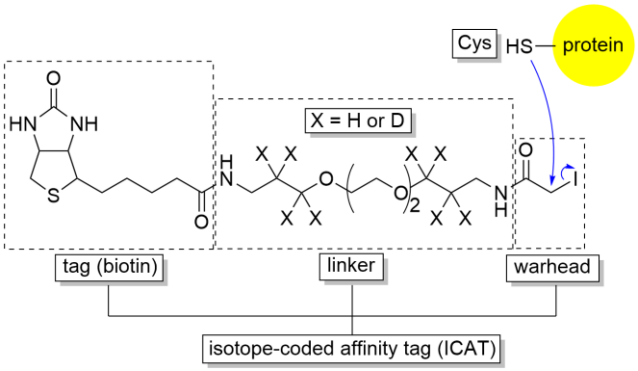


This method has a remarkably high resolution (> 10^3 "spots" per pH unit). But this is still not enough when the proteins to be separated are very similar in size and isoelectric pH. Migration is also affected by PTM, so that a single protein can undergo different PTM and be found at several spots.
One strategy to improve this complication is to reduce the size of the protein mixture and focus on specific proteins (e.g. those with similar functions or carrying similar functional groups) instead of the whole proteome.

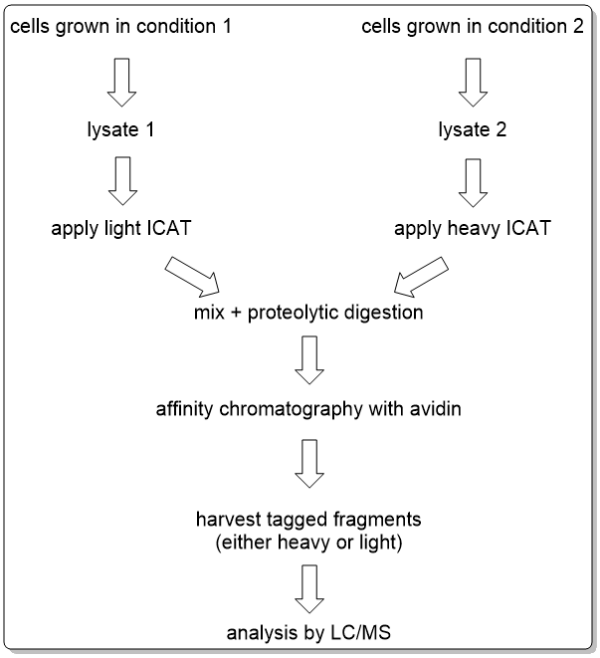
- A common approach to reduce proteome size is to selectively tag the proteins of interest and isolate it using affinity chromatography. To do this, we need to exploit the unique reactivity of proteins (based on the different side chains they carry).
Cysteine, for example, is the only AA that contain a thiol group. Among the ~350000 tryptic fragments generated by proteolysis of the yeast proteome, only ~30000 contain cysteine residues, which, however, cover ~90% of the ~6000 gene products.
A tag consists typically of three parts: the actual tag which allows affinity chromatography, the linker which prevents unwanted interactions between the tag and the protein, and the "warhead" which selectively binds to the target protein.



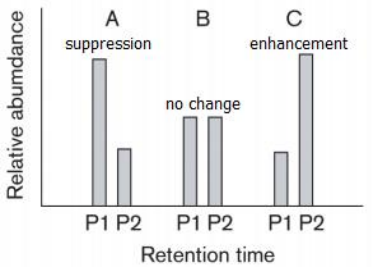
Example 1: ICAT (isotope-coded affinity tag)
The tag in ICAT is a biotin, which binds specifically to avidin or streptavidin.
The warhead binds covalently to the target protein by reacting with the thiol groups in cysteine residues.



The linker of ICAT is a long chain which can be either heavy (carrying deuterium) or light (carrying hydrogen). The two variations allow quantitative comparison of cells grown in different conditions.



The difference in hydrogen and deuterium leads to a small shift of peaks in LC/MS, allowing the two variations to be distinguished and quantified.

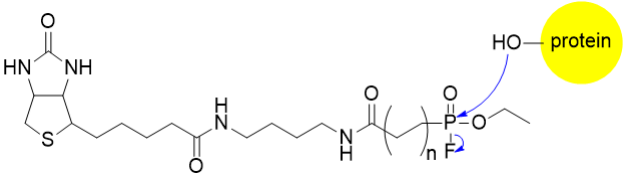


Relative abundance of proteins in two proteomes (P1 & P2) is determined by the ratios of differentially mass-tagged peptide pairs

Example 2: activity-based protein profiling (ABPP)

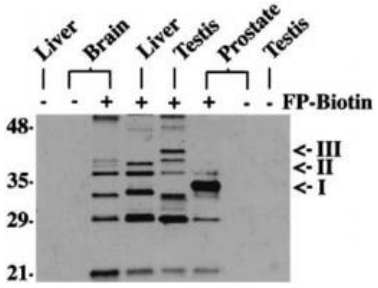
ABPP tags use a specially designed “warhead” that binds specifically and covalently to a residue in the active site of an enzyme or a class of enzymes. An enzyme that is inhibited or post-translationally modified will not react with ABPP tag.

For example, to test the activity of serine hydrolases (enzymes that possess an activated Ser residue in their active site), biotin is used as tag and a fluorophosphonate used as “warhead”.



Applications of ABPP:

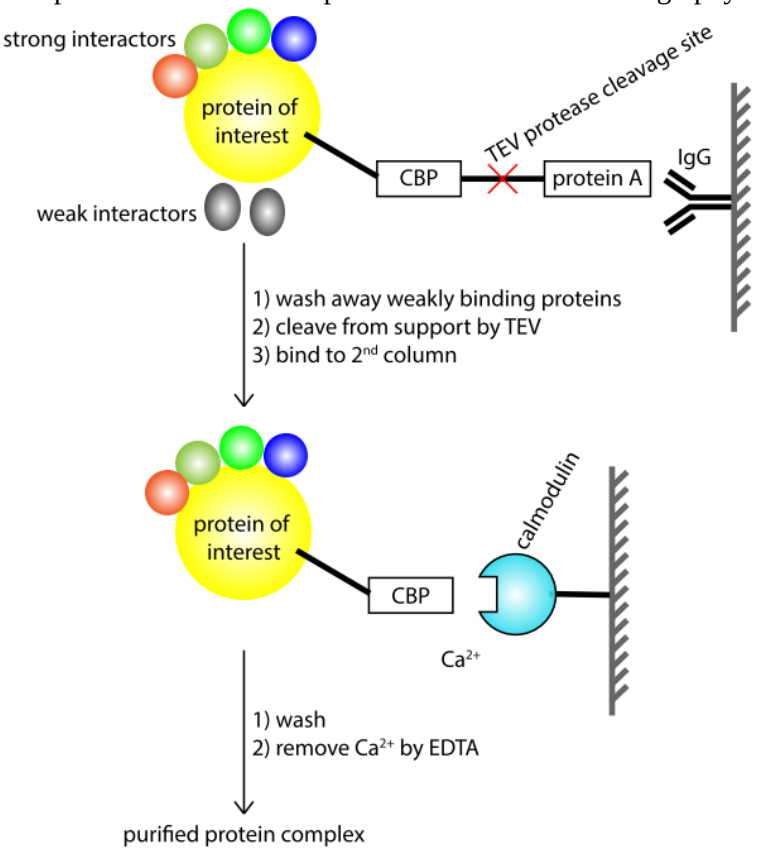
- Analyse tissue specific expression of certain class of enzymes



- Monitor temporal changes in enzyme activity
- Compare pathological and physiological states (e.g. cancer cell vs. normal cell)
- Assign function to uncharacterised proteins

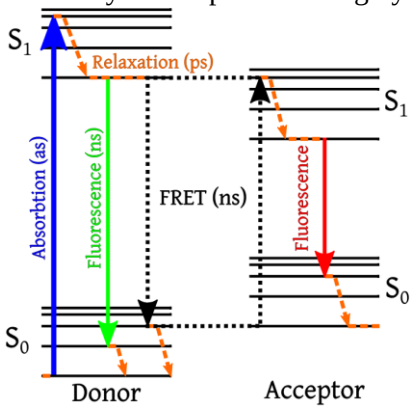
Protein-protein interactions

- Systematic analysis of complexes that proteins form in the cell is a further step to proteome analysis, which only determines which proteins are present.
- The POI is often expressed as a hybrid protein in which it is fused with a tag protein. This tag protein may serve purification (e.g. GST, His_6), localisation in microscopy (e.g. GFP) and targeting to a certain cell compartment (e.g. Arg_8).
- Tandem affinity purification (TAP)
  TAP is a technique used to purify a protein complex, in other words the POI and its binding partners.
  In TAP, the POI is tagged with a calmodulin binding protein (CBP), which in turn is linked to a protein A tag that gets recognised by an antibody. Between these two tags there is a cleavage site for tobacco etch virus protease (TEV protease).
  The procedure of TAP comprises two column chromatography steps.



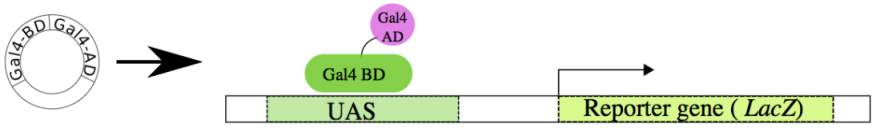
In the first column, immobilised IgG for protein A is used to pull the POI from cell lysate, together with its binding partners. Due to unspecific binding, a protein might carry many weak interactors. These unspecific binders are washed away under mild conditions, so that the real binding partners remain bound to the POI. The complex is then cleaved from the solid support by TEV protease, which cuts between the first (CBP) and second tag (protein A).
In the second column, calmodulin is immobilised on the solid support and used to pull down the protein complex. After a second washing step, the cofactor Ca^{2+} is removed by adding EDTA, leading to release of the complex.
Finally, the purified complex can be analysed by LC/MS.

- In vivo detection by fluorescence resonance energy transfer (FRET)
  In experiments utilising FRET, the POI is labelled with a fluorophore (often GFP). A second protein is labelled with another fluorophore whose excited state has a lower energy in excited state and a different excitation wavelength.
  When the two labelled proteins come together into close proximity, the first fluorophore might transfer its energy to the second one through nonradiative dipole-dipole coupling. The efficiency of this process is highly dependent on the distance between the two fluorophores.



  Therefore, if the two proteins form a complex, a change in the wavelength of the emitted light should be expected.

- Two-hybrid
  This technique is used to determine if two candidate proteins interact with each other. It is based on the activity of a naturally occurring transcription factor called Gal4, which consists of two domains: (1) the DNA binding domain (BD), which binds to a specific DNA sequence called UAS, and (2) the activating domain (AD), which binds to a nearby gene and activates its transcription. In natural Gal4, these two domains are linked by a flexible linker.

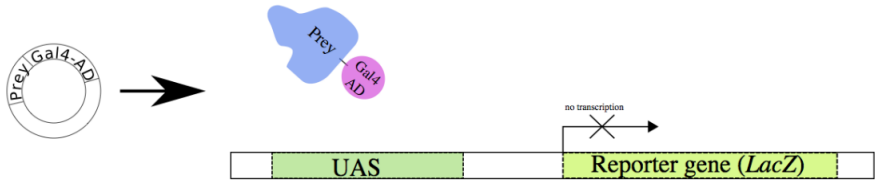


A. Regular transcription of the reporter gene

  In a two-hybrid experiment, the gene coding for the natural Gal4 protein is replaced by two hybrid genes. In the first hybrid gene, the coding sequence for Gal4's DNA binding domain is fused to the gene of the POI (the “bait”).

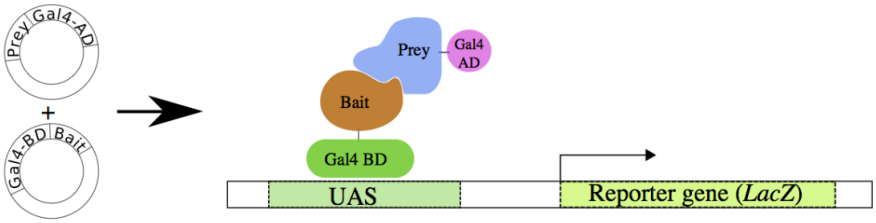
B. One fusion protein only (Gal4-BD + Bait) - no transcription

  In the second hybrid gene, the coding sequence for a potential binding partner (the “prey”) is fused to the coding sequence of Gal4's activating domain.
  One can also test a large number of potential interactors in a single experiment by using a library of proteins fused to the activation domain.



C. One fusion protein only (Gal4-AD + Prey) - no transcription

In case the "bait" and "prey" portions of the two resulting hybrid proteins bind to one another, Gal4 activating domain will be recruited to the vicinity of the reporter gene's promoter, thus stimulating transcription of the reporter gene. Reporter genes typically employed in two-hybrid experiments are genes that generate an easily detectable signal when activated, such as LacZ and GFP.

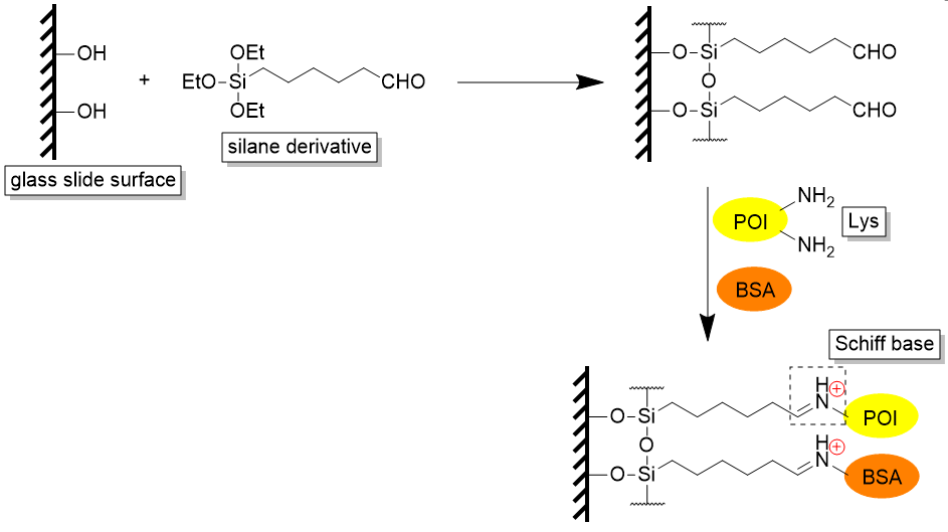


D. Two fusion proteins with interacting Bait and Prey

Despite its elegant design, this technique generates a lot of false positives because regulation of transcription is very sensitive to external influences. Also, interactions including more than two proteins cannot be detected by the two-hybrid method.

- Protein microarrays

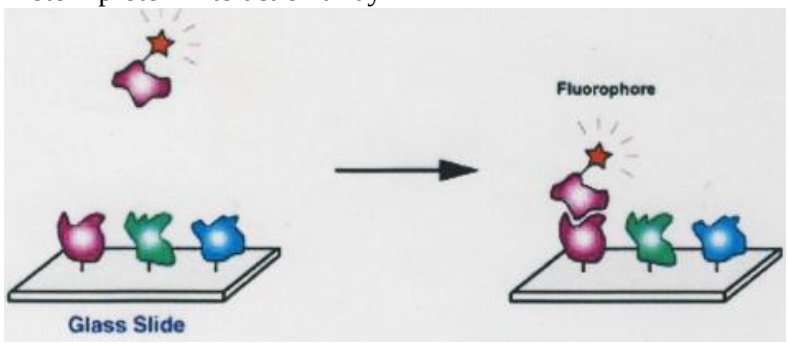
With protein microarrays, one can quantitatively measure the interactions between a large number of proteins. The technique is very flexible and provides reproducible result.
In a typical protein microarray experiment, the proteins are covalently attached onto a solid surface. For example, one can produce glass slides that are functionalised with aldehyde by using a silane derivative containing a terminal carboxyl group. POIs are attached to the surface through the reaction between aldehyde and amine (from surface Lys residues), which forms a Schiff base. Then, an excessive amount of BSA is added to cover all unreacted spots.



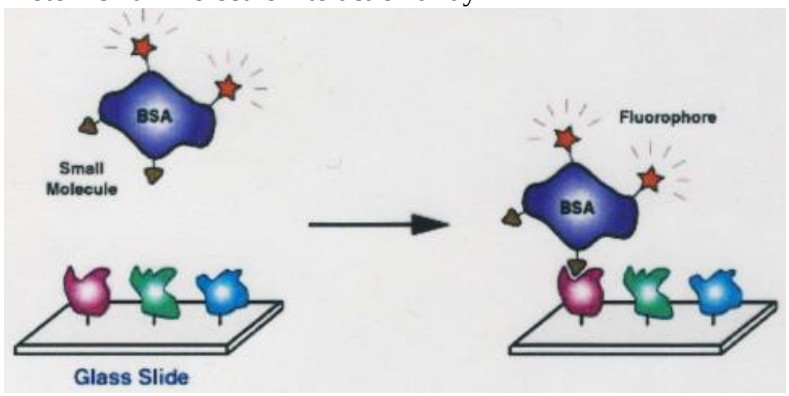
Nowadays, protein chip production can be automated by using high precision printing, with spots as small as 100 μm in diameter. In each cm^2 there can be over 1600 different spots. After immobilising the POIs, different potential binding partners are applied to the chip. These binding partners are usually labelled with different fluorophores. After washing away all unbound proteins, the binding events are directly visible under microscope and their intensities are reflected by the intensity of fluorescence on the corresponding spot. The k_D value can also be estimated by applying different concentrations of the same binding partner.

People have developed various types of protein microarray to test different binding partners and different properties.
Some examples:

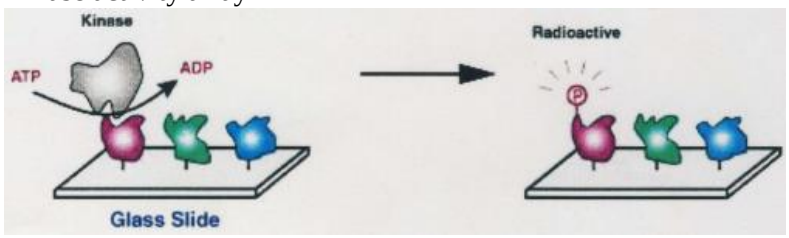
- Protein-protein interaction array



- Protein-small molecule interaction array



- Kinase activity array

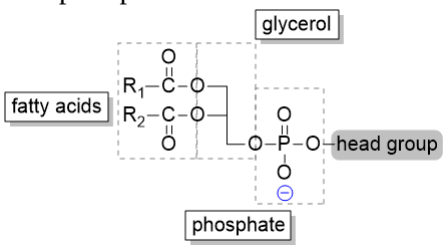


Lipid chemistry 脂类

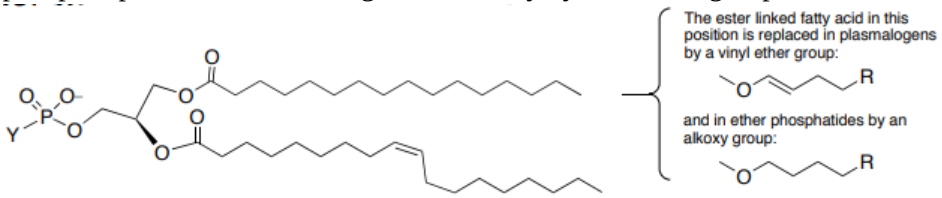
(23.04.2018)

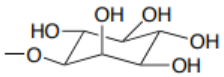
Lipid in membrane barriers

- Compartmentalisation of the cell by various membranes plays an essential role in cell metabolism. It:
  → links genotype and phenotype
  → serves communication (inside-outside, cell-cell, cell-environment)
  → determines cell shape & motility
- Membrane barriers are bilayer sheets that are 1-2 molecules thick. The main components are lipids and proteins (ratio ranges from 1:4 to 4:1) and a minor part of carbohydrate (1% - 5%).
- Three major types of lipids are found in the membrane. These lipids serve not only compartmentalisation but also energy storage and signalling.
  - Phospholipids

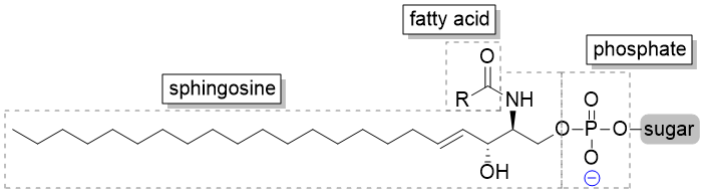


Both the tails (fatty acids) and the head group of phospholipids are variable. Different phospholipid classes are distinguished mainly by their head group.

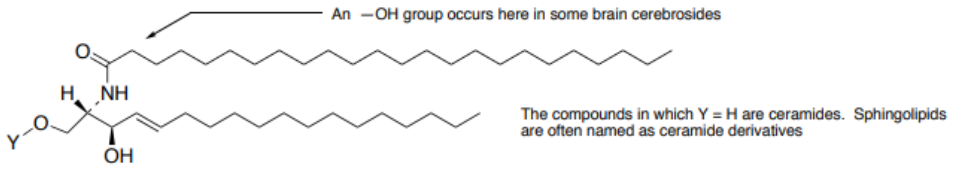


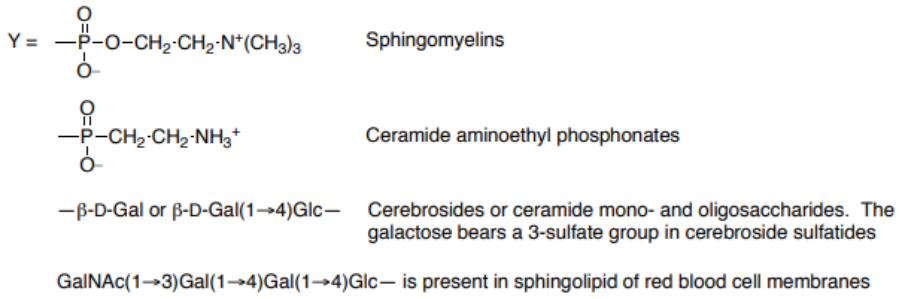
Y = —OH	Phosphatidic acid
—O—CH_2·CH_2·$N^+(CH_3)_3$	Phosphatidylcholine (lecithin, PC)
—O—CH_2·CH_2·NH_3^+	Phosphatidylethanolamine (PE)
—O—CH_2—CH(—CO_2^-)—NH_3^+	Phosphatidylserine (PS)
—O—CH_2—CH(—OH)—CH_2OH	Phosphatidylglycerol
—O—CH_2—CH(—OH)—CH_2O—	Diphosphatidylglycerol (cardiolipin), a special component of bacteria and mitochondria
	Phosphatidylinositol (PI)

  - (Glyco)sphingolipids

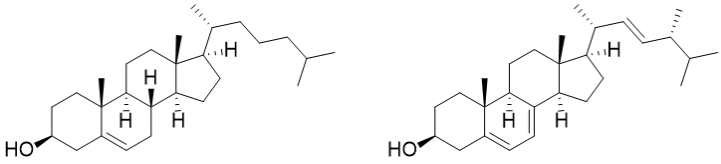


Like phospholipid, sphingolipids can also be classified by their variable head group and fatty acid residue. The head group of sphingolipid is often a sugar residue (either single sugar or oligosaccharide).





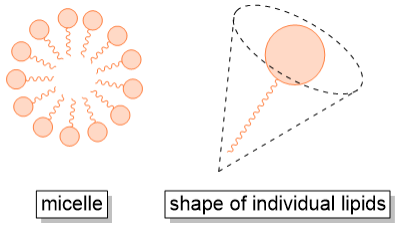
- Sterols



The most important sterol in mammals is cholesterol. In fungi and protozoa, the role of cholesterol is taken by ergosterol etc. and in plants by the phytosterols. Prokaryotes have been found to contain no or little sterol.

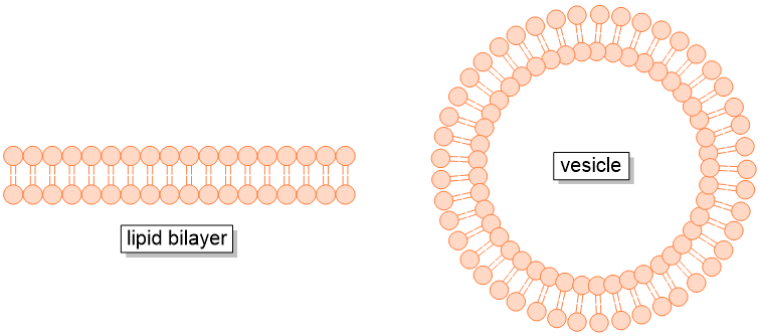
- A common feature of membrane lipids is that they are amphiphilic (i.e. consist of a polar part and an apolar part) and show properties similar to detergents. This leads to spontaneous formation of polymorphic aggregates, such as micelle and vesicle. Individual components of these structures tend to be highly dynamic.
  - Micelle

    At low concentrations, the lipids are dispersed in solution. However, when the critical micelle concentration (CMC) is reached, the lipids form spherical aggregates in which the polar groups are on the outside and the hydrophobic part buried in the interior. The driving force for this process is van-der-Waals interaction between the hydrophobic tails of the lipids. The spherical structure is obtained because individual lipids have a cone shape, which packs easily into a sphere.

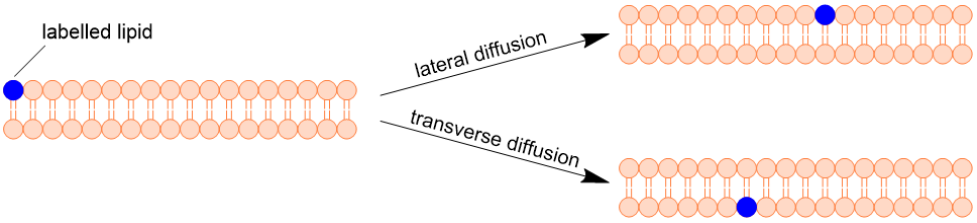


  - Vesicle / liposome

    Vesicles are bilayer structures typically consisting of membrane lipids with two hydrophobic tails. Both the interior and exterior of vesicle are aqueous.



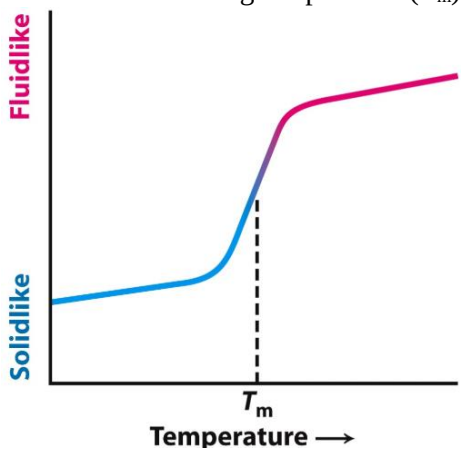
- Bilayer dynamics
  Inside the bilayers, lipids can move laterally or transversely. The speed of movement can be determined by labelling a lipid (e.g. fluorophore) in the membrane and monitoring its migration within the bilayer.
  - Diffusion



    The distance of diffusion s follows the equation:

$$s = \sqrt{4Dt}$$

    where D is the diffusion coefficient.
    Lateral diffusion turned out to be quite fast. In most lipid bilayers, D is found to be ~ 1 $\mu m^2/s$ for lateral diffusion. In contrast, transverse diffusion is much slower, with a half-life of 12 – 24 hours.
  - Axial rotation
    A lipid can rotate along its axis. This is also a very fast process.
  - Intrachain movement
    Most C-C bonds in the fatty acid chains are rotatable and are highly dynamic. This movement leads to kinks and flexes in the chain.
- Phase transition
  A lipid bilayer can be think of as a 2-D liquid. It can exist as a gel ("solid") at lower temperature or as a liquid crystal ("fluid") at higher temperature. The speed of diffusion was found to be much faster in the fluid phase.
  By monitoring the melting process of simple lipid bilayers, a sharp transition with a characteristic melting temperature (T_m) was observed.



  This melting temperature is dependent on chain length, degree of unsaturation, and head group.

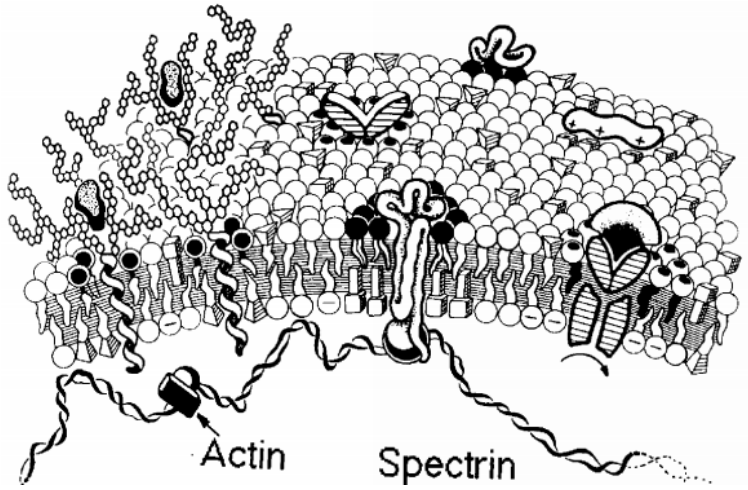
Longer fatty acids chains lead to a large surface for packing and increases T_m. Unsaturated chains, however, introduce permanent kinks into the chain and lowers the T_m. Stronger hydrogen bonding, less steric hindrance in the head group also contributes to a higher T_m.

Lipid	**Fatty acid chains (length : degree of unsaturation)**	**T_m (°C)**
dimyristoyl phosphocholine (DMPC)	di-14:0	23
dipalmitoyl phosphatidylcholine (DPPC)	di-16:0	42
distearoyl phosphatidylcholine (DSPC)	di-18:0	58
oleoylstearoyl phosphatidylcholine (OSPC)	18:1 & 18:0	3
dioleoyl phosphatidylcholine (DOPC)	di-18:1	-20
dipalmitoyl phosphatidylserine (DPPS)	di-16:0	55
dipalmitoyl phosphatidylethanolamine (DPPE)	di-16:0	63

Bilayers containing a mixture of different lipids often show very complex behaviour and a very broad melting transition. Their phase is dependent on various factors including lipid composition, temperature, pH, ionic strength, metal ions etc. Mechanisms behind these behaviours are still not fully understood.

Biological membranes

- Biological membranes are incredibly complex.



A typical biological membrane contains 10^2 to 10^3 distinct lipid species. The exact lipid composition varies greatly between different membranes.
Biological membranes have many associated proteins, which comprise up to 30% of the whole proteome.
Many biological membranes are also extensively modified with ECM, glycocalyx, cytoskeleton etc.

- Fluidity is essential for the function of biological membranes.
  In mammals, fluidity is regulated by sterols. Below the T_m, sterols prevent dense packing of the fatty acid chains and increases fluidity. Above the T_m, however, sterols decrease fluidity by preventing drastic movements of the lipids. (The melting temperature of sterols is higher so they move slower than lipids.) Sterols are thus called the "buffer" of fluidity.
  In prokaryotes, the fluidity of membrane is regulate by lipid composition. The ratio of saturated/unsaturated fatty acids in prokaryotic membrane is highly variable. There are temperature dependent promotors that activate the expression of unsaturated lipids.
- Fluid-mosaic model
  The fluid-mosaic model of membrane describes biological membrane as a 2-D lipid "sea" on which protein "icebergs" float.

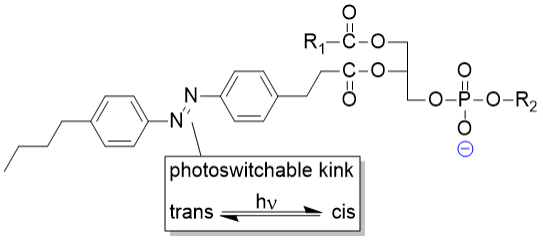
Composition of the “sea” is not consistent throughout the membrane. Individual lipids can have phase separation, i.e. complex mixture of lipids & proteins can form phase-separated microdomains, which also has an effect on their biological functions.

- Artificial liposomes

  A liposome is any lipid bilayer that define an interior volume.

  Liposomes can be produced artificially using unnatural synthetic lipids, often consisting of a charged head group and two hydrophobic tails.

  Modified phospholipids with interesting properties (such as a photoswitchable kink) have also been produced.



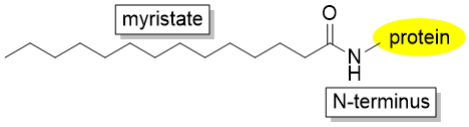
  Artificial liposomes can be used to produce:
  - model membranes (to investigate evolution etc.)
  - artificial organelles & cells
  - delivery vehicles (nucleic acid, protein, small molecule etc.)

Membrane proteins 膜蛋白

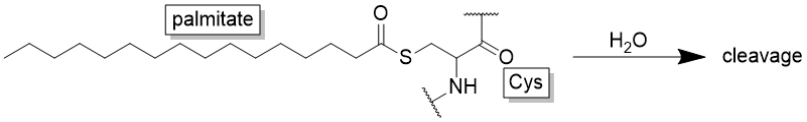
(23.04.2018 & 30.04.2018)

General

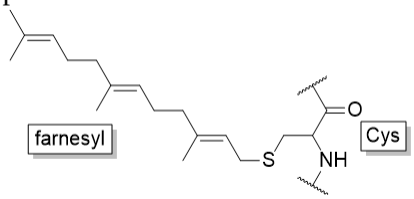
- Membrane proteins can be divided into intrinsic proteins, which are tightly associated with the membrane, and extrinsic (peripheral) proteins, which are loosely associated. Intrinsic proteins typically contain many hydrophobic residues on the surface and are integrated into the membrane through hydrophobic interactions the lipid tails. While intrinsic protein can only be isolated using detergents or organic solvents, extrinsic proteins can be removed from the membrane under certain milder conditions (e.g. pH, EDTA).
- To determine whether a protein of unknown structure is intrinsic or extrinsic, one can apply bioinformatic methods to calculate a hydrophobicity map based on AA hydrophobicity (obtained by H_2O/octanol partitioning) and compare it with sequences of known structure.
- Membrane association can also be mediated by PTM, which typically links a hydrophobic molecule to the protein to keep it anchored to the membrane.
    - Fatty acid appendage
      A protein can be appended to a fatty acid molecule.
      In the example of myristate appendage, the protein is linked to myristate through its N-terminus and then directed to the membrane.

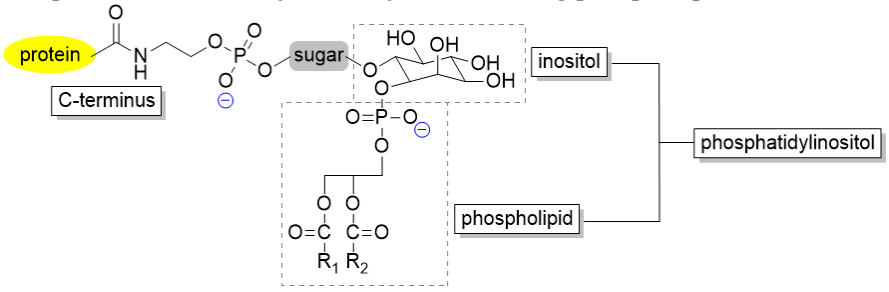
      In case of palmitate, the link occurs on a Cys residue. This link is easily hydrolysable.

    - Prenylation
      Compared to fatty acid appendage, anchoring to prenyl derivatives is a rather permanent modification.

    - GPI-anchor
      GPI-anchors are appended to the C-terminus of a protein and has the structure protein-phosphate-sugar-phosphatidylinositol. The phosphatidylinositol part of GPI anchor is part of the membrane.
      The proteins can be enzymatically cleaved using phospholipase C.

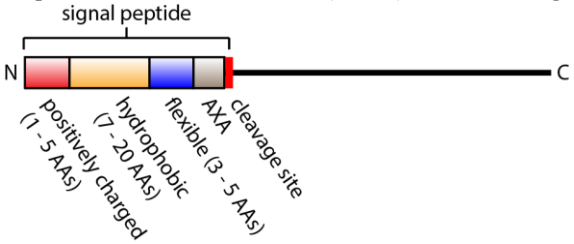
      

- Protein dynamics in membrane
  Similar to lipids, proteins show very fast axial rotation.

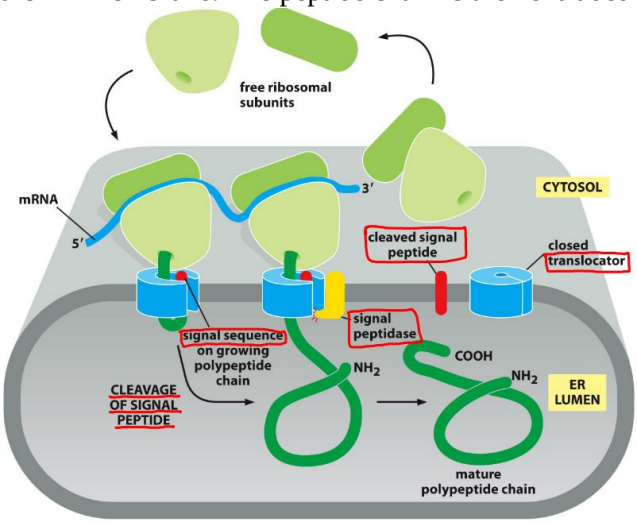
The lateral diffusion of proteins varies in rate, with diffusion coefficient D ranging from 10^{-1} to 10^{-4} $\mu m^2/s$. This rate is affected by the size (and aggregation) of the protein and its interaction with ECM and cytoskeleton. Large proteins tend to migrate slower.
Transverse diffusion is nearly impossible to proteins, unless a specialised enzyme (flipase) is present.

Protein trafficking & targeting

- Proteins are produced inside the cell, but many of them end up outside the cell, so that there is a topological problem of transporting proteins between the interior and exterior of the cell. Mechanisms to transport proteins include:
    - spontaneous insertion into membrane
    - transporter protein (translocon)
    - trafficking via coated vesicle
- Signal hypothesis
  This hypothesis suggests that proteins are directed to the correct (sub)cellular locus through peptide tags, such as the N-terminal leader (signal) peptide that directs secretion.
  Until now, more than 100 signal peptides have been found. These peptides are diverse in sequence but have similar in length (~ 30 AAs) and overall structure. During maturation of the protein, the signal peptide is proteolytically cleaved.
  N-terminal signal peptides typically begin with a few positively charged residues (1 – 5 AAs, probably helping the sequence to attach to negatively charged membrane). Next comes a hydrophobic domain (7 – 20 AAs) followed by a flexible linker (3 – 5 AAs) and a short sequence of small residues (AXA). The cleavage site comes right after the small residues.



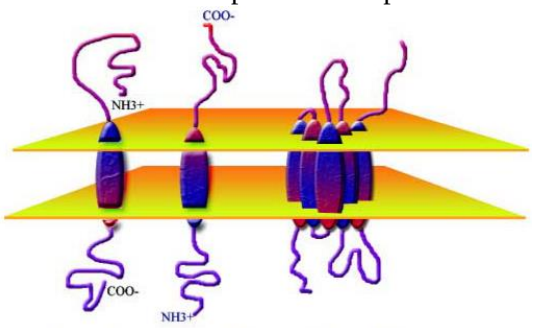
- Secreted proteins
  The first step of protein secretion is to translocate the protein into ER lumen (note that ER lumen is topologically the same as cell exterior). This often occurs co-translationally, where the N-terminal signal peptide is recognised by the signal recognition particle (SRP), which directs the translation complex (ribosome, mRNA and unfinished peptide) to a translocon on the ER membrane. The peptide chain is then extruded into ER lumen.



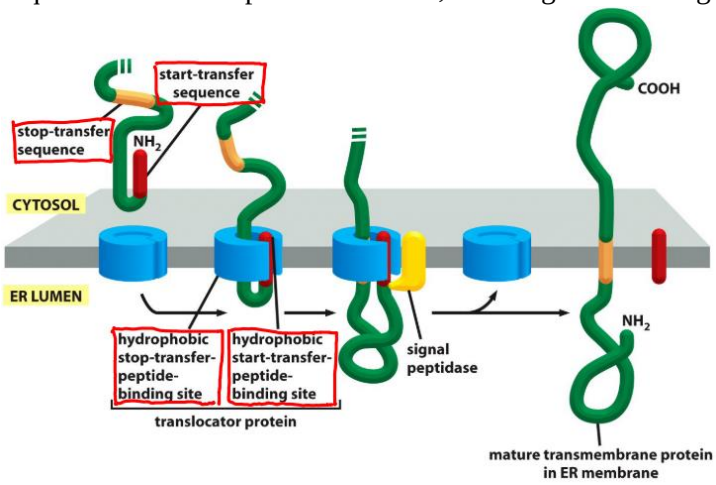
In the ER lumen, the protein undergoes co- and posttranslational modifications (glycosylation, oxidation etc.). After that, they are transported to the Golgi apparatus, where they experience more modifications before maturation. Finally, mature protein is transported to the membrane and secreted into cell exterior.

- Membrane proteins

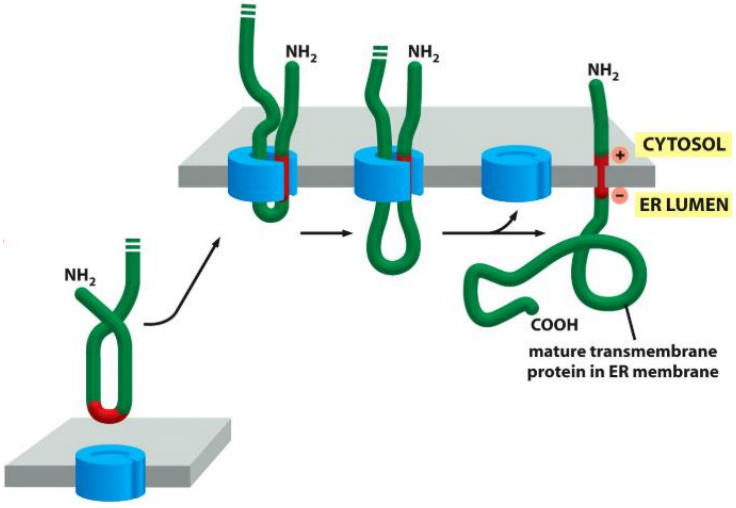
  Transmembrane proteins can have different topologies, which require different translocation mechanisms. But all transmembrane proteins tend to accumulate positively charged residues on the intracellular (cytosolic) side of the membrane ("positive inside rule").
  The transmembrane part of these proteins are usually hydrophobic helices.



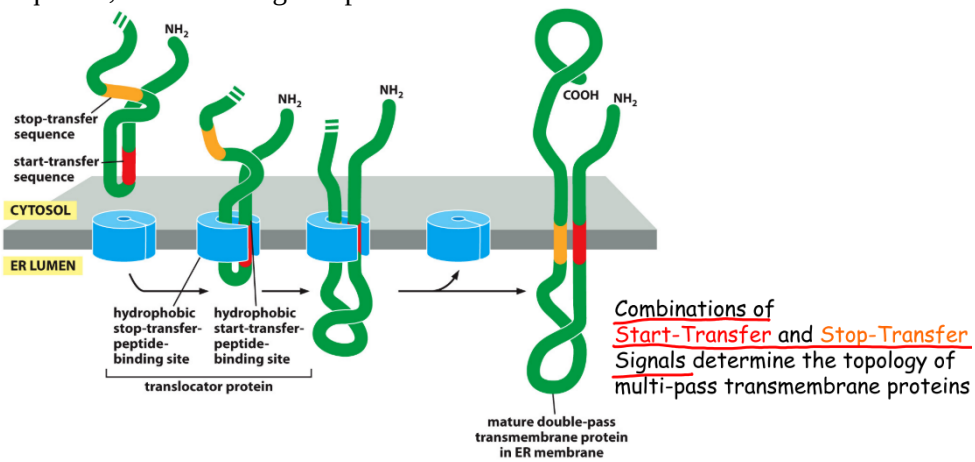
Membrane proteins with N-terminus on the extracellular side typically have a stop-transfer sequence that interrupts translocation, resulting in anchoring of the protein in the membrane.



Membrane proteins with N-terminus on the intracellular (cytosolic) can have its signal peptide not at the N-terminus but in the middle of the sequence. In this case, the signal peptide is not cleaved off during maturation, instead, it serves as an anchor in the membrane.

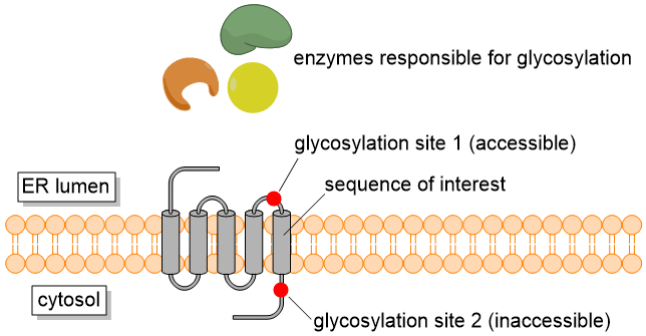


Multi-spanning transmembrane proteins possesses multiple start-stop transfer signals along its sequence, each resulting in a pair of transmembrane helices.



Multi-spanning proteins can also be cleaved during maturation or for generation of single-spanning proteins.

- Targeting peptide tags
  Peptide tags such as the signal peptides described above are very widespread. They serve as "address" for localisation of the protein. The targeted compartments include cell exterior (secretion), nuclease, mitochondria, ER, peroxisome etc.
  This provides a simple, transferable handle for practical applications, such targeting the POI to a specific location in the cell.
- Translocation vs. insertion
  It turns out that membrane proteins are equilibrium structures. For a helix sequence going through the translocon, there is always a competition between extrusion into the ER lumen and integration into the membrane. In the latter case, translocation stops and the sequence is released sideways from the translocon.
  To determine whether a transmembrane helix sits in the membrane or gets extruded, one can make use of the difference in accessibility of the extracellular and cytosolic parts.
  For example, one can put in two glycosylation sites that flank the sequence of interest. If the sequence is extruded into ER lumen, both sites will get glycosylated. In the case of membrane integration, however, only one site can be glycosylated.



The result can be quantified by determining the ratio between the di-glycosylated and mono-glycosylated forms.
Therefore, one can produce helices of different properties and determine their probability p to get inserted into membrane:

$$p = \frac{f_{1G}}{f_{1G} + f_{2G}}$$

where f_{1G} is the (relative) amount of mono-glycosylated proteins and f_{2G} the (relative) amount of di-glycosylated proteins.

In one experiment, a 19-AA peptide consisting of n Leu residues and 19 - n Ala residues was used to determine the relationship between Leu:Ala ratio and probability of insertion. The two ends of the helix are both marked with a GGPG sequence (i.e. GGPG-helix-GPGG).
The result showed that the Leu_{19} helix is nearly always inserted into membrane, whereas Ala_{19} is nearly always extruded. There is a sudden transition from p = 0 to p = 1 around n = 9. This can be explained by the fact that Leu is more hydrophobic than Ala and thus leads to more favourable hydrophobic interactions with the interior of the lipid bilayer.
This indicates that different AAs have their own propensities for membrane insertion.

- Membrane propensity of an AA can be determined by varying the identity of an AA in the middle of a transmembrane segment and calculating the free energy change of membrane insertion based on result of the two-glycosylation experiment:

$$k = \frac{f_{1G}}{f_{2G}} \quad \Rightarrow \quad \Delta G° = -RT \ln k = -RT \ln \frac{f_{1G}}{f_{2G}}$$

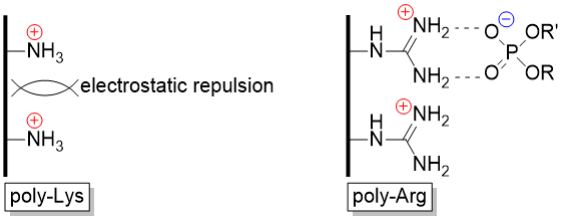
where f_{1G} is the (relative) amount of mono-glycosylated proteins and f_{2G} the (relative) amount of di-glycosylated proteins.

Amino acid residue	Transfer free energy (kcal mol-1)
Phe	3.7
Met	3.4
Ile	3.1
Leu	2.8
Val	2.6
Cys	2.0
Trp	1.9
Ala	1.6
Thr	1.2
Gly	1.0
Ser	0.6
Pro	-0.2
Tyr	-0.7
His	-3.0
Gln	-4.1
Asn	-4.8
Glu	-8.2
Lys	-8.8
Asp	-9.2
Arg	-12.3

As the table above shows, hydrophobic AAs tend to be integrated in the membrane than out in the solution. In contrast, charged AAs tend to be found out of the membrane.
There are also findings suggesting that membrane propensity of an AA is affected by its position in the helix. Introduction of a hydrophobic AA is most favourable in the middle of the helix and less favourable at the two ends (probably because the two ends are charged). For hydrophilic AAs, the situation is the opposite. Note that the membrane propensity of Gly is quite constant because of its flexibility as the smallest AA.

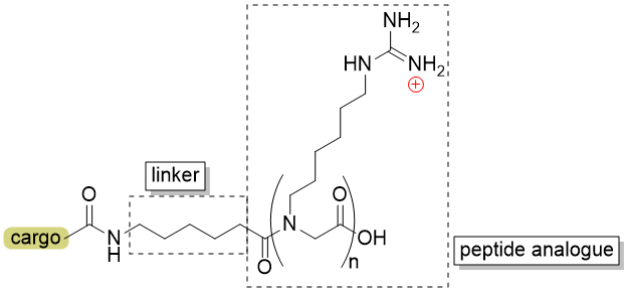
- Tags for delivery: cell-penetrating peptide
Tags that serve as “address” to deliver cargo molecules into (or out of) the cell are called cell-penetrating peptides (CPP).
Existence of this type of sequences was first discovered in an HIV protein called Tat, which is able to spontaneously cross cellular membranes. It was later found that there is a critical sequence that is responsible for this property, namely RKKRRQRRRR.
People have tried to optimise the structure of CPP by fusing different CPPs with a fluorophore and monitoring the uptake by FACS.
In one of these experiments testing a series of 9-AA CPPs, the relative intensities of uptake was: Lys_9 = Orn_9 < CPP of Tat < Arg_9 < D-Arg_9.

Arg performs better than Lys because poly-Arg has a higher pK_a (~ 14) than poly-Lys (9 – 10). Poly-Lys is tends to be only partially charged because neighbouring positive charges repulse each other, which reduces its ability to interact with the negatively charge membrane suface. This is less of a problem for the more basic Arg. In addition, Arg is bidentate, allowing it to interact with other molecules in a bidentate fashion to neutralise the effective charge.



D-Arg showed better result than Arg because they are less vulnerable to proteolytic digestion and are thus more stable in serum.

- There are many possible mechanisms of CPP-mediated uptake and these can be quite controversial. Some of the uptake mechanisms (such as that of Tat) are energy-dependent, while others are not. This can be tested by observing whether ATP deprivation or lowering temperature inhibits uptake.
  Typically, CPPs are cationic and binds to negatively charged heparin sulphate that coats the cell, bringing the cargo molecule into physical contact with cell membrane. The molecule is then engulfed into clathrin-mediated vesicles via pinocytosis or endocytosis. Later, these vesicles form endosomes.
  Main challenges for application:
  - How to prevent the molecule from being digested in endosome?
  - How to target the penetration to specific cells (e.g. cancer cells)?
- Unnatural CPPs have also been developed, such as by using peptoid, a peptidomimetic (peptide analogue) whose side chains have been moved from C_α to N.



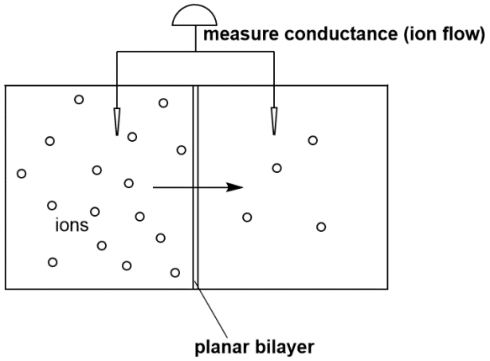
The efficacy of this artificial CPP depends on the structure of its linker and on the number n. Compared to normal peptides, peptoid CPPs are industrially easier to produce since they have no chiral centres.

Membrane transport 跨膜运输

(07.05.2018)

Ion transport & membrane permeability

- Functions of ion transport:
  - Regulation of pH and ion composition of cytosol organelles
  - Establishment of ion gradients (H^+, K^+, Na^+, Ca^{2+}, Mg^{2+}, Cl^-, etc.)
    → important fuels & building blocks
    → export hormones & toxins
    → signal transduction (e.g. in neurons)
    → ATP synthesis (e.g. mitochondria)
- Permeability is determined by how rapidly the equilibrium is achieved across a membrane. The values can vary by 10^8 in different cases.
  Permeability of a membrane for a particular ion can be measured by setting up an ion gradient across the membrane and measuring the conductance.



- Kinetics
  The kinetics of transport is highly dependent on the ion or molecule that is being transported. The membrane is nearly impermeable for ions and molecules that are charged or polar. In contrast, small molecules and hydrophobic molecules can pass the membrane quite easily.
  - Charged species (Na^+, Cl^-, etc.)
    $D = 10^{-12} – 10^{-8}$ cm^2/s (practically impermeable)
    $t_{1/2}$ = days – years
  - Hydrophobic neutral species (indole, O_2, CO_2, etc.)
    $D = \sim 10^{-4}$ cm^2/s
    $t_{1/2} < 1$ s
  - Polar molecules (urea, glucose, etc.)
    $D = 10^{-8} – 10^{-6}$ cm^2/s
    $t_{1/2}$ = minutes – hours
  - H_2O
    $D = \sim 10^{-2}$ cm^2/s
    $t_{1/2}$ = milliseconds
- Thermodynamics
  For neutral species, the free energy change during transport can be described as:

$$\Delta G = -RT \ln \frac{c_2}{c_1} = RT \ln \frac{c_1}{c_2}$$

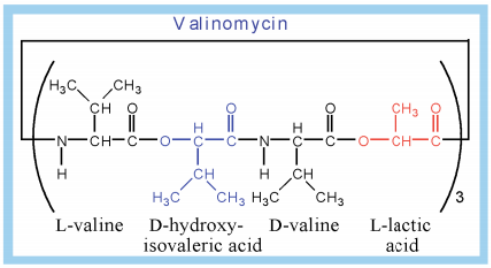
  where c_1 and c_2 are concentrations in the two compartments.
  For passive diffusion, transport can only occur when $c_1 > c_2$. For active diffusion, it is possible to transport the ion/molecule with $c_1 < c_2$ under energy consumption.
  In case of charged species, the equation for free energy is a bit more complicated:

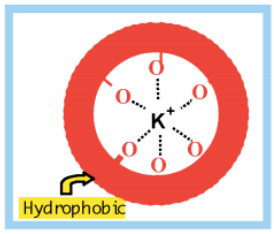
$$\Delta G = RT \ln \frac{c_1}{c_2} + zF\Delta\psi$$

where z is the charge of the molecule, F is Faraday constant and $\Delta\psi$ is membrane potential. This means that electric potential can also be the driving force for transport and that transport of charged species against electric potential requires energy.

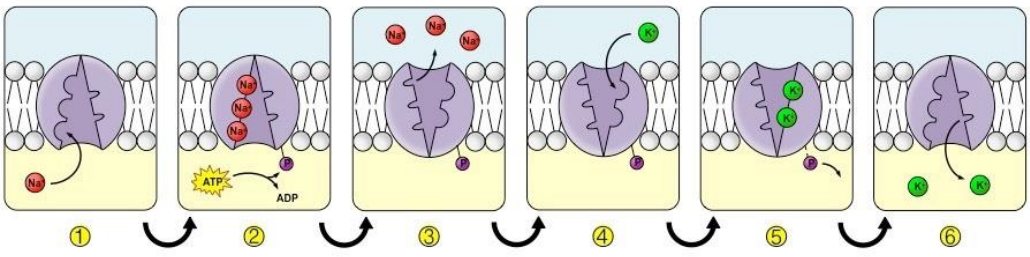
Carrier-mediated transport

- There are three major types of carrier proteins:
  - Ionophore (valinomycin, gramicidin etc.)





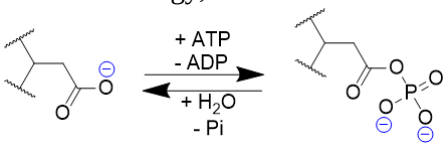
    An ionophore is a molecule that wraps itself around an ion, shields its charge and allows its to pass through the membrane. There are also synthetic ionophores available, such as crown ethers.
  - Pumps/transporters
    Pumps/transporters serve both active and passive transport. They usually convey a single or a few ions or molecules per cycle, because they have distinct binding sites that can be saturated. This also means that the maximal rate of transport is limited.
  - Pores/channels
    In contrast to pumps/transporters, pores/channels do not have specific binding sites. Instead, they just facilitate passive diffusion by creating a hole on the membrane and are thus not saturatable. Note that this hole is still highly specific.
    In addition, pores/channels typically have two different states, i.e. opened and closed. Switching between the two states is responsive to external stimuli. Pore/channels that response to change in membrane potential are call voltage-gated, while those regulated by ligand binding are called ligand-gated.
    Defects in these carriers proteins tend to cause disease, such as cystic fibrosis (deficient Cl^- channel) and chronic pain (deficient Na^+ channel).
- Example: Na^+/K^+-ATPase
  A very famous and important example of pump is the Na^+/K^+-ATPase. These pumps may consume 30% - 70% of all the energy of a resting cell.
  For most cells, cytosolic [K^+] (~ 150 mM) is much higher than extracellular [K^+] (~ 5 mM), whereas cytosolic [Na^+] (5 – 15 mM) is much lower than extracellular [Na^+] (~ 150 mM). This gradient is maintained by Na^+/K^+-ATPase using ATP as driving force.
  In each cycle, Na^+/K^+-ATPase hydrolyses one ATP and uses the energy to pump three Na^+ ions out of the cell and two K^+ ions into the cell.
  If membrane potential is $\Delta\psi$ = - 50 mV, the free energy change for each function cycle will be ΔG = + 9.9 kcal/mol, which can be covered by the free energy provided by ATP hydrolysis ΔG = - 12 kcal/mol.



Steps in each cycle:

1) Recruitment of Na^+ ions and ATP
2) Conformational change induced by Na^+ and driven by ATP hydrolysis. Release of ADP.
3) Release Na^+ ions
4) Recruit K^+ ions. Release of Pi.
5) Conformational change induced by K^+. Release of K^+ ions.

Chemically, the ATP hydrolysis step makes a PTM on an Asp residue. The two states are similar in energy, so that reaction will not get stuck in one the them.

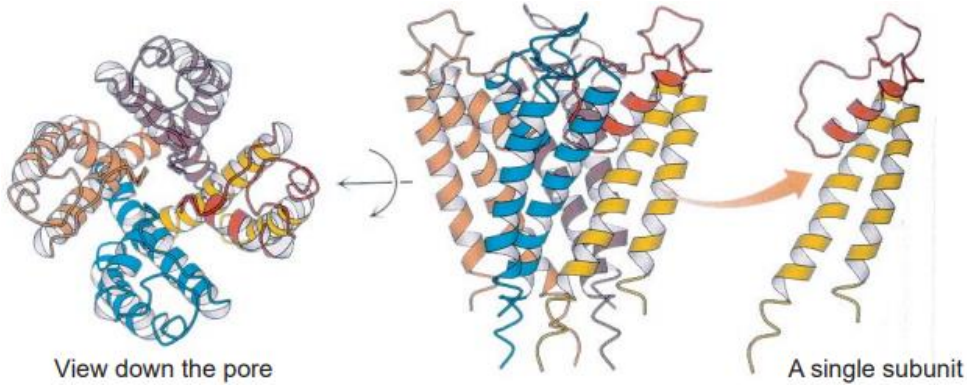


The gradient established by Na^+/K^+-ATPase plays an important role in the regulation of a cell's volume, its electric excitability etc. It also serves as the energy source for coupled transporters, such as the sodium-glucose linked transporter (SGLT) that pumps Glc and Na^+ into the cell.

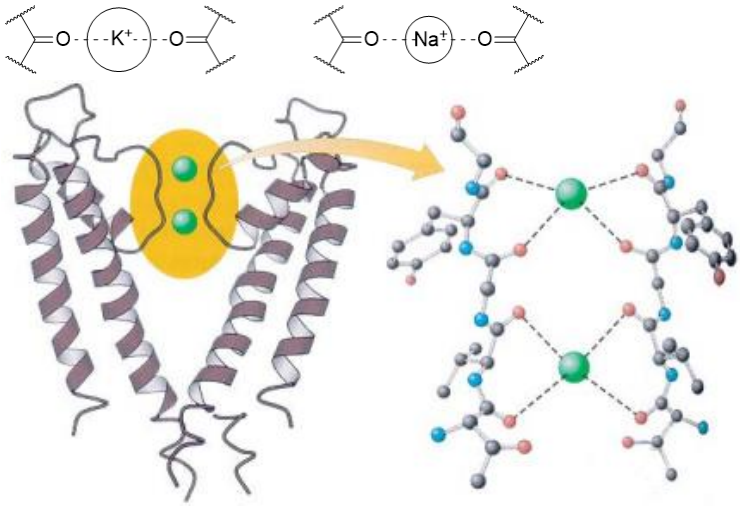
- Example 2: K^+ channel

The most famous channel, K^+ channel, is a voltage gated channel that is highly selective for K^+ over Na^+, with $k(K^+/Na^+) > 10^4$.

X-ray structure (Mackimon et al., 1998) shows that the protein is a helical bundle protein containing four helical subunits. The pore is constructed by four transmembrane helices, one from each subunit.

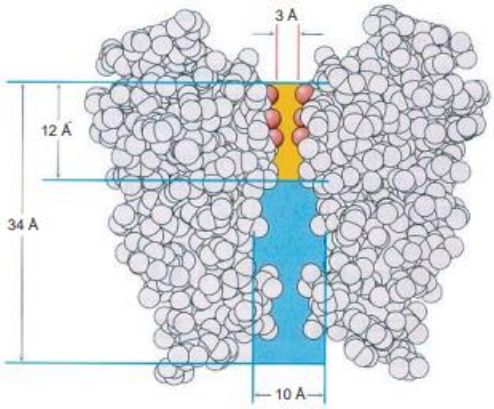


The channel distinguishes between two the ions based on their different radius (r(K+) = 1.33 Å, r(Na+) = 0.95 Å). Carbonyl groups projecting into the pore can replace the H_2O on K^+, allowing it get rid of the solvation layer and pass the channel. Na^+, on the other hand, is too small to make such contacts and is thus preventing from passing by its solvation layer of H_2O.

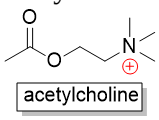


Gating the of channel is mediated by accessory proteins.

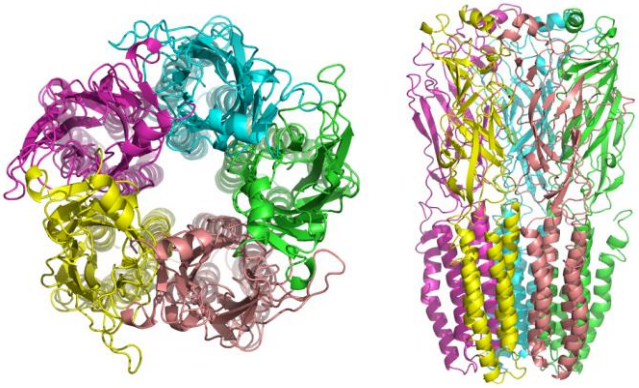
An unexpected property of the K^+ channel is that the part below the selection filter is filled with water, essentially reducing membrane width from 34 Å to 12 Å. As a result, the conductance of membrane is increased substantially.



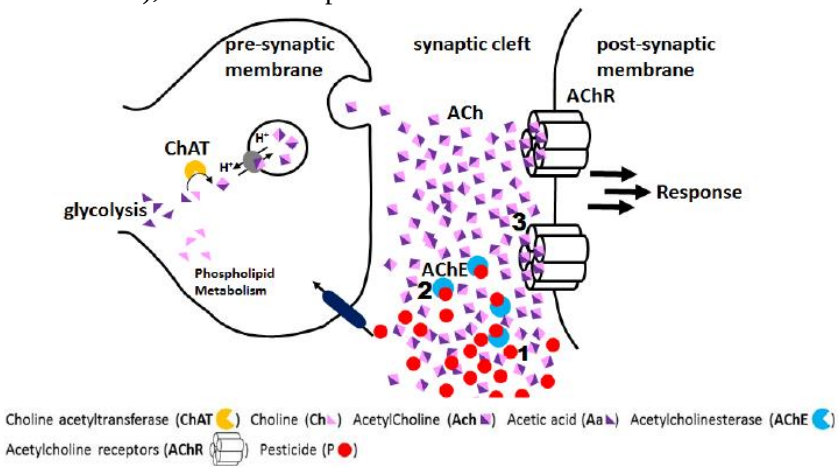
- Example 3: acetylcholine receptor (AChR)
  Acetylcholine receptors are ion channels gated by the molecule acetylcholine.



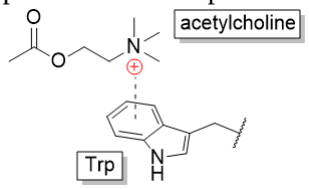
The complete structure of AChR is still not available. According to the current knowledge, the protein is very large (278 kDa) and contains multiple subunits ($\alpha_2\beta\gamma\delta$ in muscle) that are arranged in a ring. Similar to K^+ channel, the pore is formed by one helix from each subunit.



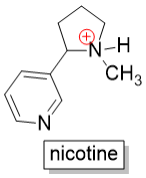
Binding of acetylcholine in the extramembrane domain induces a conformational change in AChR, which opens the pore and allows Na^+ ions to pass through into the cell.
At synapses, acetylcholine released by the pre-synaptic membrane is triggered by change in membrane potential (nerve impulse). Released acetylcholine in the synaptic cleft binds to AChR on the postsynaptic membrane, leading to collapse of its membrane potential (from -75 mV to 0 mV), so that the impulse is transmitted further.



The gating and activity of AChR by acetylcholine can be monitored by patch-clamp.
The binding of acetylcholine is thought to be the result of low-energy cation-π interactions between the ligand the a Trp residue in the protein. This interaction was proved to be crucial for acetylcholine binding because introduction of noncanonical AAs (such as F_nTrp) in the position of the Trp weakens the effect.

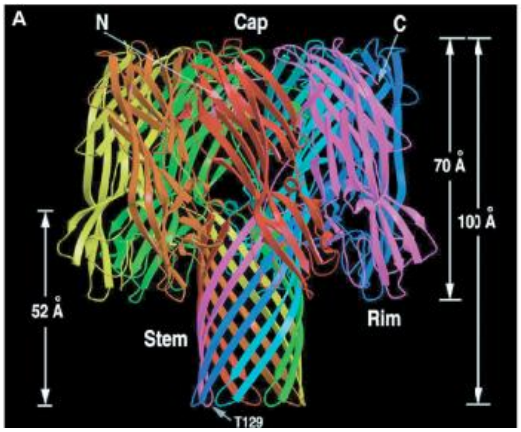


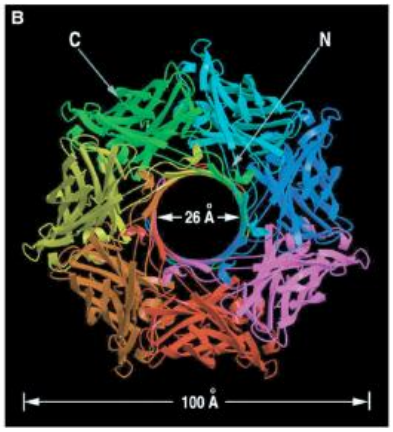
Nicotine can bind to AChR in the brain (but not in muscle) at the same site as acetylcholine does, but cation-π interaction in nicotine is much weaker. Instead, nicotine binding is mediated by H-bonding with a carbonyl group. It is believed that subtle changes in the binding pocket of muscular AChR prevents nicotine from binding.



Non-biological application of carrier protein: stochastic sensing

- The protein α-hemolysin is a bacterial toxin secreted by the bacteria *Staphylococcus aureus*. It is a heptomeric pore that spontaneously inserts into lipid membrane and is used by bacteria for defence.

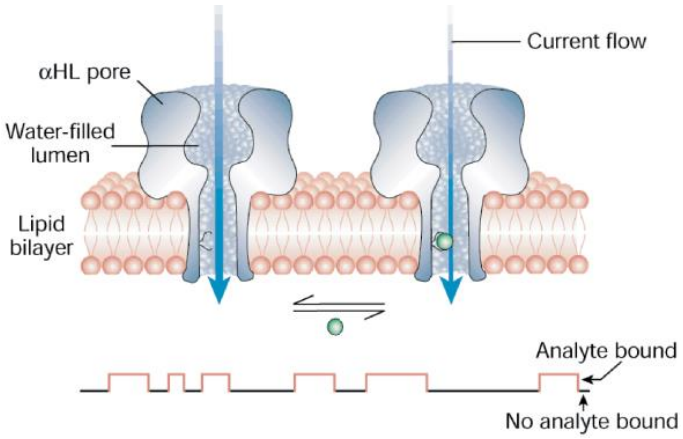




- α-hemolysin can be used to monitor ligand binding.
  To do this, α-hemolysin is first inserted into a planar lipid bilayer and the ion flow is monitored. In the resting state, a flow of 10^8 ions/s was observed, a rate that is close to diffusion.
  The interior of the pore can be engineered so that it specifically binds an analyte of interest. In the present of an analyte that can dock itself in the cavity of the pore, ion flow will be impeded when the analyte binds, and will be recovered when the analyte is released. This is a stochastic event and can be monitored using patch-clamp. From the frequency and duration of the events, binding constants such as k_{on} and k_{off} can be determined:

$$f_{\text{on}} = \frac{1}{k_{\text{on}} \cdot [\text{analyte}]} \qquad f_{\text{off}} = \frac{1}{k_{\text{off}}}$$

  where f_{on} is the frequency of events and f_{off} is the event duration.



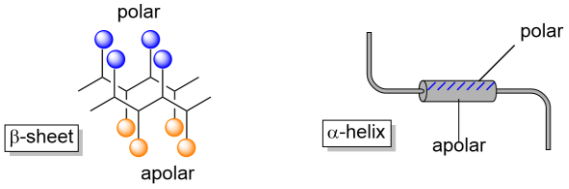
- Stochastic sensing with α-hemolysin be applied to many types of analytes, including small molecules, proteins, metal ions, nucleic acids etc.
  This method provides very sensitive binary response. The binding events are rapid, reversible and can be monitored in real-time. In addition, the dynamic range is very broad.
  Applications utilising this method include nanopore sequencing of DNA/RNA etc.

Protein design 蛋白质设计

(14.05.2018)

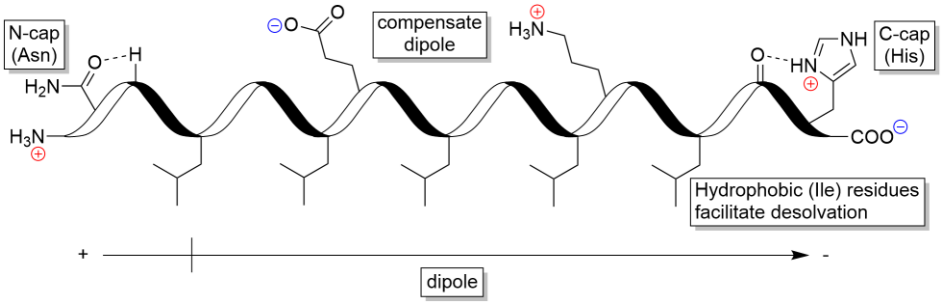
Prior knowledge

- What we know qualitatively about proteins:
  - The interior is very tightly packed
  - AAs are non-randomly distributed in a protein. Side chains have inherent propensities to adopt certain 2° structures. In a certain 2° structure, some AAs can be present with a higher probability than the other.
  - The inside tends to be apolar (hydrophobic), while the outside polar (hydrophilic). In fact, burial of apolar residues is often a driving force in protein folding.
    Corollary: 2° structures are often amphiphilic



→ We can favour the formation of certain 2° structures by alternating the properties of the AAs in such a way that we have a binary pattern of polar and apolar residues. This approach is called binary patterning.

- Interactions that stabilise α-helix

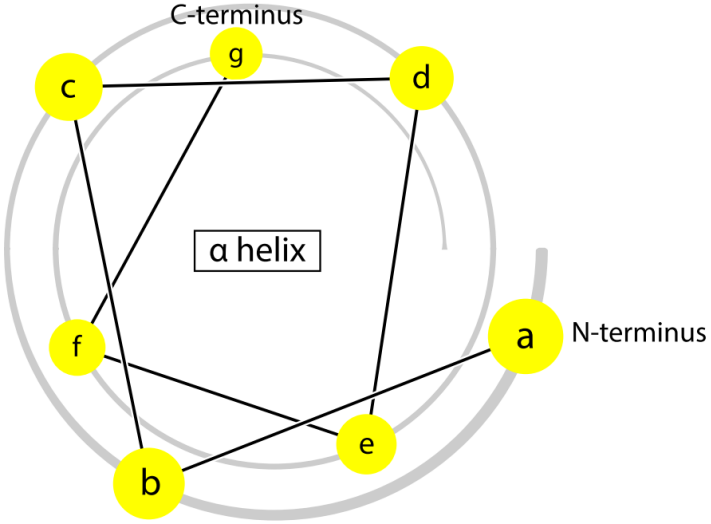


  - Helical propensities
    Residues with higher helical propensities tend to lower the energy of the helix.
  - Helix capping
    The four residues at each end of an α helix are not fully hydrogen bonded to neighbouring backbone segments. α helices are often flanked by residues such as Asn, Gln and His, whose side chains can fold back to form hydrogen bonds with one of the four terminal residues of the helix.
  - Dipole compensation
    The helix has a natural dipole, which can be stabilised by introducing charged residues (either intrachain or interchain) that compensate it.
  - Desolvation
    Van-der-Waals interactions between hydrophobic residues mediate desolvation of the helix, preventing it from being disturbed by solvent.

Feature	**ΔG contribution (kcal/mol)**
Helical propensities of residues	0 – 1
N-capping	1 – 2
C-capping	~ 0.5
Charge dipole	~ 0.5
Interchain	~ 0.5
Desolvation	2 – 5

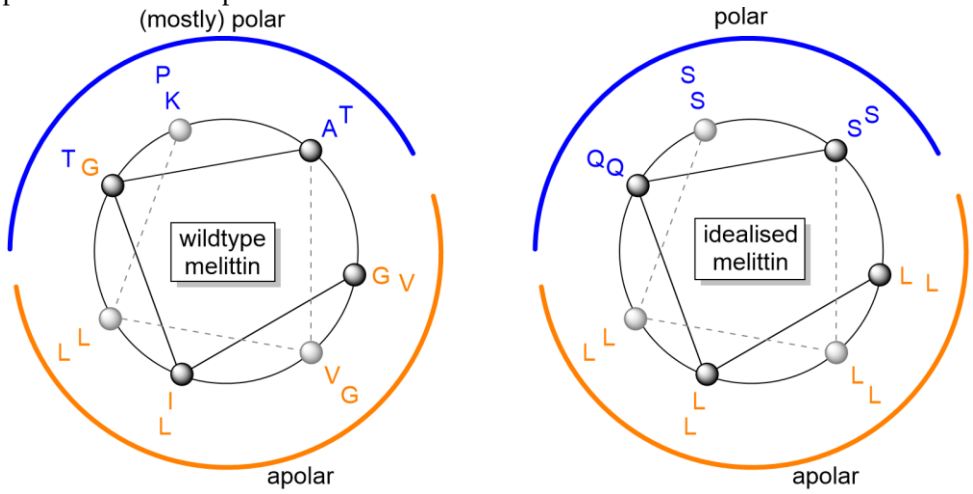
All these factors can be incorporated to bias a polypeptide to assume a helical conformation.

- Helical wheel projection
  We can draw the structure of an α-helix looking down from the end in the so-called helical wheel projection. Starting from the N-terminus, we see 7 residues (a - g) that almost come around completely (i.e. every seventh AA falls roughly onto the same position). This seven-AA loop is called binary pattern and can be used to design idealised α-helix structures.



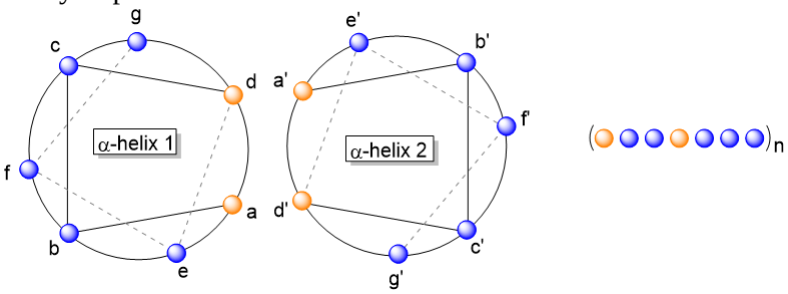
Protein design

- Melittin redesign (Kaiser et al.)
  Melittin is the major pain-producing substance in bee venom.
  The wildtype melittin peptide consists of 26 AAs, with the sequence HOOC-GIGAVLK-VLTTGLP-ALISWI-KRKRQQ-NH_2. The first 14 AAs are part of an α-helix structure, which, with the help of helical wheel projection, is shown to be hydrophobic on one side and (almost) hydrophilic on the other side. The helix is presumed to be broken by proline.
  The amphiphilic nature of this protein allows it to insert into the membrane and form bundles that orient the polar residues toward the interior, creating a hydrophilic channel (hole) on the membrane.
  People have managed to optimise (redesign) the structure of melittin in laboratory, with the new sequence HOOC-LLQSLLS-LLQSLLS-LLLSWL-KRKRQQ-NH_2. Here, partition of polar and apolar residues is clearer and the proline breaking the helix is removed. The Trp at position 19 was kept for concentration determination.

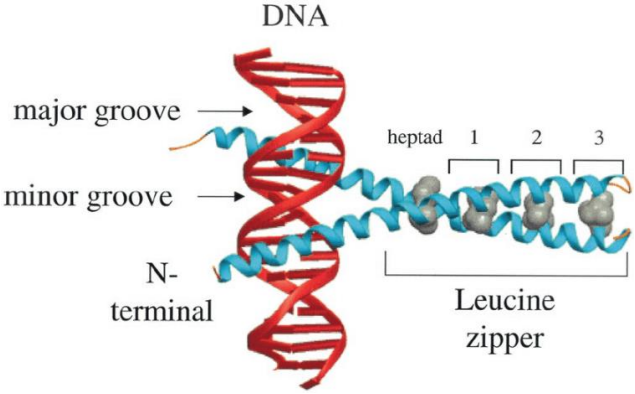


  This optimised melittin was shown to adopt a helical structure and it is also biologically more potent (i.e. performs better) than its wildtype counterpart.

- Parallel coiled coil
  A coiled coil consists of a bundle of two or more α-helices that interact with each other through one of their faces. Two α-helices coming together will tend to form a coiled coil (in aqueous solution) if the residues at their contact surface are hydrophobic and all other residues are hydrophilic.



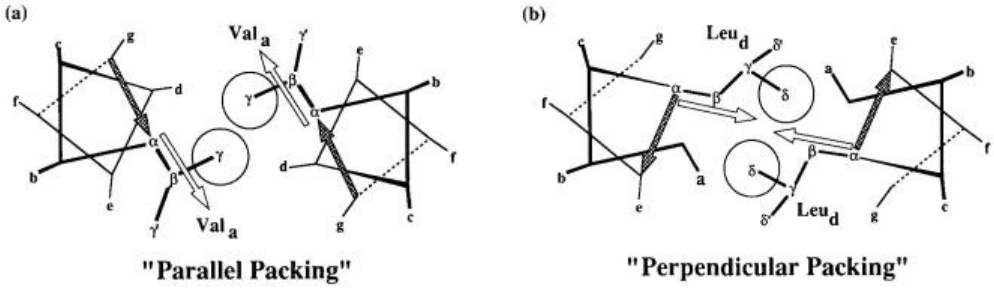
  Similar patterns have actually been observed in natural proteins. For example, a transcription factor Gcn4 was found to have a long, parallel coiled coil characterised by periodic repetition of leucine residues at every seventh position. This pattern is called a leucine zipper.



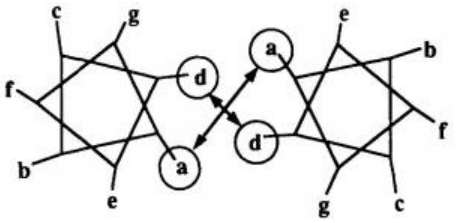
  Note that the relative positions of residues rotate after each cycle (seven residues = 700°, two turns = 720°). For long coiled coils, this rotation is compensated by (left-handed) supercoiling.
- The overall structure of coiled coil is influenced by specific choices of residues at the contacting face (positions a & d).
  In one experiment, a 34-AA long peptide fragment from Gcn4 with varied a/d residues was used to construct coiled coils (Kim et al.). The structure of the formed coiled coil turned out to be highly dependent on the choice between Ile and Leu (which differ from each other only sterically).

Position a	**Position d**	**Structure**
Ile	Leu	dimer
Ile	Ile	trimer
Leu	Ile	tetramer
Leu	Leu	diverse

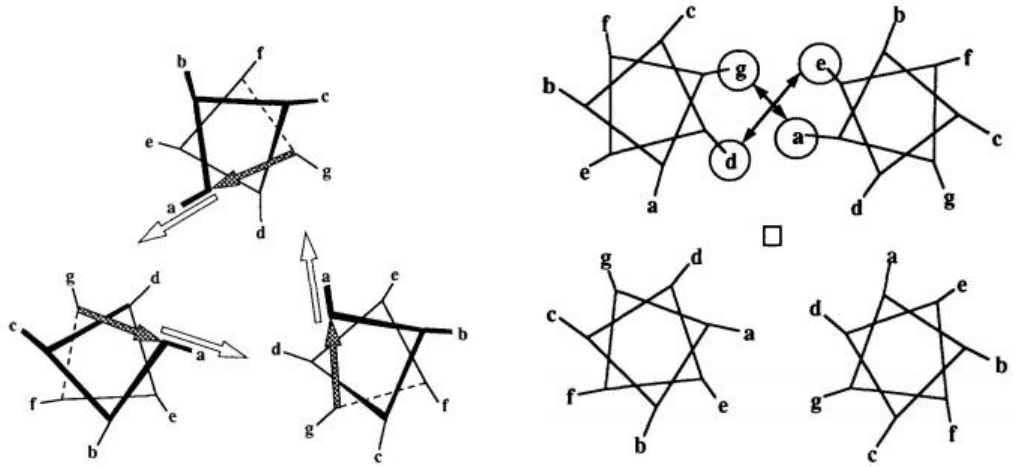
  This result indicates that packing of residues is critical to the topology of coiled coil.
  Packing modes in Gcn4 (notice the positions of the α-β vectors):



  Two-stranded coil:



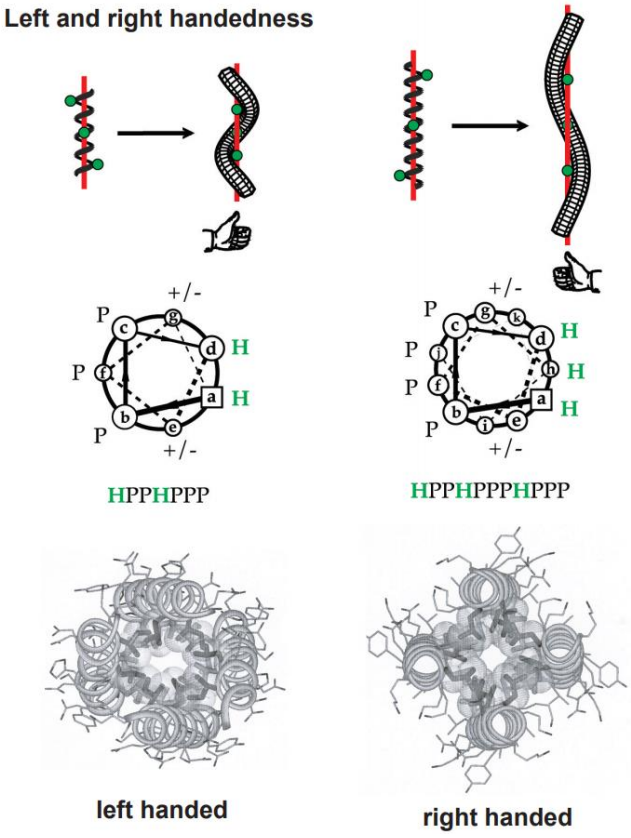
Three- and four-stranded coil:



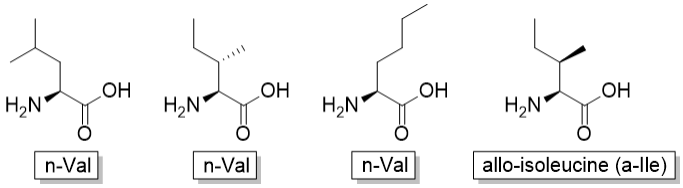
- Right-handed coiled coil (Kim et al.)
  Coiled coils with the 7-AA pattern are left-handed because the seven residues provides a rotation (700°) that is 20° less than the rotation required by two turns (720°) and this difference has to be compensated by left-handed supercoiling. Therefore, to make a right-handed coiled coil, an 11-AA pattern was chosen that leads to a 20° more rotation (1100°) than three turns (1080°).

  $(abcdefghijk)_n$ (a, d, h)

  This pattern has three residues at the contact face, namely a, d and h.



Steric interactions in the interface were modelled by computational methods. The residues used in the simulation were Leu, Ile, n-Val and a-Ile.



Position a	Position d	Position h	Predicted structure
Ile	Leu	n-Val	dimer
a-Ile	a-Ile	a-Ile	trimer
Leu	a-Ile	Ile	tetramer

All three predictions were later proved correct in experiments. In addition, the tetramer was shown to be a right-handed coiled coil by X-ray.

- Toward globular proteins: four-helix bundle

Proteins with four-helix bundles are widespread in nature. From the simple considerations described above, it is possible to get peptides that fold up into structures with four α-helices bound together in a parallel fashion.

To get a globular protein, three linkers need to be introduced to connect the four helices. Experiments have revealed that the exact sequence of the linker is unimportant, as long as the helices are in register and can form the required bundle. (In an experiment testing all 8000 sequences of a 3-AA linker, 92% turned out to be functional)

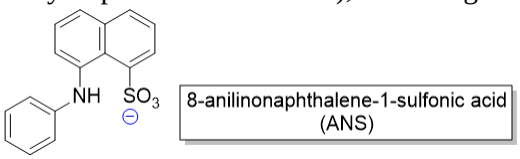
Idealised four-helix bundles can be designed by hierarchical design. This has been done using Leu for hydrophobic residues and Lys and Glu for hydrophilic ones (Hill & DeGrado, 1998). Adjacent helices were linked by a flexible linker PRR (Pro for turn and Arg for solubility).

Name	Sequence	Structure
[α_1B]	G-ELEELLK-KLKELLK-G	tetrameric bundle
[α_2B]	[α_1B]PRR[α_1B]	dimeric 2-helix bundle
[α_4B]	[α_2B]PRR[α_2B]	stable, monomeric 4-helix bundle

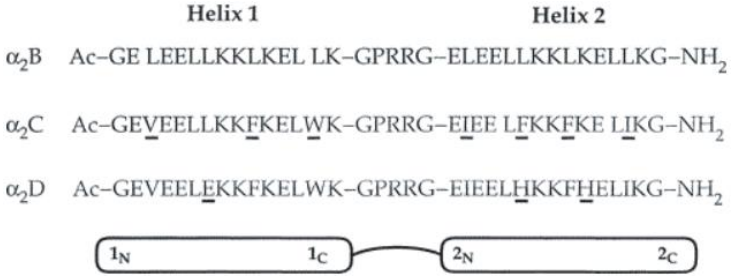
[α_4B] was considered a very successful de novo design of helix bundle. This protein is not native-like and turned out to be a "molten globule" (an ensemble of rapidly interconverting conformers) with dynamic, fluctuating tertiary structure.

Evidences:

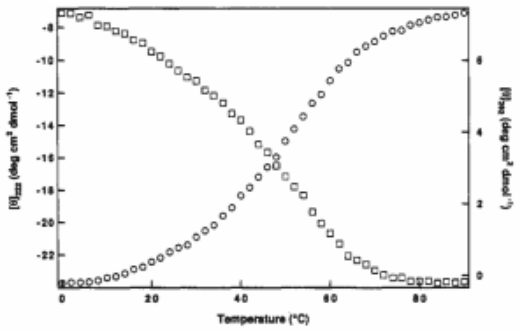
- Poor signal dispersion and broad lines in NMR
- Noncooperative thermal denaturation (i.e. no clear state transition)
- Ability to bind hydrophobic dyes such as ANS (which shows increased fluorescence in hydrophobic environment), indicating that the hydrophobic core is exposed



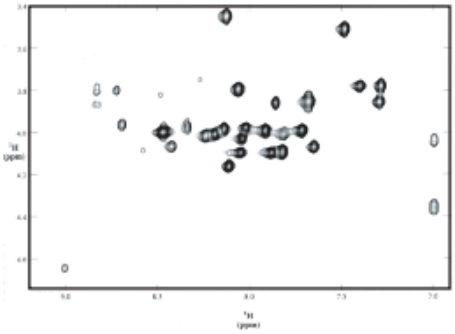
The molten property of [α_4B] is considered a result of poor packing of the buried hydrophobic residues. Leu, the only residue in the hydrophobic core, can adopt a large number of different conformations, and there is no specific interactions that exclude alternative conformations. This problem was solved by designing more specific packing.



The variant [α_2D] was finally shown to be native-like. It binds no more ANS and shows cooperative unfolding and disperse NMR spectra.

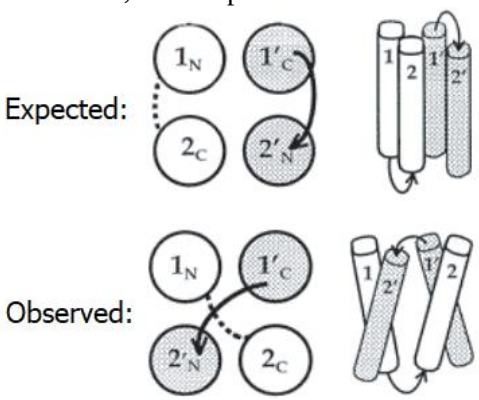


Temperature dependence of the near- and far-UV CD signals



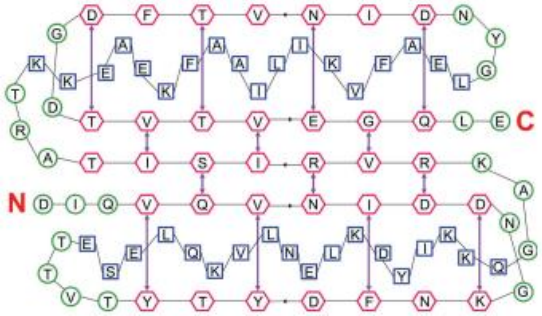
Fingerprint region of a 1H- IH TOCSY spectrum

[α_2D] dimer (four-helix bundle) showed an unexpected structure that was unknown at the time. Later, natural proteins with this kind of structure was discovered.



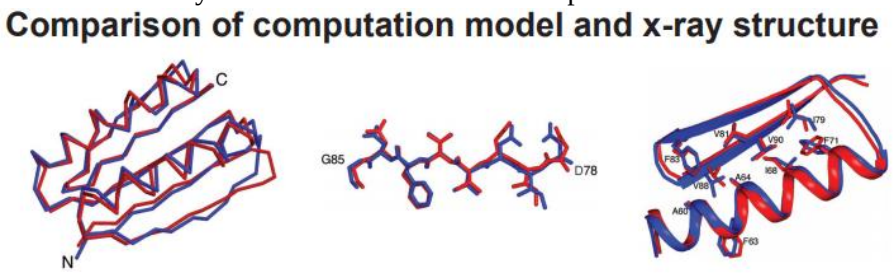
Future

- The efforts to design native-like proteins were successful, but did not work exactly in the expected way. Further progress requires better control to the interactions between residues, so that the expected structure is favoured and alternatives are excluded.
- Computational methods (both hardware and software) are becoming more and more important. One of the most widely used software in this field is Rossetta.
  Example: computational design of a completely unnatural α/β fold
  1) Sketch topology using 3-D structural fragments from protein data bank (PDB)



  2) Define an initial sequence (protein length: 93 AAs)
     Hydrophobic core: 19 AA species (Cys is excluded to avoid redox problems) at 71 positions (~100 rotamers considered per position)
     Polar residues: all standard AAs at the remaining 22 positions (~75 rotamers considered per position)

3) Iterative two-step optimisation (~15 cycles)
   In each cycle, the sequence is first optimised for a fixed backbone. Once the sequence is optimised, the backbone is relaxed and optimised for the optimised sequence.
4) Protein isolation and biochemical characterisation
   The best design, labelled TOP7, is very stable (ΔG°(unfolding) = 13.2 kcal/mol), unfolds cooperatively, and has disperse NMR. X-ray structure of the backbone shows an RMSD of only 1.17 Å between model and experiment.



Implications:

- Unnatural protein folds can exist.
- Models with simple force fields may not be physically realistic but are good enough for designing highly idealised protein structures
- Iterative optimisation is much more successful than ab initio predictions of 3-D structures based only on sequence

- Next goal: progress from designing structures to designing functions

Laboratory evolution 蛋白质进化

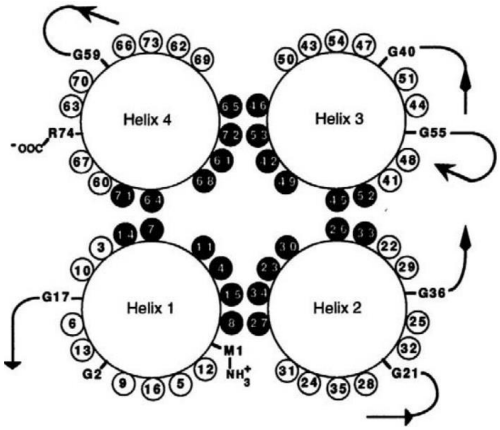
(28.05.2018)

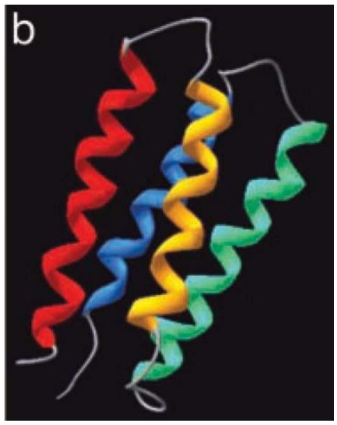
Creation of functional proteins

- Problem: astronomical number of possibilities (20^n variants for an n-AA long protein), meaning that most protein sequences have never been examined by evolution.
  We need efficient methods to navigate the sequence space (i.e. to reduce the size of the sequence space that we have to look into).
- Proteins are combinatorial objects. → We can use combinatorial approaches to make proteins.
- Binary patterning
  This method relies on the premise that the exact residue identity is less important than the overall distribution of residues with distinct properties (usually polar and nonpolar). Thus, we can design patterns that fold into specific secondary structures.
  Sheets: $(\circ\bullet)_n$
  Coiled coil: $(\circ\bullet\bullet\circ\bullet\bullet\bullet)_n$
  …
  With these patterns, we can create a library by maintaining the property (e.g. polarity) and varying the identity of each residue. This library is much smaller than the one generated by stochastic methods and most peptides in it should be able to fold into a reasonable, protein-like tertiary structure.
  To do so, we can take advantage of the inherent degeneracy of the genetic code.

Codon	**AA**	**Polarity**	**Noteworthy**
NTN (N = A, C, G, T)	F, I, L, M, V	apolar	
VAN (V = A, C, G)	D, N, E, Q, H, K	polar	T is avoided in the first base because it may result in a stop codon (and a truncated product)

- Example of binary patterning: design of 4-helix bundles
  Idea:



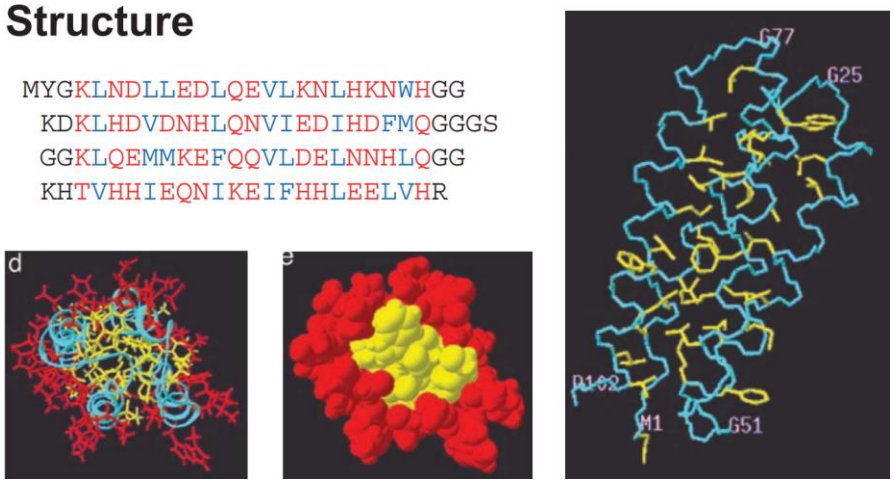


1) Create a binary patterned gene library encoding the four helices and some constant linkers
2) Express the genes in *E. coli*
3) Isolate & characterise all "survivors" of peoteolysis (since unfolded peptides are easily digested by the cellular proteolytic system)

Statistics:

- more than 60% of the sequences survived and most of which are folded as shown by CD spectroscopy.
- Generation 1 are mostly molten globules that do not adopt a single, well-defined tertiary structure. It was thought that the length of the helices (14 AAs) are too short to form an organised structure.

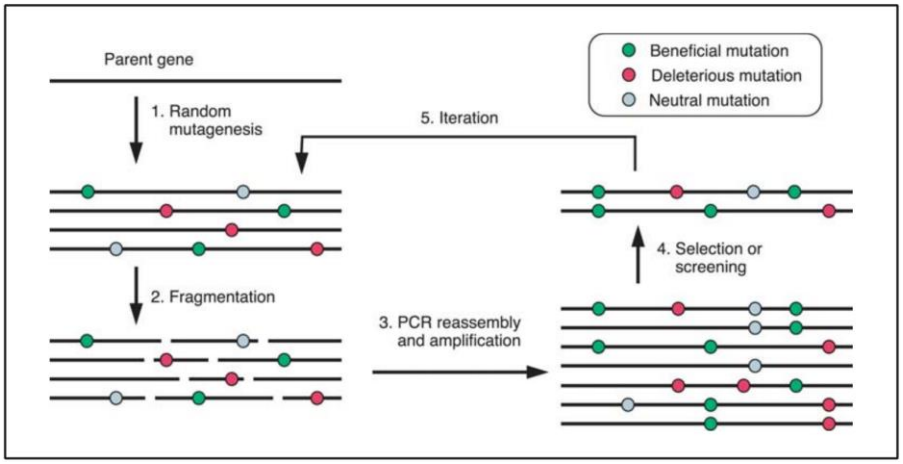
- In generation 2, the length of helices were extend to 20 AAs and native-like, folded structures were observed.



Although functionally naïve, these proteins exhibit both binding & catalytic activities. They work even in vivo and were found to be able to compensate the loss of an essential enzyme (although much less efficient). One explanation is that accumulation of residues with similar properties (polar, nonpolar, etc.) in a small space (such as a pocket) leads to such activities.
(Check out this paper for more details: Ann E. Donnelly, Grant S. Murphy, Katherine M. Digianantonio & Michael H. Hecht, A de novo enzyme catalyzes a life-sustaining reaction in Escherichia coli, *Nature Chemical Biology*, **2018**, 14: 253–255)

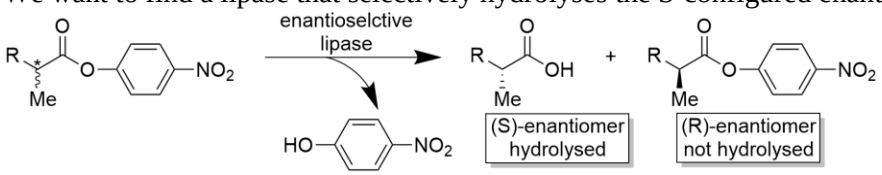
Laboratory evolution

- Darwinian algorithm: apply directed evolution to single molecules
  1) Mutation: create molecular diversity
  2) Screen/select: pick a "winner" out of the large population
  3) Amplification (yield doesn't matter in biology)
  4) Repeat
- Features of directed evolution
  - Easy, robust, reliable
  - Many ways of mutagenesis (smart libraries targeting specific regions are now also available)
  - Typically many iterations needed
- Generation of diversity
  There are many methods available. Some examples:
  - Cassette mutagenesis
    Site-directed mutagenesis. Uses a short, double-stranded oligonucleotide sequence (gene cassette) to replace a fragment of target DNA.
  - Error-prone PCR
    Mutations are distributed evenly.
  - DNA shuffling
    First, random mutations are introduced to a parent gene. The parent gene is cut into small fragments and reassembled. Since the fragments have overlap each other, there is a good chance that favourable mutations are be combined during this process.



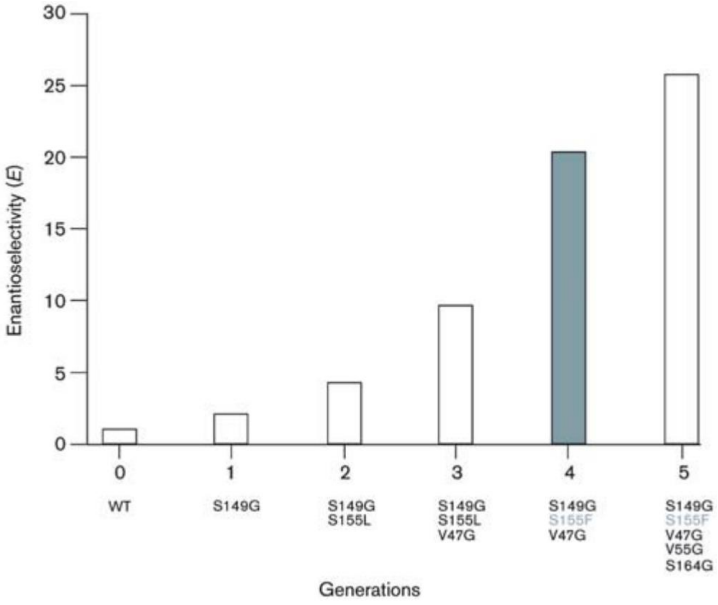
- Identification and isolation of "winners" (critical step)
  Distinguishing between screening and selection:
  - Scope: In screening, we look at every molecule in the library, while in selection we only look at the winners.
  - Limitation: Screening is limited by efficacy of the assay, while selection is limited by size of the library.
  - Suppose you are a bee that wants to find nectar but don't know which flowers are nectar producers. In a screen you have to fly to every flower in the field and check if it produces nectar. In a facilitated screen the nectar producers have some characteristics (such as colour) that allow you to identify them very easily. In a selection, the flowers are grown in such a way that only the nectar producers survive.

  Some common screening/selection methods:
  - Genetic complementation
    Replace a vital function in a cell with a candidate from the library under survey. This method is very powerful but often difficult to set up. Also, there is no guarantee that a candidate compensating the loss of a vital gene has the same function as the product of that gene.
  - Microtiter plate assays
    Experiments done in parallel on microtiter plates (usually 96 wells) and assay them using a microtiter plate reader. This is particularly useful when there is a spectroscopic handle (e.g. increased absorbance, fluorescence) in the reaction.
  - FACS and droplet-based microfluidic sorting
    Cells or picolitre aqueous droplets in oil are sorted according to selected criteria (e.g. fluorescence). This method has a very high throughput.
  - Display technologies
    Display molecules on bacteriophage, cell surface, ribosomes etc.
  - Fuse the gene product to the gene itself
- Example 1: evolution of an enantioselective esterase
  (Easy problem with abundant solutions → simple plate screen)
  We want to find a lipase that selectively hydrolyses the S-configured enantiomer of an ester.



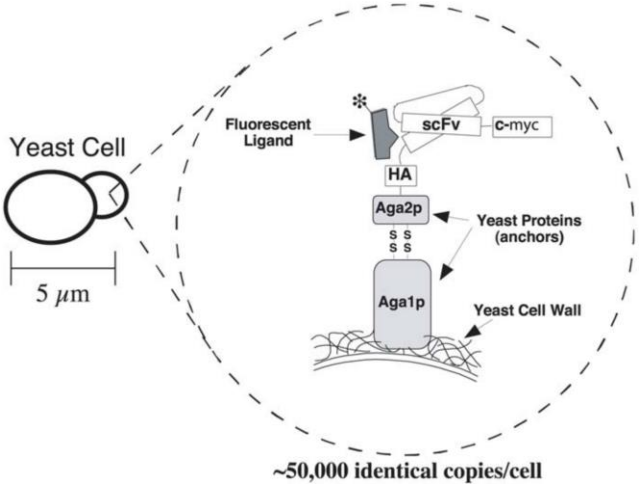
  Starting lipase has the enantioselectivity $E = \frac{k_S}{k_R} = 1.1$

To evolve this lipase, it is first mutated and then screened with the S enantiomer. Since 4-nitrophenol has a yellow colour and the original compound does not, the reaction can be easily monitored by observing the change in colour.
This process is repeated many times, during which the "hotspots" are mutated on purpose by cassette mutagenesis to explore all possibilities. An enantioselectivity of around 51 was finally achieved. Gene shuffling is also performed to accumulate favourable mutations.

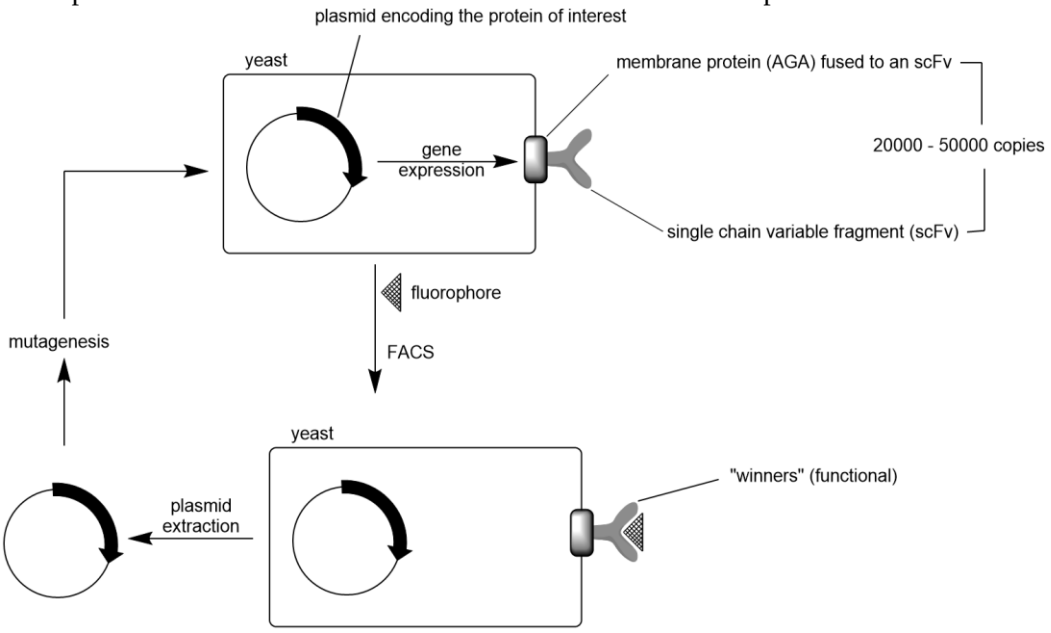


The same procedure can be done to find a lipase that selectively hydrolyses the R enantiomer. In this case, the final enantioselectivity was around 30.
An interesting phenomenon is that the mutations causing such selectivity are often far away from the active site.

- Example 2: affinity maturation of an antibody fragment
(Rare solutions → facilitated screen)
The immune system produces antibodies that recognise the target molecule. However, the affinity of natural antibodies is not always very high. In this case, we can improve the performance of the antibody by evolve it using yeast cell surface display combined with FACS.
In this experiment, a plasmid expressing our protein of interest is introduced into an yeast cell. The proteins of interest is a fusion protein consisting of a membrane protein and a single chain variable fragment (scFv, a fusion protein of the V_L and V_H domains of an antibody). Once expressed, the protein of interest presents its scFv part on the surface of the cell.
If the cell is then exposed to a fluorophore that is specifically recognised by the scFv, the cell will become fluorescent and can therefore be sorted by FACS.



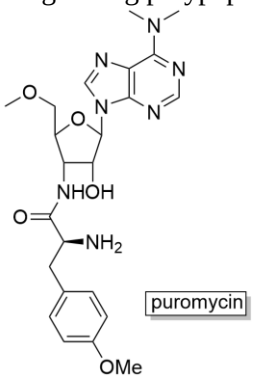
Instead of a single molecule, we can also make a library. In other words, we make a large number of variances in a population of yeast, treat them all with the fluorophore, and pull out the fluorescent "winner" cells (i.e. cells whose variances have higher affinity) with FACS. After that, we generate mutations in the winner plasmids and repeat the whole process until a variance with enough affinity is obtained.
In order to make sure that only the tightest binders are isolated, we can expose the cells to fluorophore for shorter and shorter time or we can introduce competitive molecules.



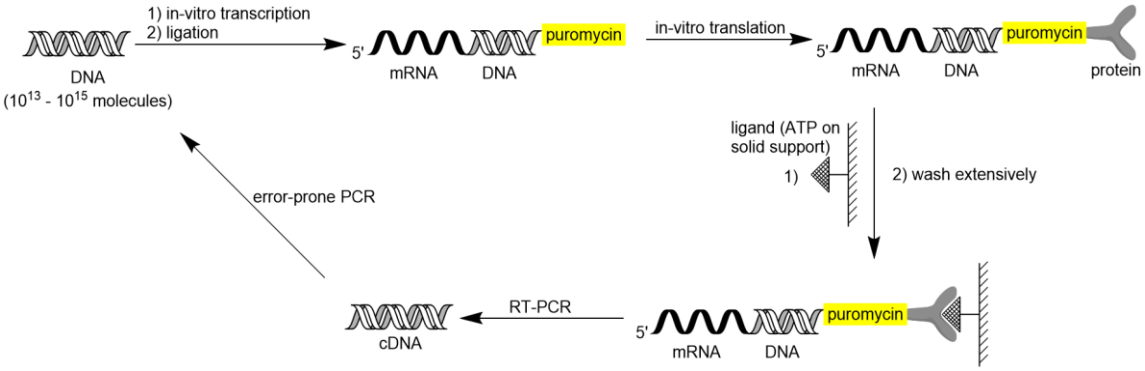
This approach has been used to make a fluorescein-binding antibody.
At the beginning, the dissociation constant was around $K_d = 1$ nM. After multiple rounds of affinity maturation, it became $K_d = 50$ fM.
Again most beneficial mutations are found distant from the active site.
(Check out this paper for more details: Boder et al., *Proc. Natl. Acad. Sci.* USA **2000**, 97, 10701)

- Distributed mutations
  As described, beneficial mutations are mostly not found in the active site. This phenomenon is actually observed in most directed evolution experiments. However, it is difficult to identify these beneficial mutations by inspection and computational methods.
  The beneficial effect of these distant mutations is probably a result of subtle changes in conformation, electrostatics etc. that are transmitted to the binding pocket. It is believed that mutations at all positions can have a beneficial effect on the active site, and that distant mutations are more common simply because the chance for a mutation to hit a distant site is much higher than to hit the active site (which normally comprises <10% of all AAs in the protein).
  In terms of thermostability, active site mutations and distant mutations are equally beneficial because thermostability is a global property and most mutations are additive (in the sense of $\Delta G°$). As for binding/catalytic activity, active site mutations typically have greater effect (either positive or negative) than distant mutations, because binding/catalytic activity is a local property and residues in the active site tend to interact in a synergistic fashion (i.e. the overall effect is much larger than the sum of individual parts).
- Example 3: discovery of ATP receptor from a random sequence space
  (Very rare solutions → selection)
  Selection is achieved by mRNA-protein fusions.

The first step is to create a library of random polypeptides encoded in a synthetic oligonucleotide. mRNA is produced with in-vitro transcription and linked to a short piece of DNA, which in turn is attached to puromycin.
Puromycin is an antibiotic that has an amine at 3' position, to which a tyrosine derivative is attached. Puromycin can bind to the A-site of a ribosome translating an mRNA and capture the growing polypeptide chain, causing it to stuck.



The mRNA-DNA-puromycin hybrid is translated in vitro. When the ribosome reaches the end of the mRNA, puromycin will reach into the active site, capture the polypeptide chain and link itself to it, resulting in a mRNA-DNA-protein hybrid molecule.
Proteins that bind ATP are pulled out by fixing the ligand (ATP) to a solid support, running the protein mixture generated in the previous step over it, and washing it extensively.
The mRNAs encoding high-affinity proteins are then reversely transcribed into cDNA and amplified by error-prone PCR to give a new DNA library serving as the starting point for the next round.



Following is the result of this experiment.
Start: randomised gene encoding 80-AA long proteins
Ligand: ATP (attached to solid support)
Number of selection rounds: 18
Statistics: Molecules that can bind ATP are very rare ($1/10^{11}$). Those that bind to ATP can be divided into four families, the best of which is a Zn^{2+}-dependent ATP binder. It is highly selective ($K_d(\text{ATP}) = 100$ nM, $K_d(\text{ADP}) = 200$ nM, $K_d(\text{AMP}) = 900$ nM, $K_d(\text{2dATP}) = 4$ µM, $K_d(\text{A}) = 60$ µM, $K_d(\text{GTP/CTP/TTP}) > 200$ µM). X-ray revealed a unique structure that is unknown in nature. Although the molecule was crystallised with ATP, ADP was detected in the crystallised compound, suggesting that this molecule might also be able to catalyse the hydrolysis of ATP.
Advantages of mRNA-protein hybrids:

- Huge libraries
- Compatible with non-canonical AAs
- Useful under many selection conditions

Summary of the course

- Proteins and lipids are macromolecules essential for life.
- Proteins and lipids are (re)designable for new applications.
- Chemical and biological approaches have their own advantages and disadvantages, depending on the problem addressed. Combined, they are incredibly powerful.