

# BCH 304 - Protein Biophysics

## Summary

Let's learn some CBB

02/02/2020

## Part I - Caflisch

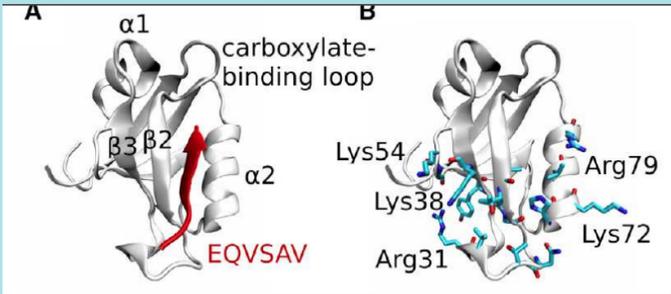
### 1 Interactions between atoms

The Pauli principle states that all atoms attract each other when they are only a small distance apart but repel upon being squeezed.

#### 1.1 Classification of interactions

Interactions depend mainly on the charge state and on the electronic structure (open vs. closed).

**Example**



What we see in this picture is in red the carboxy terminus of the substrate. There are negative charges on the COO<sup>-</sup> group and on the glutamic acid (E). The substrate binding site on the other hand is positively charged by all the Lys residues. This shows nicely the importance of interactions. When the whole protein is shifted in a surface charge plot we see, that the side opposite of the binding site is highly negatively charged. This is important in repelling the substrate if it is on the wrong side.

Principles of atomic interactions are not only important in substrate binding but in any form of ligand binding, as for instance small-molecule inhibitor binding.

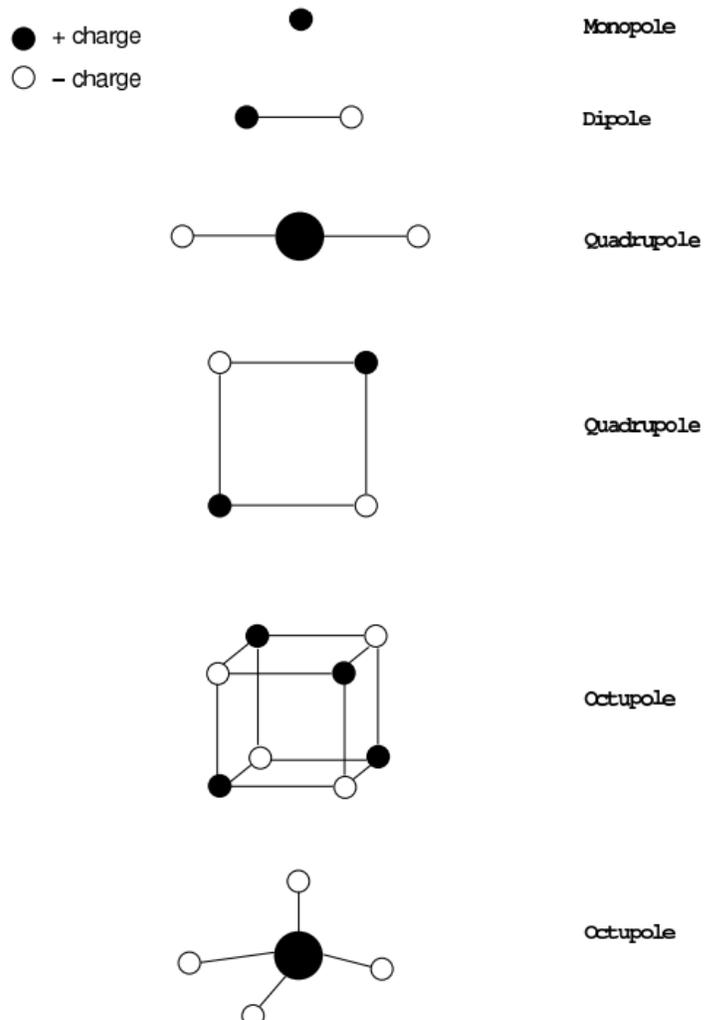
#### 1.2 Electrostatic interactions in molecules

The coulombic energy  $E$  in medium with the constant  $\epsilon$  is:

$$E = 332 \frac{q_i q_j}{\epsilon r} \quad (1)$$

The dipole between the charged particles is a vector  $\vec{\mu}$  that points away from the negative charge to the positive charge. We distinguish permanent electric dipole moments (water) and induced dipoles (methane in presence of ammonium). The field vectors of an electrostatic field are orthogonal to the equipotential surfaces (where the potentials remain the same).

**Multipole expansion:** Is basically a Taylor series of our function. The multipole expansion provides a description of the electrostatic potential. If the charges are localised close to their origins the coefficients in the expansion are called multipole moments.



In an electrostatic field the dominant molecular multipole is the dipole. The dipoles will align against an external field. This is called the screening effect which is summarised in the dielectric constant  $\epsilon$ .

Type of Interaction	Distance-dependence	Typical energy [kcal/mol]
Monopole-Monopole	$1/r$	-50 to -4
Monopole-Dipole	$1/r^2$	-3.5
Dipole-Dipole	$1/r^3$	-0.5
Dispersion	$1/r^6$	-0.1

The distance-dependence is proportional to the Coulomb energy for  $2^n$  poles:

$$E \propto \frac{1}{r^{n+m+1}} \quad (2)$$

### 1.3 Dipole-Dipole interactions

The dipole-dipole interactions lead to the following energy dependence if the distance  $r \gg l$ :

$$E = -664 \frac{\mu_i \mu_j}{\epsilon r^3} \quad (3)$$

This is the sum of the four contributions (consult script). If the two dipoles are directed into the same direction it is the most favorable (-1.32 kcal/mol), if they face against each other, the least (+1.32 kcal/mol). upwards in opposite direction (-0.66 kcal/mol) and in the same direction (+0.66 kcal/mol).

### 1.4 Dipole-induced dipole interactions

Dipoles can as well be induced like in the case of Benzene. There we include the polarisability  $\alpha$ .

$$E \propto -\frac{\mu_i^2 \alpha_j}{r^6} \quad (4)$$

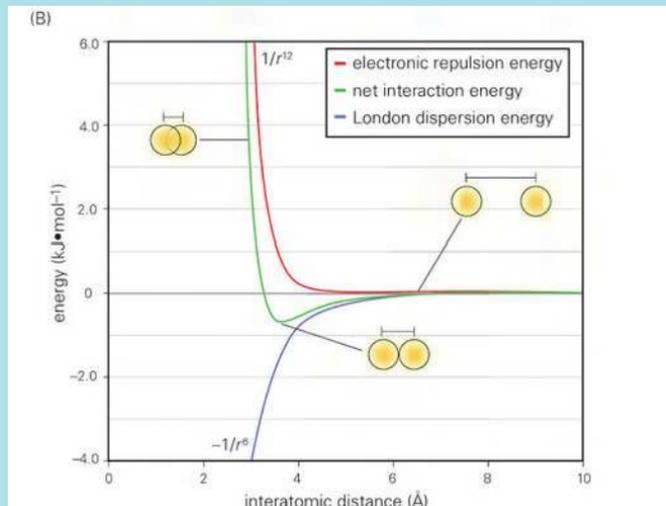
in the equation we have the inducing species  $i$  and the polarised species  $j$ .

### 1.5 Van der Waals interactions

All atoms show dispersion interaction (van der waals) which is proportional to  $r^{-6}$ . The attractive dispersion is balanced by the electron repulsion (Pauli) which dominates over short distances. The van der Waals energy is then the sum of repulsive and attractive interactions:

$$E_{vdW} = E_{min} \left[ \left( \frac{r_{min}}{r} \right)^{12} - 2 \left( \frac{r_{min}}{r} \right)^6 \right] \quad (5)$$

### Example



The following picture summarises the repulsion, the dispersion and the net interaction energy. This net energy shows a clear minimum with rather low energy.

This phenomenon can as well be found in a pair of carbon atoms and oxygen atoms. There we find that the Carbon interaction need more space in Å. Energetically, the carbon-carbon interaction is more favorable than the oxygen-oxygen interaction.

### 1.6 Hydrogen bond

Hydrogen bonds are a consequence of a positive partial charge on the H and an electron withdrawing donor D and the electrons of the acceptor A. The distances in H bonds are shorter than in vdW and depend on the angles of D, H and A. The optimal case for D-H ···A is 180°. Between H ···A-AA (anterior acceptor) is 135°, with a carbonyl carbon it should be zero (in the plane).

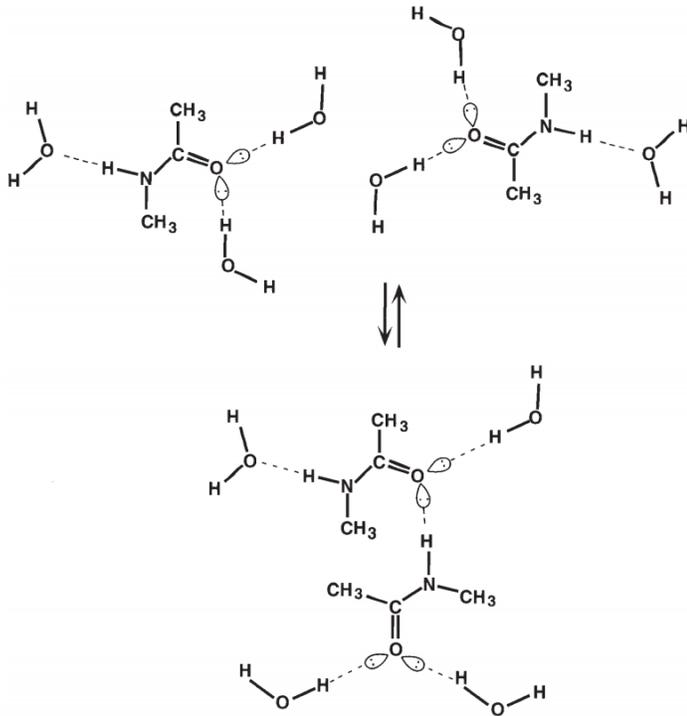
#### Typical distances between atom pairs:

Covalent bond between two carbon atoms: 1.5 Å.

Hydrogen bond between amide nitrogen and carbonyl oxygen 3.0 Å.

vdW between two carbon atoms 4.5 Å.

#### Example



The protein can either form H bonds with water or with itself, like in  $\alpha$  helices. The formation of H-bonds with water occurs only if the water molecules are closely oriented. Dimerisation is only possible if the loss of H-bonds with free water is favourable.

## 1.7 Properties of water

Each water molecule can be an acceptor for two hydrogen bonds and a donor for two hydrogen bonds. This means in the liquid phase and optimal conditions water will form four hydrogen bonds. This peculiarity is responsible for the high dielectric constant and the high melting and boiling points of water. The high heat capacity indicates high degree of organisation. The more hydrogen can be broken, the higher the heat capacity. This decay is constant in the sense of having a constant gradient.

## 1.8 Hydration

The gibbs free energy of hydration of an ion can be approximated:

$$\Delta G_{hydr} = -\frac{332Q^2}{2R_{ion}} \left( \frac{1}{\epsilon_{vacuum}} - \frac{1}{\epsilon_{water}} \right) \quad (6)$$

This describes the change in free energy upon moving an ion from vacuum to water. In the case of proteins charged amino acids are on the outside to interact with water. The formula can be derived as follows

$$\phi = \frac{332q}{\epsilon R_{ion}} \quad (7)$$

$$w = \int_0^Q \phi dq = \frac{332}{\epsilon R_{ion}} \int_0^Q q dq = \frac{332}{\epsilon R_{ion}} \frac{Q^2}{2} \quad (8)$$

In protein hydration water molecules involve in hydrogen bond formation with hydrophilic groups. There are usually only few buried polar groups which are involved in intra-protein hydrogen bonds.

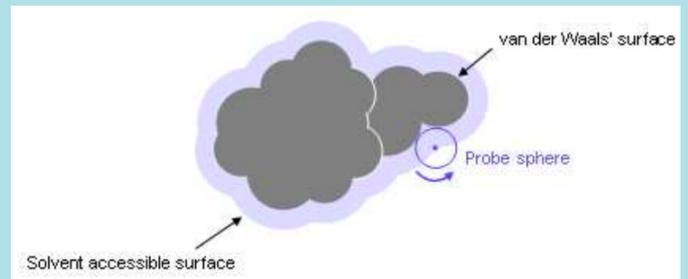
## 1.9 Hydrophobic effect

The main reason for the hydrophobic effect is entropy. In a hydrophobic solute water molecules cannot form hydrogen bonds with the solute and therefore there is an entropy loss due to the water molecules being highly structured to maximise the number of hydrogen bonds among them. The structure of the water molecules is cage-like and called clathrate.

At physiological temperatures the hydrophobic effect is the main cause of the folding of proteins. The hydrophobic side chains are preferred to be buried within the hydrophobic core.

Protein-protein association and peptide-protein binding are as well often driven by the hydrophobic effect.

### Example



Here we see how the solvent accessible surface was calculated. The vdW radius was determined from the protein. Then a probe with the radius of water "rolled" along the vdW radius and like this the solvent accessible surface was calculated. the relationship is negatively linear, meaning that with increasing size the solvation process is entropically disfavored.

The hydrophobicity of amino acids is anti-correlated with the free energy of the water-vapor or water-cyclohexane (cyclohexane  $\epsilon$  is similar to the interior of a protein) transfer. Subtracted from these values is the hydrophobicity of Gly, so that the hydrophobicity is mainly due to the side-chains and not to the backbone.

## 1.10 Hydrophobic Effect in Micelles and Membranes

The shape of micelles and bilayers depends on the conical (monolipid) or cylindrical (bilipid) shape of the amphiphilic molecule.

## 1.11 Hydrophobic Effect and Membrane Proteins

The transmembrane region of proteins is hydrophobic on the outside. One performs an analysis on the hydrophobicity of the amino acids. In order to do this, one averages over the scores of a pre-defined sliding window. Transmembrane domains either form  $\alpha$  helices or they form  $\beta$  barrels (only in OM of bacteria and mitochondria).

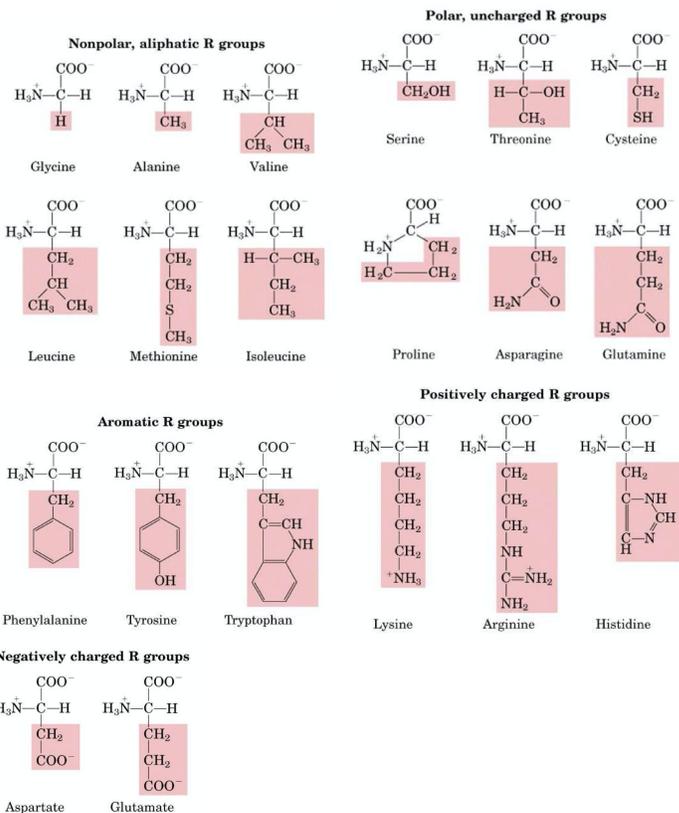
## 2 Protein Structure

Amino acids are the basic structural units of proteins, they consist of an amino group, a carboxyl group, a hydrogen atom and a residue termed R. The conformation is tetrahedral around the  $C_{\alpha}$ -atom, this results in an optical activity. (L- and D-isomers) In nature only L-amino acids are found.

### 2.1 Properties of Amino Acids

Twenty different amino acids are found in proteins, they differ in size, shape, charge, hydrogen bonding ability, hydrophilicity and chemical reactivity.

- **Gly:** Flexible conformation
- **Ala, Val, Leu:** Aliphatic (non-aromatic hydrocarbons), hydrophobic
- **Pro:** Aliphatic, ring structure, side-chain bound to  $C_{\alpha}$  and N atoms, amino acid with secondary amino group, found in bends of folded chains
- **Phe, Tyr, Trp:** Aromatic, the rings have clouds of delocalized  $\pi$ -electrons that allow them to interact with other  $\pi$ -systems ( $\pi$ -stacking) and transfer electrons
- **Cys, Met:** Have a sulfur atom, reactive  $-SH$  group, form disulfide bridges important for stability and shape of tertiary structure
- **Ser, Thr, Asn, Gln:** polar, charged, residues often involved in hydrogen bonds (through side chain hydroxy groups (Ser, Thr)) and amide groups
- **Asp, Glu:** Acidic, side chain carboxyl group is usually negatively charged under physiological pH, if these residues are located within the protein they are most likely involved in salt bridges with Arg or Lys
- **Arg, Lys:** Basic, the guanidinium (Arg) and amino (Lys) groups are usually positively charged under physiological pH, if these residues are located within the protein they are most likely involved in salt bridges with Arg or Lys
- **His:** Basic, aromatic, the pK value of the imidazole ring lies in the physiological pH range, present in the active center of serine protease (imidazole ring switches between ionization forms)



### 2.2 Post-Translational Modification

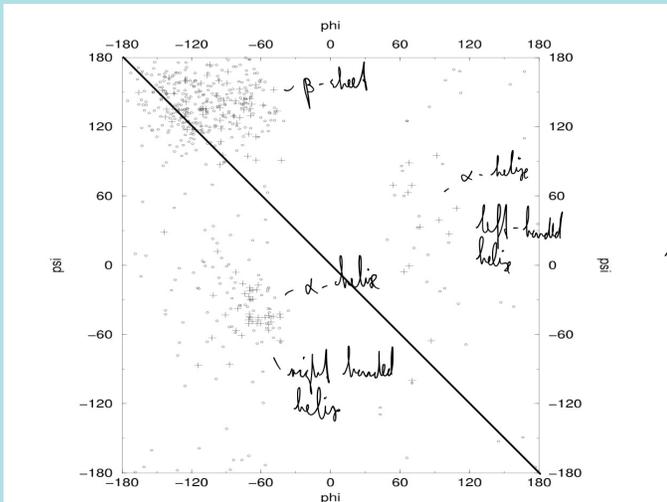
Some PTMs change the charge states of the side chains. Examples are phosphorylation and acetylation. Phosphorylation of the hydroxyl group of Ser, Thr and Tyr results in an additional negative charge. Acetylation of the amino group on Lys residues on the other hand results in a neutral amide. This is important in the context of histone-DNA packing. Addition of acetyls results in a less densely packed DNA by interaction with bromodomains. There are as well PTMs which don't affect the charge state like e.g. methylation. Methylation can happen several times at a residue (mono-, di- and trimethylation). This PTM works by changing the size of the residue via steric hindering.

### 2.3 Dihedral angles and Ramachandran plot

The dihedral angles are the angles that are created when looking at static chemical bonds as planes. The angle between two planes is a dihedral angle. For each residue three backbone dihedral angles are considered,  $\phi$ ,  $\psi$  and  $\omega$ . Due to the partial double bond character peptide groups are planar and the  $\omega$  angle is in *trans*. In Pro the probability of finding  $\omega$  in *cis* is 10%.

The Ramachandran plot is a two-dimensional representation for the conformation of individual dipeptide units in the  $\phi - \psi$ -plane.

## Example



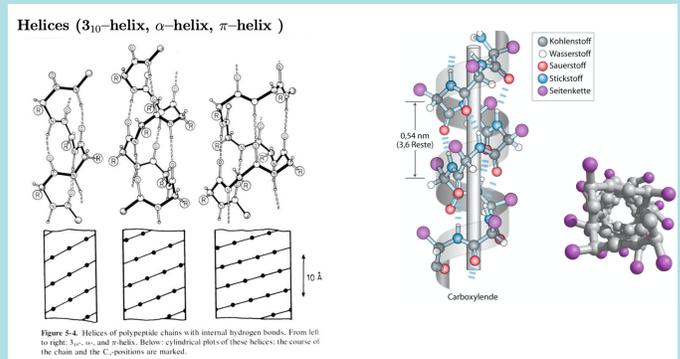
In this picture the Ramachandran plot is visible. The  $\beta$  sheet region is a region with higher density on the right side as right the individual strands have a right-handed chirality. The  $\alpha$  helical region has a higher density on the left side due to the right-handed chirality of the helix. Glycine is the only amino acid which will show a symmetrical ramachandran plot since it is the only achiral amino acid. Amino acids like proline have only a small  $\phi$  region because of steric hindering via their ring-structure

## 2.4 Secondary Structure

The secondary structure is governed by the backbone angles  $\psi$  and  $\phi$  and can be classified into regular ( $\alpha$  helices and  $\beta$  sheets) and irregular (loops) elements.

In regular elements the dipeptide units show the same combination of  $\phi - \psi$  angles. In helices the R-groups point outwards. There are three different helices we distinguish between.  $3_{10}$ -helix,  $\alpha$ -helix and  $\pi$ -helix. The optimal stacking of turns can be found in  $\alpha$ -helices with an optimal vdW distance of 2.3 Å per helix turn. In  $3_{10}$ -helix this radius is too compact with 1.9 Å and too wide in  $\pi$ -helices with 2.8 Å.

## Example



The picture shows the different types of helices and their stacking. The optimal case is the  $\alpha$ -helix with 2.3 Å.

$\beta$ -sheets can be either parallel or antiparallel. The structure of antiparallel  $\beta$ -sheets is stretched as the  $\phi$  angles are stretched. They have a lot of H-bonds which are nicely ordered. The R's are always alternating in their orientation. In parallel  $\beta$ -sheets the orientation of R's is not perfectly aligned, therefore the H-bonds are different as well, meaning that the stacking is a little bit different.  $\beta$ -sheets can aggregate very closely meaning they can form very densely packed structures like amyloid fibrils, which  $\alpha$ -helices can't do.

Another aspect is the propensities of building  $\alpha$ -helices and  $\beta$ -sheets by different amino acids. There are some amino acids like glycine or proline that are poor helix builders. Alanine on the other hand is very favorable for building  $\alpha$ -helices. The same analysis can be performed for valine or isoleucine which are really good  $\beta$ -sheet builders. This is mainly due to the branching at the  $C_{\beta}$  of V and I. Glycine is as well here very unfavorable, since it disfavors the entropic component in the secondary structure formation.

Loops and turns are irregular elements of secondary structure. Loops have usually a length of 5 to 30 residues with a majority at 10. Tight turns are found in antiparallel  $\beta$ -sheets where they connect neighbouring  $\beta$ -strands.

Supersecondary structure describes the combinations of secondary structures that show a high degree of structure but do not form entire structural domains yet. Examples are hairpin motifs and cross-over folds.

## 2.5 Tertiary Structure

Arrangement of regular secondary structure and loops into folded three-dimensional structure. The side chains are important here as they can e.g. determine a hydrophobic core for folding.

Taxonomically, proteins can be ordered according to their  $\alpha$ -helical /  $\beta$ -sheet content. ( $\alpha$ )-proteins contain mainly  $\alpha$ -helices, 50% to 95%. This value can never reach 100% as there is always the need for loops to connect the helices. ( $\beta$ )-proteins contain mainly  $\beta$ -sheets. Two of these (mainly antiparallel) are packed on top of each other. ( $\alpha + \beta$ )-proteins contain helices and sheets in separate parts of the sequence

whereas ( $\alpha/\beta$ )-proteins have alternating helices and sheets.

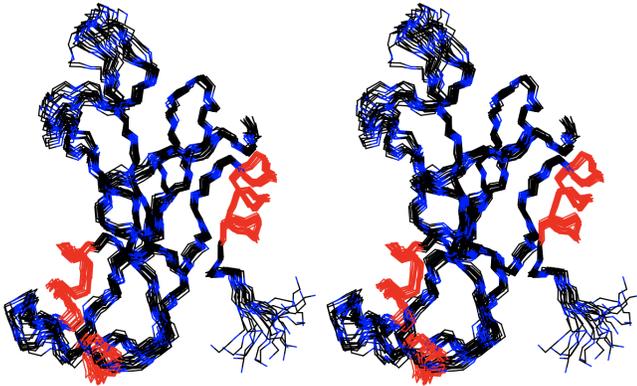
## 2.6 Quarternary Structure

Proteins that consist of more than one polypeptide chain (subunits) adopt a new level of organisation called the quarternary structure. The subunit contactpoints are often of physiological importance, as in the case of hemoglobin's coordination of  $O_2$  and  $CO_2$ . If the subunits are identical, we speak of a homodimer, otherwise of a heterodimer.

## 2.7 Experimental Approaches

There are high resolution methods, single molecule methods and other methods for determining experimentally the biophysical behaviour of a protein.

### Example



This is an NMR-derived structure. We see 20 NMR conformers and the differently flexible regions. The N-terminus is very flexible, leading to a wide range of possible conformations.

## 3 Ordered Aggregation and Amyloid Fibrils

The most important protein when studying amyloid fibrils and Alzheimer's disease is the alzheimer polypeptide precursor (APP). This is a transmembrane protein where the  $A\beta$  section crosses the membrane. This  $A\beta$  protein can be cleaved at multiple sites, by BACE at the N-terminus and by  $\gamma$ -secretase at the C-terminus. The differences in cleavage and the resulting different lengths of APP show differences in pathogenicity. One approach to modulate or inhibit thus the pathogenic function of APP is to block the malignant cleavage by BACE and  $\gamma$ -secretase.

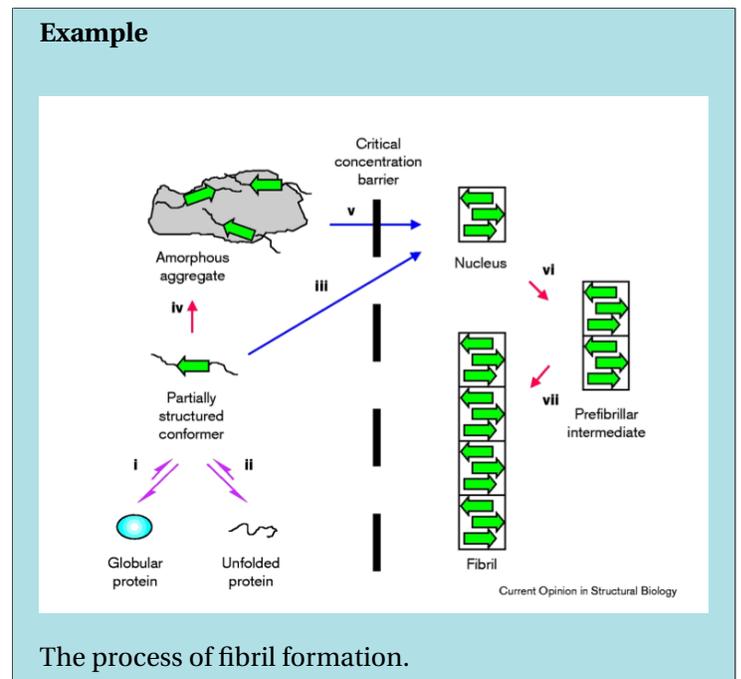
### 3.1 Amyloid Fibrils

Amyloid fibrils are organised in cross- $\beta$  structure. This means that they consist of  $\beta$ -sheet structures with the strands being perpendicular to the fibrillar and the backbone hydrogen

bonds parallel to the axis. Stabilising are the hydrogen bonds, vdW and the hydrophobic effect. The distance for the hydrogen is optimal - perfect packing. The variation between the sheets is due to the different side chains. An individual unit of the fibril is a homodimer with a double horseshoe topology. They contain a floppy N-terminal domain as seen in NMR spectroscopy. There are certain mutations that favor the formation of amyloid fibrils like K16N which takes away a positive charge by mutating to a neutral Asn. This mutation is common in younger patients. The broad structure of amyloids is always the same, the difference lies in the tips as one might grow faster than the other.

### 3.2 Kinetics of Fibril Formation

Some conditions favor the formation of insoluble fibrils like high concentrations and/or temperature. Fibril formation requires partial unfolding of globular proteins. The critical step is the nucleation, so the formation of a primary fibrillar structure.



The process of fibril formation.

The kinetics of fibril formation can be separated into a lag phase, a growth phase and a saturation phase - see page 66 for more info. There is an inverse correlation between the lag phase and the slope of the curve. The formation is strongly sequence dependent. Conservative mutations have no effect. Amino acids that differ only slightly can already influence the rate formation drastically and mutations with Pro mutations can't aggregate at all. The dock and lock mechanism describes the growing of two steps (docking) where the molecules associate with the tip followed by a second step (locking) where the intermolecular  $\beta$ -sheet is established.

### 3.3 A unique Fibrillar Structure?

Some amyloids like  $\beta$ -solenoid fold into a unique fibrillar structure whereas other amyloids (including disease-associated amyloids) can assume different structures. Fib-

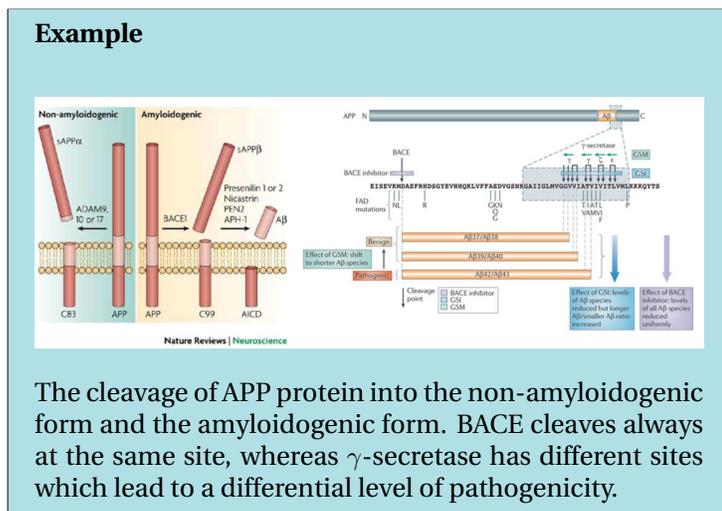
rillar structures are influenced by environmental factors whereas globular proteins have evolved to be quite stable in a range of conditions. In contrast to protein folding, amyloid formation is under kinetic control, not thermodynamic control. Molecular recycling describes the process of dynamically constructing/deconstructing the amyloid fibril in question until the whole thing is recycled eventually, which can happen because the process is asymmetric.

### 3.4 Amyloid Related Diseases

Several neurodegenerative diseases have been associated with either extracellular amyloid deposits or intracellular amyloid-like inclusions. Consult page 73 for more information.

Prions are infectious agents composed of proteins in a misfolded state. The folded structure of a prion can convert to an amyloid prone structure that is able to act as a template to guide the misfolding of more protein into prion form and to subsequently build amyloid fibrils.

**Alzheimer's disease** is the most common form of dementia. The accumulation of Amyloid- $\beta$ -protein is associated with the outbreak of the disease. The early oligomers are expected to be the toxic agents. The A $\beta$  peptide originates from an abnormal cleavage of the amyloid precursor protein (APP) by  $\beta$ -secretase (BACE) and  $\gamma$ -secretase. Hydrophobicity of the APP peptide is associated with the velocity of polymerisation of the amyloid fibrils and makes A $\beta$  as well more toxic *in vivo*. The N-terminal segment is mainly hydrophilic. Most of the cleavages are not pathological as the pathological part of APP remains within the membrane. If however BACE and  $\gamma$ -secretase cut both, we get the pathological A $\beta$  peptide in solution.



There are several different **therapeutic strategies** like the inhibition of BACE and  $\gamma$ -secretase, small molecule modulators of A $\beta$  aggregation, specific targeting of soluble oligomers and stabilisation of less toxic fibril products by antibodies.

### 3.5 Functional Amyloid Fibrils

There are as well amyloidic structures that have a function and are by this the functional state of some specific proteins. In the marine snail a prion CPEB is even associated in memory function. Amyloid formation can thus have a physiological function, provided it is regulated and allowed to take place under highly controlled conditions.

Most amyloid fibrils are made up of  $\beta$ -sheets but there are exceptions like in sychel cell anaemia which is due to  $\alpha$ -aggregation. Another non  $\beta$ -sheet fibril is the actin filament made up of folded proteins.

## 4 Protein Folding

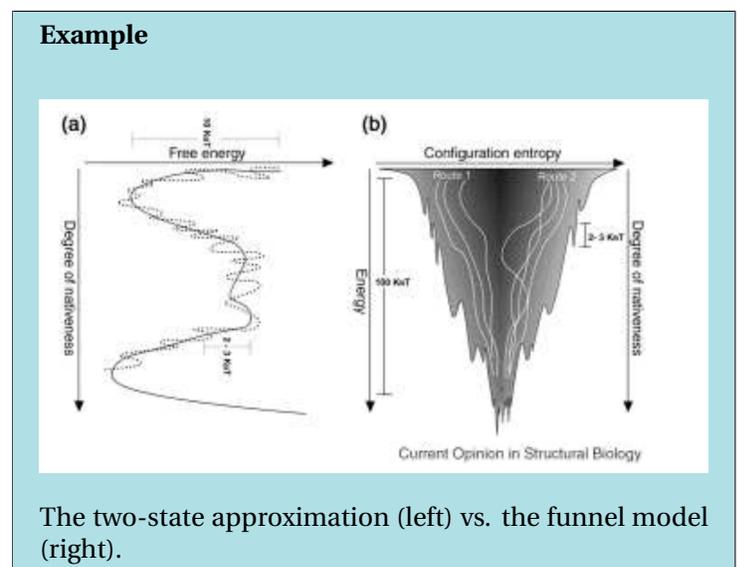
In protein folding we must distinguish thermodynamic and kinetic control.

**Thermodynamic control** is just the minimum of free energy. This stands in contrast to the **kinetic control** which takes into account the activation barrier of the reaction. A protein can be thermodynamically more stable but kinetically less stable. The folding/unfolding process is important for the action of proteins, the transfer across membranes and protein degradation. There are as well many diseases that are related to protein misfolding due to inherent properties of WT proteins, changes in the environment and mutations.

Small proteins up to 150 Aas do not require chaperones for their folding.

### 4.1 The funnel model of folding

The funnel of folding states that the process of folding is not a simple pathway but rather process over a free energy landscape with many intermediates that can lead to local optima. This is contrasted to the "golf course" model that looks at folding as a "hole in one" process.



## 4.2 Protein folding mechanisms

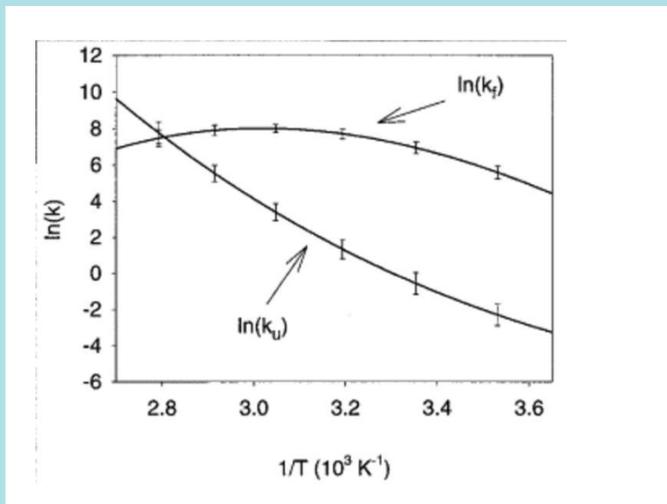
Proteins fold via several different pathways whereas variations of the nucleation-condensation mechanisms describe the overall features of folding of most domains. There are the two extreme cases, the framework if the helical propensity is very high (unlikely to happen) and the hydrophobic collapse, where the propensity is very low (as well an extreme case, as there would be no order at all). Secondary structures are not stable and are stabilised by tertiary interaction due to the low helical propensities for  $\alpha$ -helices.

In most of protein folding *nucleation* is the first step which leads to an initial nucleus with a few native contacts. What happens afterwards is the process of *condensation* which leads to an extension of the native contacts and to a fully folded molecule.

## 4.3 Folding Kinetics

The Arrhenius behaviour describes the reaction rate of a monomolecular reaction as an exponential decay. The unfolding rate ( $k_u$ ) shows Arrhenius behaviour whereas the folding rate ( $k_f$ ) only shows Arrhenius behaviour at low temperatures and even anti-Arrhenius behaviour at high temperatures.

### Example



The temperature dependence of  $k_u$  and  $k_f$ . Note the reciprocal value of the temperature T along the x-axis.

The contact order is the separation along the primary structure of pairs of residues that will be in proximity in the tertiary structure. There is an anticorrelation between the contact order and the rate of folding, meaning that mainly  $\alpha$ -helical proteins fold faster than  $\beta$ -sheets, there in particular parallel  $\beta$ -sheets.

*Note : when a reaction has a rate constant that obeys Arrhenius equation, a plot of  $\ln(k)$  versus  $1/T$  gives a straight line with negative slope. That line can be used to evaluate the activation energy  $E_a$  and the pre-exponential factor  $A$*

## 4.4 Two-state folding

In small proteins the two-state model of protein folding seems accurate but the larger proteins the two-state model is only an approximation and there are several intermediates that are more or less populated. The transition states between U/N and the intermediates is an ensemble of relatively similar structures.

## 4.5 Complexity of the unfolded state

The most obvious reason against the two-state model is that the unfolded state is not clearly defined but rather a complex state populated by many different possibilities. The folded state is as well heterogeneous but far less than the unfolded state. The ensemble is however not completely random but there are some native and non-native interactions, transient hydrophobic clusters and residual but unstable regular secondary structure and turns.

$$\Delta G_{folding} = \Delta H_P + \Delta H_{H_2O} + \Delta H_{P-H_2O} - T\Delta S_P - T\Delta S_{H_2O} - T\Delta S_{H_2O-P} \quad (9)$$

This is the Gibb's free energy of folding where the most important component to make the process spontaneous is the hydrophobic effect ( $-T\Delta S_{H_2O}$ )

## 4.6 Protein denaturation in vitro

Due to the low conformational stability of a folded protein (-15 to -5 kcal/mol) denaturing proteins via a variety of techniques is possible by altering the weak non bonding energy contributions.

- Heating: altering entropy/enthalpy balance
- pH variations: altering the ionisation state of amino acid R's
- Detergents: associate with nonpolar residues
- Guanidinium ion or urea: chaotropic agents that bind to both polar and apolar protein groups.

## 5 Molecular Dynamics

Very fast processes in biophysics can only be simulated via molecular dynamics because the time resolution of the imaging techniques is too low. The lowest molecular dynamics can model is  $\approx 1fs$ . After molecular dynamics, NMR has the highest time resolution.

Computer simulations have two major problems:

- huge size of configuration space leads to the fact that we can't sample all conformations - **statistical error**.
- accuracy of the molecular model and the force field - **systematic error** (protein simulations are simplifications).

# Part II - Vitalis

## 6 The Basics

### 6.1 X-Ray Crystallography

**Collected Data** In X-Ray Crystallography the collected data is a 2D pattern of intensities of scattered beams at multiple incidence angles.

**Strengths** After solving the phase problem (required for Fourier synthesis), one can acquire 3D electron densities. X-Ray crystallography allows for a very high atomic spatial resolution and was used for most of the known structures up until now.

**Disorder Tolerance** is not very high in X-Ray as it is quasi inapplicable to disordered systems mainly due to the process of crystal formation. Partial disorder can be resolved by either resulting in crystallographic disorder or in artificial order. If the protein crystal has a lot of water, disorder is favored as the actual signal is averaged out and we get thereby no or only weak signal.

### 6.2 Transmission EM

**Collected Data** of transmission EM are 2D electron micrographs. These are acquired from very thin grids of flash-frozen samples.

**Strengths** are e.g. the determination of a 3D electron density after solving the reconstruction problem. With EM the spatial resolution of acquisition is very good and high and the states obtained are realistic as the freezing process is very rapid and there is little distortion.

**Disorder tolerance** is rather high in EM but the resulting images will contain close to no high-resolution information, as the disordered part tends to fall below the noise level in the averaging step.

### 6.3 Solution NMR

**Collected data** are the FID curves which tell us about the time-dependent magnetisation relaxation that is recorded after/while applying the RF pulses.

**Strengths** are e.g. that NMR gives a lot of information about several aspects like conformational heterogeneity etc. Other than this NMR signals are sensitive to electronic and spatial environments and rather high resolutions are possible. Since the recording is in solution, realistic states can be analysed.

**Disorder tolerance** is high if the interconversion between conformers is sufficiently slow. If the interconversion is fast, the spectra tend to overlap and can't be assigned. Ordered and disordered regions can be distinguished in the 3D projection.

In order to really get a complete 3D structure one needs as well information about the molecular force field to simulate the annealing.

Different methods allow for different information, for instance COSY allows for through-bond 2D correlation spectroscopy and NOESY allows for 2D through-space correlation spectroscopy.

### 6.4 Solid state NMR

Fundamentally similar to solution NMR. Solid-state NMR eliminates the inherent tumbling in the solution, thus allowing control over how exactly orientations are measured or averaged out. This control is exerted by a magic angle spinning (MAS).

The primary use of ss-NMR in biology have been in the structural characterization of **amyloid fibrils and membrane proteins**.

### 6.5 Small-angle X-ray/Neutron scattering

**Collected data** The 2D pattern of intensities of a scattered beam is collected.

**Strengths** It allows for the detection of well-defined oligomeric states of large enough particles and detects average size and shape features of particles.

**Disorder tolerance** SANS and SAXS can be applied to partially or fully disordered systems. However, the normalized information content becomes very low. Disorder means that we have to assume different species are present: this is called a polydisperse system.

### 6.6 Circular dichroism spectroscopy

**Collected data** Absorption spectra of circularly polarized light.

**Strengths** Detection of average secondary structure contents of polypeptides. Detection of onset of (partially) ordered (amyloid-like) aggregation (visible as  $\beta$ -secondary structure).

**Disorder tolerance** CD spectroscopy can be applied to disordered systems. As for SANS/SAXS, the normalized information content becomes very low. Artifacts from aggregation can be a problem due to the moderately high concentration requirements. The CD result is always particle averaged. Spectral decompositions will not work very well unless the sample is homogenous.

## 6.7 Förster resonance energy transfer

**Collected data** Photon arrival times for two wavelengths (donor and acceptor fluorescence of two added labels).

**Strengths** Very good time resolution combined with single molecule resolution gives a dynamic view of individual molecules albeit projected onto a single coordinate.

**Disorder tolerance** FRET can be applied to disordered systems and the raw signals can theoretically resolve this disorder. While artifacts from aggregation are unlikely and identifiable, dye-molecule interactions can be an issue, especially for hydrophobic moieties. An effective particle- and time-averaging occurs if secondary analyses like transfer efficiency histograms or correlation functions are produced.

	X-ray	Cryo-EM	NMR	ss-NMR
Atomic res	✓	✓	✓	✓
Particle avg	✓	✓	✓	✓
Time avg	✗	✗	✓	✗
Disorder avg	✗	✗	✓	✗
Dynamic	✗	✗	✓	✗
Cheap, fast	✗	✗	✗	✗
Artifacts	✗	✗	✓	✗
Labeling	✓	✓	✓	✓
Radiation	✓	✓	✗	✗

	SAX(N)S	CD	FRET
Atomic res	✓	✗	✗
Particle avg	✓	✓	✗
Time avg	✗	✓	✗
Disorder avg	✗	✓	✓
Dynamic	✗	✓	✓
Cheap, fast	✗	✓	✓
Artifacts	✗	✓	✗
Labeling	✓	✗	✓
Radiation	✓	✗	✓

## 6.8 Fundamental caveats

- The experimental methods listed here study the macromolecules under heterogeneous conditions. This heterogeneity is sometimes more pronounced (crystalline vs. soluble states in X-ray vs. NMR) and sometimes less pronounced (different buffers and sample concentrations).
- In isolation or even in combination with other experimental results, these experimental results are insufficient to produce an atomic resolution 3D model.
- It is critical to understand the effects of **averaging**, both across **particles** in ensemble experiments and across **time** (pulses of laser light or radio-frequency RF oscillations have finite lengths; signals need to be collected for some time to have sufficient length). Otherwise, these experiments can lead to very *misleading* interpretations.

## 6.9 The PDB format

Some general properties of \*.pdb files:

- Most protein structures can be found in the protein data bank, where files are stored in a \*.pdb format ([www.rcsb.org](http://www.rcsb.org)). The PDB uses a 4-character code to uniquely

identify a structure. Currently, the first character is always a number and the code is assigned sequentially.

- For a structure to be released to the public, it has to pass **validation** tests.
- It is a **fixed column** format.
- The first **6 characters** determine the type of record on every line of the \*.pdb file. The most important are the ATOM and the HETATM records.

**ATOM records** :

RECORD	SERIAL	NAME	X-COOR.	Y-COOR.	Z-COOR.	ELEMENT
1-6	7-11	13-16	31-38	39-46	47-54	77-78
ATOM	1214	CA	7.889	-22.223	-32.139	C
ATOM	1215	C	6.692	-23.191	-33.257	C
ATOM	1216	O	7.552	-23.740	-33.964	O
ATOM	1217	CB	7.437	-20.829	-32.712	C
ATOM	1218	CG1	6.254	-20.230	-33.487	C
ATOM	1219	CG2	7.876	-19.891	-31.591	C
ATOM	1220	N	5.393	-23.421	-33.392	N
ATOM	1221	CA	4.870	-24.198	-34.510	C
ATOM	1222	C	3.540	-23.603	-34.947	C
ATOM	1223	O	2.911	-22.850	-34.193	O
ATOM	1224	CB	4.702	-25.669	-34.115	C
ATOM	1225	CG	3.564	-25.949	-33.119	C
ATOM	1226	CD	3.752	-27.291	-32.418	C
ATOM	1227	NE	4.745	-27.217	-31.338	N
ATOM	1228	CZ	5.144	-28.256	-30.602	C
ATOM	1229	NH1	4.651	-29.472	-30.818	N
ATOM	1230	NH2	6.043	-28.004	-29.643	N
ATOM	1231	N	3.119	-23.930	-36.166	N
		THRA			1.00	25.12

RESIDUE NAME	CHAIN	RESIDUE NR.	OCCUPANCY	B-FACTOR
18-20	22	23-26	55-60	61-66

**HETATM records** :

HETATM	1829	S	S04	A	601	-4.077	25.683	21.846	0.50	11.36	S
HETATM	1830	01	S04	A	601	-4.403	27.069	22.085	0.50	21.79	O
HETATM	1831	02	S04	A	601	-3.097	25.356	20.815	0.50	11.67	O
HETATM	1832	03	S04	A	601	-3.596	25.057	23.049	0.50	14.89	O
HETATM	1833	04	S04	A	601	-5.247	24.968	21.330	0.50	13.21	O

These records work the same way as ATOM records with a few adaptation. HETATM records are used for everything that is **not a standard biopolymer (protein/DNA/RNA) residue**. This includes nonstandard residues like hydroxyproline and acetyllysine, covalent modifications like sugars, prosthetic groups, noncovalent specific binders, metal ions, water, crystallisation buffer components. For covalent species, the chain letter is only sometimes assigned, i.e. there is no rigorous definition.

More details:

- Spatial units are in Å.
- Serial numbers are 1-indexed and used by CONECT records.
- the precision of the \*.pdb file format is 0.001 Å, meaning that many software programs developed their own file formats.
- Residue number refer to an underlying exon; they are difficult to interpret if there are in/dels for artificial constructs (header required). This is particularly problematic if coordinates are missing for entire residues, which are simply skipped in the ATOM records.
- Residue insertion can be highlighted using column 27; they are necessary because the residue numbering remains fixed to the exon, which would create identical residue numbers for different residues.

- Residue names for standard biopolymers are the 3-letter codes; for everything else, there is a limited but standard naming (e.g. HOH, SO4, EDO)

Additional information can be provided to ATOM and HET-ATM records, such as anisotropic temperature factor information, which is included in ANISOU records (and bloat file size).

**Multiple models** In crystallography, it may be that two alternative solutions of comparable quality exist for the coordinates of a particular group, This is explicitly annotated in the ALTLOC column, and you must choose one of them.

**CONNECT records** CONNECT records explicitly specify the bond connectivity matrix for all nonstandard entities redundantly. They **do not carry information about the type of bond**, which is not present in crystals anyway.

**SEQRES records** This contains the exact sequence of amino acids that was present in the experiment for biopolymer chains. The HET records list only those other groups that are resolved in the structure, and they exclude water molecules.

**DBREF records** These entries provide a sequence reference database reference per chain.

**SEQDAV records** They are the records describes differences between SEQRES and DBREF entries.

**REMARK records** more free-format entries; but the codes are free and REMARK 465 is the one reporting the residues that are **missing** from a 3D structural model.

**TER records** They signal the end of a biopolymer chain and the presence of a free carboxylic acid for a 3'-OH. The record is found at the end of the corresponding coordinate section (ATOM record).

**END records** Signal the end of an input file

**MODEL/ENDMDL records** they delineate alternative conformations when the entire coordinate section adtrops different values. They are wrapping multiple coordinate sections into the same file. This is how NMR structures are usually deposited, i.e. an ensemble of similar structures, which are all consistent with the experimental data.

**CRYST1, ORIG and SCALE records** Specifies the unit cell geometry for crystallography.

## 6.10 Molecular representation

It makes no physical sense to pretend that molecules can be represented as macroscopic objects, but insights can be gained from their representation; yet, it is important to recalibrate our understanding of these representations with the experimental reality. The ways by which these substances can be represented is as follows:

- Macromolecular (=cartoon, ribbon) representation
- Space-filling (=sphere, surface) representation
- Diagrammatic (=sticks, wire, licorice) representation

Each representation depends on what you want to achieve with the visualisation.

**Software** various software exists to visualize molecules on machines (UCSF Chimera, PyMOL, VMD) and in browsers (NGL viewer, JSmol). Other software also enables the simulation, calculation and modelling, such as Maestro. In order to produce publication quality graphics, tools such as OpenGL are often not enough, and tools offering ray tracing could be used to enhance the quality of images.

### Ways of packaging molecular visualizations

**Static images** : *Strengths*: easiest to customize, immutable (i.e. publishable), compact. *Weaknesses*: only a few structures can be visualized at once, projection inevitably hides something, poor depth perception.

**Movies** : *Strengths* Easy to customize, immutable, can be used to show multiple structures in sequence or the same structure from different angles. *Weaknesses* File size, blur artifacts, difficult to make comprehensive.

**Interactive sessions** : *Strengths* Maximum information disclosure. *Weaknesses* Not publishable, difficult to customize, file size.

## 7 The Data Scientist's View

### 7.1 Structural biology techniques with a focus on disorder

A 3D model can be seen as fitting a function with  $3N$  parameters, where  $N$  is the number of atoms and each atom has a 3-element position vector. An ensemble of  $M$  models can be understood as a fitting function with up to  $3MN + M - 1$  parameters, with  $M - 1$  being the weights of each member of the ensemble. **More data is needed** to constrain the structural ensemble. However, the reasons why such a model is needed is when **less data is available**.

### 7.2 Probabilistic interpretation of structure prediction

The resultant model should be evaluated with reference to a particular set of observations, for example a set of diffraction patterns from a crystallography experiment. Let's call this set of observations  $\mathcal{O}(z)$ .  $P(\mathcal{O}(z)|\mathbf{R})$  is the prior probability (Likelihood) of obtaining a particular diffraction pattern from an assumed 3D model ( $\mathbf{R}$ ) of the molecule believed

to be crystallized. The observations in place can be modelled as a likelihood function  $\mathcal{L}(R; O(z))$ , which can be defined as the product of individual likelihoods given the observations are independent:

$$\mathcal{L}(R; O(z)) = \prod_{i=1}^N f_r(z_i) \quad (10)$$

$$\ln \mathcal{L}(R; O(z)) = \sum_{i=1}^N \ln f_r(z_i) \quad (11)$$

*Ansatz:* If we can predict  $z(R)$  and we assume  $\mathcal{L}(R; O(z))$  is normally distributed around where the predictions  $z(R)$  and the observation  $O_i(z)$ , then we have:

$$\mathcal{L}(R; \{z_i, \sigma_i\}) \propto \prod_{i=1}^N \exp \left[ \frac{-(z(R) - z_i)^2}{\sigma_i^2} \right]$$

and

$$\mathcal{L}(R; \{z_i, \sigma_i\}) = - \sum_{i=1}^N \left[ \frac{(z(R) - z_i)^2}{\sigma_i^2} + f(\sigma_i) \right]$$

Due to the complex nature of representing a diffraction pattern, numerical optimization techniques need to be used to find the maximum of this equation. Also  $\sigma_i$  are not constrained, so they become free parameters.

**The likelihood is not enough** Given there are some assumptions (prior knowledge) that can be derived from e.g. the primary sequence about local covalent geometries, we want to include them in our model and check if they are verified in our observations, which is not always the case. Due to the fact that we rely on prior information, we would also want to evaluate the probability of our model given that we made some observations:  $P(R|O(z))$ .

**Back to the Bayesics (cringe alert!)** The prior probability of observing values given a model  $P(O(z)|R)$  and the posterior probability of observing the model given the observations  $P(R|O(z))$ :

$$P(R|O(z)) = \frac{P(O(z)|R)P(R)}{P(O(z))} = \frac{\mathcal{L}(R; \{z_i\})P(R)}{P(O(z))} \quad (12)$$

where  $P(\mathbf{R})$  is the model prior, it is a way to quantitatively describe that a given model fulfills some basic requirements. They are subjective and are meant to incorporate prior data on primary sequence and covalent geometries of chemical groups into estimation tasks. The posterior essentially interpolates between prior and likelihood.

Prior to public release of a structure, a given structure needs to pass validation tests, which evaluates whether a structures has sound covalent geometries and does not contain steric clashes.

**Maxwell-Boltzmann statistics** can be used to formulate a more universal model prior. Consider a collection of  $N$  point masses with sets of positions  $\mathbf{R}$  and momenta  $\mathbf{P}$ . They define a microstate as a realization of the system with specific values

for all independent degrees of freedom  $\mathbf{P}$  and  $\mathbf{R}$ ; its total energy is composed of potential and kinetic energy. The probability density to find a specific microstate at a given temperature  $T$  is exponentially distributed with its total energy. Specifically, the kinetic energy is defined as:

$$E_k = \sum_{j=1}^N \frac{\mathbf{p}_j^2}{2 \cdot m_j}$$

and

$$E_t = E_k + E_p = f(\mathbf{R}, \mathbf{P}).$$

where

$$f_i \propto \exp [-E_t(\mathbf{R}^i, \mathbf{P}^i)/k_b T]$$

which implies

$$\frac{f_i}{f_k} = \exp [-\Delta E_{t,i-k}/k_b T].$$

Note that:

$$f_i = \frac{1}{Q} \exp [-E_t(\mathbf{R}^i, \mathbf{P}^i)/k_b T]$$

Where

$$Q = \int_{-\infty}^{+\infty} \exp[-E_t(\mathbf{R}, \mathbf{P})/k_b T] dr_{1,x} dr_{1,y} dr_{1,z} dp_{1,x} dp_{1,y} dp_{1,z} \dots$$

The ensemble average of an observable  $X$  defined as:

$$X = g(\mathbf{R}, \mathbf{P})$$

results in:

$$E[g(X)] = \int_{-\infty}^{+\infty} g(x) f(x) dx.$$

concretely:

$$\begin{aligned} \langle X \rangle &= \sum_{\text{microstates}} X_i \mathbf{P}_i \\ &= \left( \frac{1}{Q} \right) \cdot \int_{-\infty}^{+\infty} g(\mathbf{R}, \mathbf{P}) \cdot \exp \left[ -\frac{E_t(\mathbf{R}, \mathbf{P})}{k_b T} \right] d\mathbf{R} d\mathbf{P} \end{aligned}$$

Now, if we consider a separable Hamiltonian (where  $X = h(\mathbf{R})$ ), the potential energy is only a function of the  $\mathbf{R}$  and the kinetic energy is only a function of the  $\mathbf{P}$ , then:

$$Q = f_r \int_{-\infty}^{+\infty} h(\mathbf{R}) \cdot \exp [-E_p(\mathbf{R})/k_b T] d\mathbf{R}$$

The probability distribution function defines entirely the thermodynamics of the system. We can perform a change of variables, for example from coordinates to energies, and this entails a density of states in the form of a Jacobian.

$$f(E = E_i) \propto g(E_i) \cdot \exp [-E_i/k_b T]$$

External constraints can be incorporated in a thermodynamic ensemble, which is a partition function which includes these constraints. These constraints are the number of particle  $N$ , the total volume  $\mathbf{V}$  and the temperature  $\mathbf{T}$ , which is called the canonical ensemble.

**Putting things together** We start with Bayes:

$$P(\mathbf{R}|O(z)) = \frac{P(O(z)|\mathbf{R})P(\mathbf{R})}{P(O(z))} = \frac{\mathcal{L}(\mathbf{R}; \{z_i\})P(\mathbf{R})}{P(O(z))}$$

With the help of the aforementioned *Ansatz* for the likelihood and the statistical mechanics results of the prior, we get:

$$P(\mathbf{R}|O(\mathbf{z})) \propto \frac{\prod_{i=1}^N \exp\left[-\frac{(\mathbf{z}(\mathbf{R})-\mathbf{z}_i)^2}{\sigma_i^2}\right] \exp[-E_p(\mathbf{R})/k_B T]}{P(O(\mathbf{z}))}$$

Rearranging yields:

$$\begin{aligned} P(\mathbf{R}|O(\mathbf{z})) &\propto \frac{\exp\left[\sum_{i=1}^N \frac{-(\mathbf{z}(\mathbf{R})-\mathbf{z}_i)^2}{\sigma_i^2} - E_p(\mathbf{R})/k_B T\right]}{P(O(\mathbf{z}))} \\ &= \frac{\exp\left[-\frac{1}{k_B T} \left(\sum_{i=1}^N \frac{k_B T (\mathbf{z}(\mathbf{R})-\mathbf{z}_i)^2}{\sigma_i^2} - E_p(\mathbf{R})\right)\right]}{P(O(\mathbf{z}))} \end{aligned}$$

The model prior  $E_p$  represents quasi-classical biomolecular force fields. A force field is a functional form of  $E_p(\mathbf{R})$ , which is ideally differentiable w.r.t.  $\mathbf{R}$ , and is the relevant representation of a molecule's identity. An example of irrelevant representation would be to consider helices as springs due to their cartoon representation. A quasi classical force-field contains both **soft** (dispersion, electrostatic interactions and rotational barriers) and **stiff** terms (covalent interactions, excluded volumes). Besides, bonded terms describe local interactions governed by electronic structure. Nonbonded terms describe nonlocal interactions, which govern both inter and intramolecular interactions. Exclusion rules determine what are the nonlocal interactions in the same molecule. As a result, the model prior can be expressed as follows:

$$\begin{aligned} E_p(\mathbf{R}) &= \underbrace{\sum_{\text{bonds}} U_{ij}^b(|r_{ij}|)}_{\text{Stretch of cov. bonds}} + \underbrace{\sum_{\text{angles}} U_{ijk}^\theta(\theta_{ijk})}_{\text{Bend of cov. bonds}} + \underbrace{\sum_{\text{dihedrals}} U_{ijkl}^\phi(\phi_{ijkl})}_{\text{Rotational barriers}} \\ &+ \underbrace{\sum_{\text{nonbonded}} U_{ij}^{LJ}(|r_{ij}|)}_{\text{Dispersion and excl. vol.}} + \underbrace{\sum_{\text{nonbonded}} U_{ij}^{Cb}(|r_{ij}|)}_{\text{Electrostatic terms}} \end{aligned}$$

where  $|r_{ij}| = f(\mathbf{r}_i, \mathbf{r}_j)$  and  $\theta_{ijk} = g(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k)$  and  $\phi_{ijkl} = h(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k, \mathbf{r}_l)$ . LJ stands for Lennard Jones and Cb stands for Coulomb's law, which are simplifications of the real non-bonding interaction to due omission of a number of properties.

**The forward prediction problem** This problem addresses whether or not we can predict  $z(\mathbf{R})$  (*the observables*) from a set of parameters  $\mathbf{R}$  (*the model parameters*). Due to the fact that most of the experimental techniques interact with matter, thus containing quantum effects due to the interactions, the only reliable way to make a structural prediction is to make an *ab initio* simulation engine. The latter is computationally unfeasible, therefore scientists rely on fully or partially empirical heuristics. If not fully empirical, these heuristics are often derived from or motivated by analytical results

for highly stylized cases. This can be illustrated by the nuclear Overhauser effect, which intensity is proportional to  $1/r^6$  of the distance between two NMR-active nuclei for independent sites tumbling freely in solution. However, this is never realistic scenario, and the NOE distance dependence can be anywhere from  $1/r$  to  $1/r^6$ , because the assumption that two independent sites tumble freely in solution is often violated.

**Sampling from the posterior distribution** Estimating  $P(\mathbf{R}|O(\mathbf{z}))$  is done by any number of numerical optimization techniques, especially finite temperature sampling algorithms. We can either:

- find the maximum (maximum a posteriori, MAP).
- find the subspace where the posterior is large, giving us a quantification of uncertainty of the predicted 3D structure.

To get candidate models, we draw samples from  $\mathbf{R}$  and thereby estimate  $z(\mathbf{R})$ . We would like to sample from the posterior directly, alternatively we can sample from the prior or the likelihood or uniformly. Usually, sampling from the prior is done since it is more discriminative.

Since the inference problem can be viewed as a thermodynamic ensemble with an augmented potential energy:

$$E_{p,\text{total}} = \sum_{i=1}^N \frac{k_B T (z(\mathbf{R}) - \mathbf{z}_i)^2}{\sigma_i^2} + E_p(\mathbf{R})$$

The term  $\frac{1}{\sigma_i^2}$  is often treated as free parameters but should ultimately originate from the statistical errors in the observations. The most widely used technique is molecular dynamics which relies on the numerical integration of suitable equations of motion. The main alternative is to use random perturbations and accept them according to rules that yield a Boltzmann distribution, i.e. Monte Carlo sampling.

**Molecular dynamics** Molecular dynamics essentially try to find solutions for Newton's equations of motion for multiple particles. Newton's equations of motion link the time derivative of the momentum (acceleration) of every particle to the gradient of the potential energy w.r.t. the absolute position of that particle. Constraints on temperature (thermostats) and pressure (manostats) need to be added to not sample from a constant energy ensemble. An example of such equations include the Langevin equation:

$$\begin{aligned} \dot{\mathbf{p}}_i &= \mathbf{F}_i = \underbrace{-\nabla_i U(\mathbf{R})}_{\text{Newton}} + \underbrace{\sqrt{2\gamma_i k_B T m_i} \mathbf{C}(t)}_{\text{Noise}} - \underbrace{\gamma_i m_i \dot{\mathbf{r}}_i}_{\text{Friction}} \\ \text{and } \dot{\mathbf{r}}_i &= \underbrace{m_i^{-1} \mathbf{p}_i}_{\text{Newton}} \end{aligned}$$

which offers to sample a canonical ensemble directly.

**Monte Carlo sampling** Monte Carlo samplers generate random perturbations of the system's coordinates and calculate the difference in potential energy between the two configurations. Then, the sampler applies an acceptance criterion such

as the Metropolis one, where the probability of acceptance is defined as:

$$p_{\text{accept}} = \min \left[ 1, \exp \left( -\frac{\Delta U_{p,\text{total}}}{k_b T} \right) \right]$$

$$= \min \left[ 1, \exp \left( -\frac{U_{p,\text{total}}(\mathbf{R}_{\text{new}}) - U_{p,\text{total}}(\mathbf{R}_{\text{old}})}{k_b T} \right) \right]$$

The algorithm used to propose new configurations is known as the move set, which is an empirical move set of rules and parameters, which quality is problem-specific.

**Constraints on numerical optimization** In addition to the *ensemble quantities* constraint inserted to the thermodynamic constraints, additional geometric constraints are applied. It essentially sets the prior probability of all models violating the constraint to 0. There are three types of geometric constraints:

**Trivial constraints** are those that simply do not consider particular coordinates as modifiable by the sampler. They include the degrees of freedom of the structure, and include absolute positions of atoms or sets of molecular coordinates for rigid body motion and dihedral angles for internal motion.

**Nontrivial constraints** are those who do not correspond to individual degrees of freedom, and include fixing the bond lengths in molecules when using the absolute positions of atoms as degrees of freedom.

**Implied constraints** are those that arise from choosing a set of degrees of freedom that is not just the position of atoms, and include choosing dihedral as explicit degrees of freedom, which implies that the bond lengths are fixed.

## 8 The Pharmaceutical View

### 8.1 The drug discovery pipeline

The pre-clinical drug discovery pipeline looks as follows:

1. **Target selection**
2. **Hit identification:** High-throughput screening of 3D multicellular, primary human and iPSC models at a scale suitable for large chemical library and genome-wide screening:
  - Microfluidic/miniaturized screening formats
  - Defined media or substrate
  - Novel 3D/multicellular assays with standardized analysis
3. **Lead identification**
4. **Lead identification**
5. **Preclinical candidate nomination**

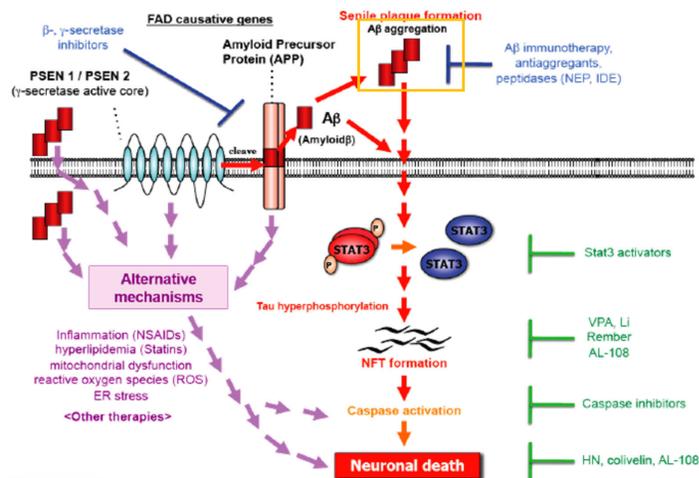
General considerations for drug discovery:

- The identified target should be modulated in the desired direction in humans. Target validation is tricky because almost all targets do not just modulate one particular cellular process, and all processes do not just involve a single target.

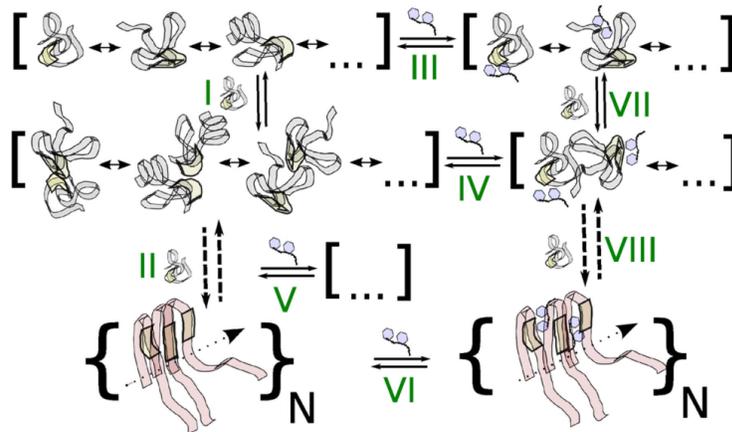
- Absorption, distribution, metabolism and excretion (ADME) properties should be fulfilled. They can be incorporated as prior information.
- Control side effects: short/long term toxicity to long-term interference with other drugs
- Make assays cheap, broad & more accurate.
- Avoid PAINS (pan-assay interference compounds) which show a response in assays for the wrong reason.
- Quantify the effects of the drug.

### 8.2 Disorder vs. drugs: the case of Alzheimer's disease

For Alzheimer's disease, the approved drugs are symptom-alleviating but not disease-altering. No approved drug follows directly the traditional view of the progression of Alzheimer's disease, which is the Amyloid hypothesis and the accumulation of tau proteins.



Overview of the amyloid hypothesis and MOA of (potential) drugs.



The many states of oligomerization and conformations of amyloid β make it a difficult target.

**Structural aspects of binding** If the targets or the drugs are disordered, the complex cannot be described by a single conformation. In this case, the difficulty stems from the fact that

there is no clear structural interpretation despite being thermodynamically and kinetically well-defined. In the case of A $\beta$ , several compounds interfere with fibril formation of amyloid  $\beta$ , but the mechanistic understanding of how this works or whether it is likely to be beneficial is very poor.

The models for understanding drug complexes with biopolymers are **conformational selection** where the dissociated entities sample the same conformations as in the complex with reasonable likelihood is *higher* than **induced fit**, where the likelihood of finding the unbound complex in the same conformation is very *small*. The **lock-key** model is an extreme case of conformational selection with just single structures for both partners. Structural interpretations form the heart of **structure-activity relationships** and rational drug design. However, many drug targets are partially disordered like transcription factors. Still, disorder does not imply low affinities. In the cell, the **functional binding** of disordered species *increases order* (a process called folding upon binding) unlike in disease, where **pathological binding**, or intracellular aggregation, is accompanied by a *decrease in order*.

**Empiricism** in drug discovery is required to try to understand how a drug works as insights are gained from the effect it has on the organism from acting on immediate target, as the **complexity of predicting the effect** of altering a target in the organism is too difficult. This is shown by the fact that the **MOA of many approved drugs and natural compounds is still unknown**. **Computational drug discovery** is built on the idea that it is possible to predict important physiochemical properties of small molecules from limited information in silico. This includes properties like solubility, binding affinities to various targets, or metabolic stability. **Chemoinformatics** refers to applying computational methods to small molecules. *The diversity of small molecules make the determination of atomic detail parameters much more challenging* relative to the limited diversity of amino acids. Knowledge-based approaches are gaining traction for making predictions as the available data grows. The main bottleneck is data availability for target-specific scenarios. Inference methods seems to be powerful in prediction organism-level effects such as drug compatibility and long-term side effects.

### 8.3 High-throughput screening

**High-throughput screening** means that many molecules are tested in one or more target-specific assays consuming as little time and as little material as possible with reproducible outcomes. Molecules tested by HTS represent a **library**, and can be virtualized in the case of virtual HTS. **Curation** is an essential feature of a library. In experimental libraries, this means that unambiguously identified compounds are available in soluble stable forms with well-defined stereoisomeric populations. In virtual libraries, this means that compounds are represented as *realistic* tautomers and protomers, they are *chemically stable* and easy to *synthesize* with reasonable effort. For structural applications, the library also needs conformers, which requires knowing the **covalent geometry**. In this context, the conformer problem is the challenge to propose reasonable three dimensional molecules for molecules that have relevant internal degrees of freedom, which can be

treated as prior information and implies the use of a force field.

### 8.4 Computational drug discovery

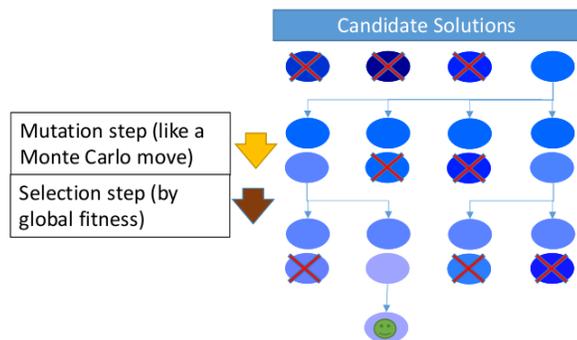
Computational drug discovery faces several problems:

- **The library problem.** Despite the near-exponential growth of libraries, not all chemotypes are well-understood structurally or chemically (large, flexible rings strained systems, multicyclic heteroaromatic rings) which makes curation difficult
- **The search problem.** The aim with computational drug discovery is to find a specific structure that has the highest effect and the lower toxicity. Each piece, however, are deformable, and in the case of small molecules can be *deformable*. What this yields is dozens of approximate fits and this decomposes into two subproblems: on the one hand, it is *difficult to see if a single molecule may interact* and on the other hand it is *not simple to exclude molecules easily*.
- **The scoring problem.** Discussed further below

Searching for optimal hits mostly relies on numerical optimization. Several properties hold for virtual HTS:

- For each compound, there are usually comparatively **few degrees of freedom**
- The actual degrees of freedom of the molecule are subject to significant constraints, such as the position of a target molecule is usually **constrained to what is desired/believed to be the binding site**.
- In general, there is **no well-defined distribution** makeup that is used or sought after in searching for a molecule.
- The results of searching as **biased towards physically meaningful orientations**. This relies on highly empirical energy functions, and are a low fidelity approximation of the binding thermodynamics. This problem is referred to as the **scoring problem**.
- The approximation is motivated by the fact that many molecules should be evaluated quickly, which introduces historical bias.
- Due to the low fidelity, **pose diversity** is sought after as well to avoid false negatives from approximation errors.

**Search and optimization algorithm** GLIDE can use *exhaustive search* predominantly for the ligand on its own. Other approaches, such as SEED, use *exhaustive search for rigid rotations* of ligands around heuristically selected vectors. For pharmacophore matching, Pharmit uses very fast *matching algorithms on simplified descriptors* to identify putatively interesting results. *Optimization methods like the Broyden-Fletcher-Goldfarb-Shanno algorithm combined with Monte Carlo sampling* such as Autodock Vina to search for optimal molecules. Algorithms can also be *parallelized* using for instance genetic algorithms score and then select multiple variants for further scoring based on the selected match. rDock uses this approach.



Representation of a genetic algorithm.

**The scoring problem** The primary goal of a score or scoring function in a drug discovery workflow is to be able to predict which molecules would be active and which would not. Experimentally, a score is derived from the **relative binding free energies** of different compounds,  $L$  and the target (receptor)  $R$ . The relative free energy given by any  $i^{th}$  pair of receptor and ligand is given by:

$$\Delta G_i^0 = -k_b T \ln \frac{\text{Concentr. at equilibrium } [LR]_{eq}}{[L]_{eq}[R]_{eq}} = \overbrace{\Delta H_i^0 - T\Delta S_i^0}^{\text{Enthalpy-entropy decomp.}}$$

and  $\Delta\Delta G_{jk} = \Delta G_k^0 - \Delta G_j^0$

Although predicting  $\Delta G_i^0$  is a perfect scoring function, it is to computationally expensive. Instead, a practical scoring function is implemented that must perform significantly better for inferring  $\Delta\Delta G_{jk}$ . Several assumptions are made when scoring molecules:

- A primary and implicit assumption is that binding is sufficient to achieve the desired effect. This assumption is not at work in experimental HTS.
- The entropy terms are ignored by assuming they are included in enthalpy terms and not discriminative, which works for homogenous libraries.
- Single molecular forms is representative (i.e. alternative protonation states are ignored).
- No conformational averaging is required to calculate a meaningful score, because the selected conformation is drastically energetically favorable compared to the others.

**Caveats of scoring and searching** The performance of a scoring function often varies across different sets of compounds and across different receptors. *There is no specific reason to use one energy function over another for searching and scoring.* The problem is that we sample from the energy function used for searching and have to rely on the assumption that the poses found this way are poses of high probability also in the energy function used for scoring. Combining ways to resolve these issues include:

- The energy function is chosen to be the same
- The scoring energy function can be viewed as containing the energy function with corrections. This can also be viewed as sample exclusively from the prior but score

according to the posterior, which incorporates corrections.

- The energy function can only be used to exclude clear cut cases, i.e. only use the energy function for crude guidance.
- The poses during searching are refined by applying the scoring energy function.

**Evaluation of scoring performance** During the hit identification stage, it is unlikely that molecules with very high affinities are found. Thus, the primary information taken from this stage is whether molecules pass a certain threshold. This means that the score has to be converted into a binary classifier. Every compound that passes the threshold are a predicted hit. Unless these prove to be hits experimentally, they are not an actual hit. This allows for the construction of a confusion matrix and the calculation of performance metrics. None of the screened molecules will have a significantly high value above the threshold (usually between 1 – 100  $\mu\text{M}$  for  $K_d$ ), and many will usually fail.

The ROC curve can be computed using the docking score as a control parameter. This curve can be easily biased due to the unbalanced nature of the dataset, which contains a lot of easy false negatives. The AUC and Youden's J statistic need to be reported.

**SEED case study Objective:** SEED is a program for fragment docking with force-field based evaluation of binding energy. **Input:** a conformer library, which contains multiple, pre-generated conformers of the same molecule. **Procedure:** SEED positions the ligand molecule in direct contact with selected sites on the receptor. This procedure aligns vectors formed, for example, by hydrogen bond-donating groups like OH or NH. Once the molecule is aligned to a given vector, it is rotated systematically around the vector axis and stores the most energetically favourable combinations of vector and rotation. The resultant similar poses are clustered to eliminate redundancy from the result, such that only a few remain. To predict the binding free energies, the following is evaluated:

$$\begin{aligned} \Delta G_i^0 &= \Delta H_i^0 - T\Delta S_i^0 \\ &= \Delta H_{i,\text{elec}}^0 + \Delta H_{i,\text{np}}^0 + \Delta H_{i,\text{strain}}^0 \\ &\quad - T(\Delta S_{i,\text{elec}}^0 + \Delta S_{i,\text{np}}^0 + \Delta S_{i,\text{strain}}^0) \end{aligned}$$

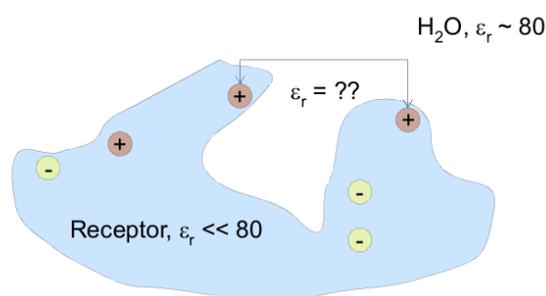
and individual  $\Delta H$  and  $\Delta S$  terms can be split further into solvent, solute and solute-solvent contributions. SEED neglects all entropy terms related to receptor or protein and water entropy, disadvantaging hydrophobic fragments from a scoring perspective. SEED approximates enthalpy using the Lennard-Jones form. The main result of SEED is a prediction of the electrostatic contribution to the binding free energy.

**Scoring:** to calculate the predicted electrostatic contribution to the binding, the process is split into 3 terms:

- *The receptor desolvation*, which is always unfavourable and approximated by the Poisson equation (see below)

- *Fragment desolvation*, which is always unfavourable and handled in a generalized Born model (see below)
- *Screened electrostatic interaction*, which is usually favourable and also handled by the GB model (see below)

**Continuum Electrostatics** The idea of CE is that there is a separation of length and time scales that allows the effects of a surrounding medium like water to be described like a mean field in relation to the solute of interest. The equation predicting the potential created by a distribution of charges given a spatially inhomogeneous dielectric is the **Poisson equation**. In the classical approximation of molecules, the spatial heterogeneity is a *low dielectric cavity* defined by the receptor immersed in a high-dielectric medium like water (see below). The charges stem from the partial charges from the atoms. The fact that Coulomb's law cannot be applied here comes from the fact that the dielectric is not **homogeneous**.



**Generalized Born Models** Solving the Poisson equation is costly, so many approximations have been developed: the most widely used being the Generalized Born equation. It approximates the receptor as a nonpolarizable sphere with with Born radius  $\alpha$  and a single charge at the center.

## Part III - Jelezarov

### 8.5 Thermodynamics of protein folding

The native conformation of a protein is stable in a narrow range of conditions. When proteins denature they lose their native, 3D structure. For most small monomeric proteins this process is reversible. Proteins fold back when transferred into begin conditions. The folding can be visualized with a funnel shape. When going down the funnel, the amount of possible conformations decreases and higher degrees of structures form. There is a difference between folding/unfolding and denaturation/renaturation.

**Folding** is an ideal concept - The folded state is formed starting from the unfolded state, which by definition lacks intramolecular non-covalent interactions and is fully accessible to the solvent (water).

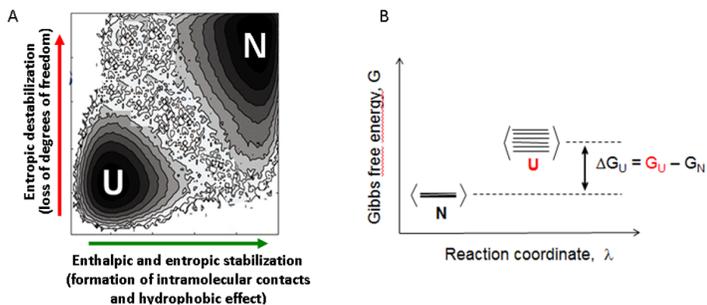
**Renaturation** is a real phenomenon - The native state is formed starting from the denatured state, which might have residual intramolecular non-covalent interactions (sometimes quite developed) and might be partially shielded from the solvent (water).

In this section we mainly aim to answer the question: "Why and how do proteins fold?" We want to answer this by thinking about the folding code, the folding mechanism and whether we can predict the native structure of a protein from its amino acid sequence. We will try to do this by introducing different models of protein folding.

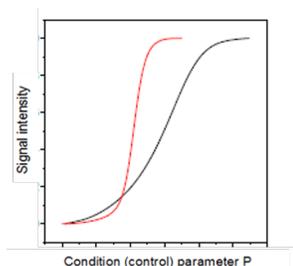
**The classical view** There is a well defined pathway that a protein passes before it's completely folded. This can be characterized by multiple intermediate states. In order to go from the folded to the unfolded state the protein must pass several energy barriers.

**The new view (funnel)** Multiple pathways are available, but overall the energy and entropy when going to the native state decreases.

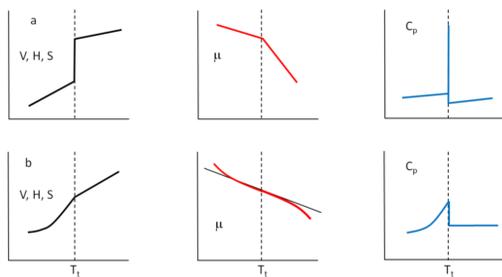
**The two state approximation cooperativity** The protein can only visit a limited number of states that are sufficiently stable and separated by high energy barriers. High cooperativity means that disrupting one interaction in the protein will destabilize other interactions. The protein can only be fully folded or fully unfolded. In the two state model we assume that all intermediate states between the unfolded and the native state are so short lived, that they don't contribute to the free energy. The two state model can be represented in two ways.



In the left image the entropic and enthalpic properties of the two states are shown. The native state is balanced by the enthalpy of the intramolecular bonds and the hydrophobic effect (favorable enthalpy). It is entropically penalized. The unfolded state has unfavorable enthalpy since it has few stable contacts, but it is entropically favorable since it also has few fixed states. The right image shows the free energies of the two states.



The signal intensity of a protein tells us something about the cooperativity of the protein. The red line indicates high cooperativity. There is a fast transition between the folded and the unfolded state and there are very few intermediate states. The dotted line indicates low cooperativity. The transition is slow because there are many intermediate states.



The top three images are plots for a highly cooperative protein and the bottom three images for a non-cooperative protein.

## Thermodynamic forces

**Conformational entropy increase** favours unfolding

**Enthalpy increase** favours folding because the native state contains many covalent bonds.

**Solvent entropy increase (hydrophobic effect)** favours the native state

**The reference state** The choice of the reference state is arbitrary. In protein thermodynamics the native state is defined as the reference state. This is because the structure of the native state is usually well characterized.

**Stability of two state monomeric proteins** Monomeric proteins consist of only one subunit. The proteins can be in the folded or unfolded state. The unfolding rate constant is defined as.

$$K_u = \frac{[U]}{[N]} \quad (13)$$

The fraction folded and unfolded protein can be derived as follows.

$$f_n = \frac{[N]}{[N] + [U]} f_u = \frac{[U]}{[N] + [U]} \quad (14)$$

From this we can define the Gibbs free energy.

$$\Delta G_u = -RT \ln \left( \frac{[U]}{[N]} \right) \quad (15)$$

$$= -RT \ln \left( \frac{f_u([N] + [U])}{f_n([N] + [U])} \right) \quad (16)$$

$$= -RT \ln \left( \frac{f_u}{f_n} \right) \quad (17)$$

$$= -RT \ln \left( \frac{f_u}{1 - f_u} \right) \quad (18)$$

**Stability of two state oligomeric proteins** Oligomeric proteins consist of multiple subunits. In the unfolded state we have  $n$  times as many molecules as in the folded state (where  $n$  is the amount of subunits).

$$K_u = \frac{[U]^n}{[N]} \quad (19)$$

$$f_n = \frac{[N]}{[N] + \frac{[U]}{n}}, \quad f_u = \frac{[U]}{n[N] + [U]} \quad (20)$$

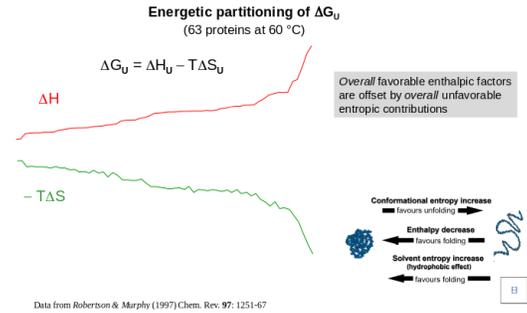
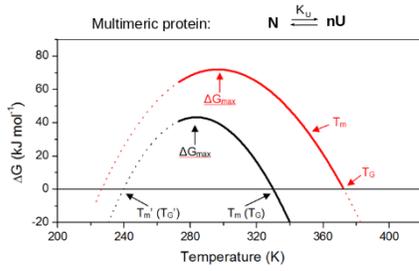
$$\Delta G_u = -RT \ln \left( \frac{[U]^n}{[N]} \right) \quad (21)$$

$$= -RT \ln \left( \frac{f_u(n[N] + [U])}{f_n([N] + \frac{[U]}{n})} \right) \quad (22)$$

$$= -RT \ln \left( \frac{n f_u^n U_0^{n-1}}{1 - f_u} \right) \quad (23)$$

$$(24)$$

With  $U_0$  being the total concentration of unfolded protein.



We see a stability curve in which the dependence of the Gibbs free energy on the temperature is shown for both monomeric and multimeric proteins.  $T_m$  is the temperature at which  $f_U = f_N = 0.5$  and  $T_G$  is the temperature at which the free energy is zero. For monomeric proteins we have at  $T_m = T_G$ :

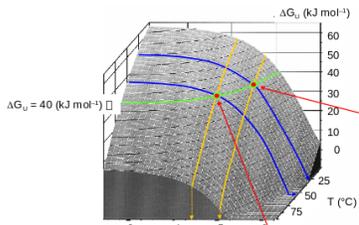
1.  $[N] = [U]$
2.  $f_U = f_N = 0.5$
3.  $K_u = 1$

For multimeric proteins this does not hold since  $T_G$  is defined at the temperature at which  $\Delta G = 0$ , which has a different definition for multimeric proteins:

$$\Delta G_U = -RT \ln \left( \frac{[U]^n}{[N]} \right) \quad (25)$$

At low temperatures there is also some unfolding. This is called cold denaturation. The reason for this is that at low temperatures the hydrophobic effect diminishes and the unfolded state is favored. However, for most proteins this occurs well below water freezing temperatures and is therefore not observed.

**Stability is not an intrinsic property of proteins** We measure stability by measuring  $\Delta G$ . This stability depends on many different parameters (pH, temperature, ionic strength, etc.). We can plot  $\Delta G$  with respect to different parameters.



**Energy properties of folding and unfolding** If we define  $\Delta G = G_U - G_N$

1. Folding is exergonic ( $\Delta G_U > 0$ ) while unfolding is endergonic ( $\Delta G_U < 0$ )
2. Folding is exothermic ( $\Delta H_U > 0$ ) and unfolding is endothermic
3. Entropy decreases upon folding and increases upon unfolding ( $T\Delta S_U > 0$ ).

We usually observe that enthalpy and entropy effectively cancel out each other.

$$\Delta G_U = \Delta H_U - T\Delta S_U \quad (26)$$

So overall we observe that proteins, regardless of structure, follow the same mechanism.

**Shape of the stability curve** The shape of the stability curve is determined by enthalpy, entropy and the heat capacity. They are given by the following equations, where  $T_R$  is a reference temperature.

$$\Delta H_U(T) = \Delta H_U(T_R) + \int_{T_R}^T \Delta C_P dT \quad (27)$$

$$\Delta S_U(T) = \Delta S_U(T_R) + \int_{T_R}^T \frac{\Delta C_P}{T} dT \quad (28)$$

$$\Delta C_P = C_P^U - C_P^N = \left( \frac{d(\Delta H)}{dT} \right)_P = \left( \frac{T d\Delta S}{dT} \right)_P \quad (29)$$

The heat capacity is the ratio of the total absorbed heat ( $dQ$ ) to the resulting increase of temperature ( $dT$ ). We have two definitions:

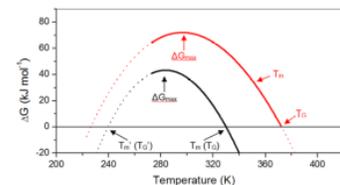
1.  $c_v = \left( \frac{dU}{dT} \right)_v$  for an isochoric process (volume remains constant)
2.  $c_p = \left( \frac{dH}{dT} \right)_p$  for an isobaric process (pressure remains constant)

The heat capacity is an extensive property, but for our purpose we only use intensive definitions like:

1. Molar heat capacity, units:  $JK^{-1}mol^{-1}$
2. Specific heat capacity (mass), units:  $JK^{-1}g^{-1}$

### The Gibbs-Helmholtz equation

$$\Delta G_{(T)} = \Delta H_{(T)} - T\Delta S_{(T)} = \Delta H_{(T_m)} + \int_{T_m}^T \Delta C_p dT - T\Delta S_{(T_m)} - T \int_{T_m}^T \frac{\Delta C_p}{T} dT$$



Integrated form:

Monomeric protein:

$$K_u(T_m) = 1, \Delta G(T_m) = 0, T_m = T_G$$

$$\Delta G_{(T)} = \Delta H_{(T_m)} \cdot \left( 1 - \frac{T}{T_m} \right) + \Delta C_p \cdot \left( T - T_m - T \ln \frac{T}{T_m} \right)$$

Multimeric protein:

$$K_u(T_m) \neq 1, \Delta G(T_m) \neq 0, T_m \neq T_G$$

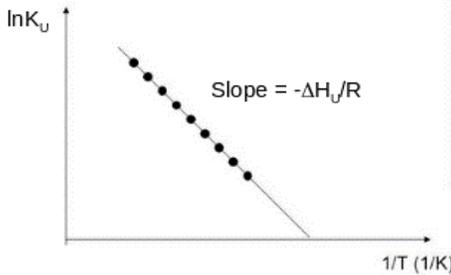
$$\Delta G_{(T)} = \Delta H_{(T_m)} \cdot \left( 1 - \frac{T}{T_m} \right) + \Delta C_p \cdot \left( T - T_m - T \ln \frac{T}{T_m} \right) - RT \ln K_u(T_m)$$

**Van 't Hoff Enthalpy** There are two ways to measure the unfolding enthalpy. One of them is the Van 't Hoff enthalpy.

$$\Delta G_U = -RT \ln K_U = \Delta H - T \Delta S \quad \ln K_U = -\frac{\Delta H_U}{RT} + \frac{\Delta S_U}{R}$$

$$\frac{d \ln K_U}{d\left(\frac{1}{T}\right)} = -\frac{\Delta H_U}{R}$$

Measure  $K_{eq}$  as a function of T:



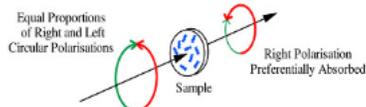
$$\frac{d \ln K_U}{dT} = \frac{\Delta H_U}{RT^2}$$

$\Delta C_p$  is assumed to be independent. This is a but the temperature var small between 0 °C and be neglected.

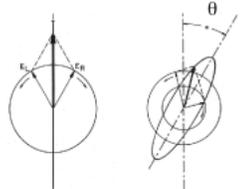
The plot can only be created at temperatures around  $T_m$ .

### Circular dichroism spectroscopy

Circular dichroism is the difference in absorption of left and right circularly polarised light by an asymmetric or chiral molecule.



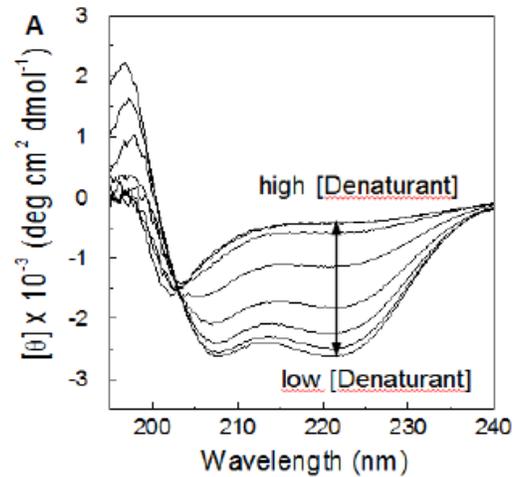
Mean Residue Ellipticity MRE (also called mean ellipticity per residue)



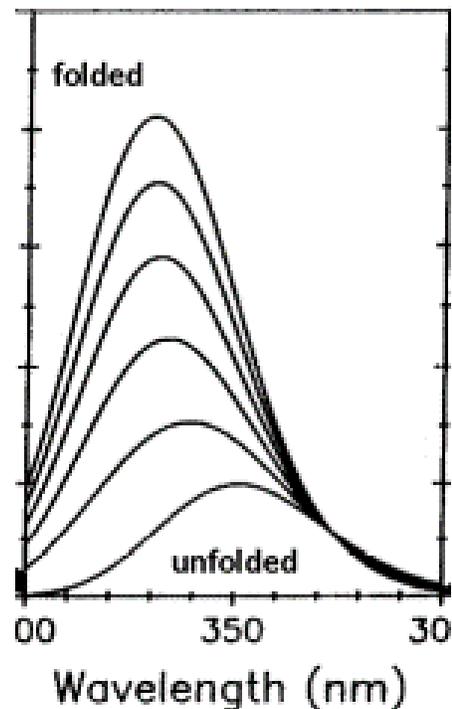
**CD in the far UV region**  $\lambda < 250nm$  This UV region gives information about the secondary structure. Alpha helices have a peak at 180 nm and beta sheets at 200 nm. We often look at the difference in the spectra between the folded and the unfolded protein.

**CD in the near UV region**  $\lambda > 250nm$  In this region we can find information about aromatic side chains (especially containing Trp). These side chains have a peak at about 280 nm. However, the intensity of the peak is very low and therefore we need high concentrations to observe anything.

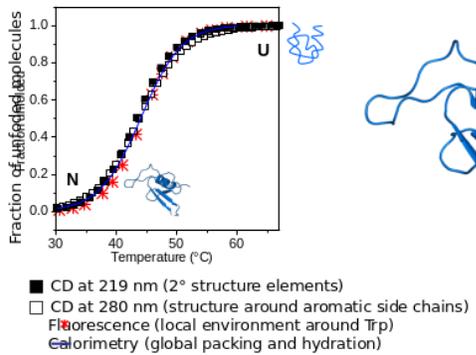
Often we make CD curves for many different temperatures or solvent compositions in order to observe the unfolding of a protein. In the image below we observe that upon higher urea concentration (thus unfolding) the signal for the secondary structures disappears.



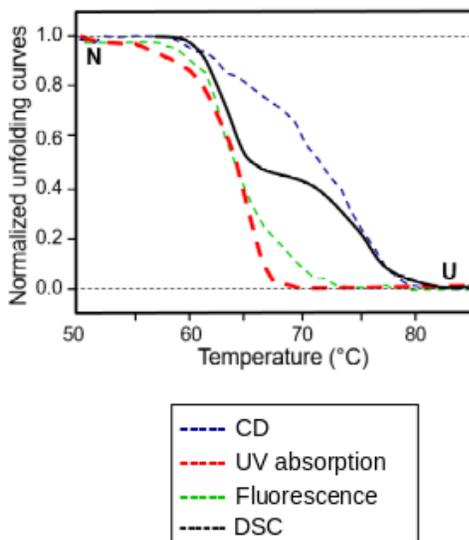
**Fluorescence spectroscopy** Fluorescence spectroscopy gives information about the region around Trp or an artificially introduced probe. Low concentrations are needed for fluorescence experiments. The downsides are that often intensities may cancel out and if the Trp is involved in an interaction in both the unfolded and the native state. You will not observe a difference between native and folded protein.



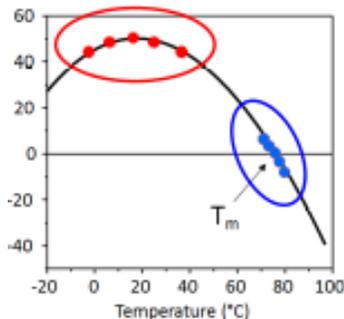
## Testing the two state approximation



Usually a combination of these techniques is used to validate the two state approximation of a certain protein. If the protein does not follow a two state folding process, the curves will resemble the image below. The results of the different experiments are not consistent, which probably means there are more than two states.

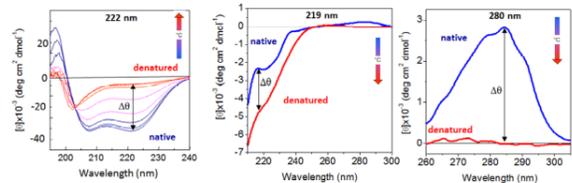


**Constructing protein stability curve** by thermal unfolding experiment and denaturant induced unfolding



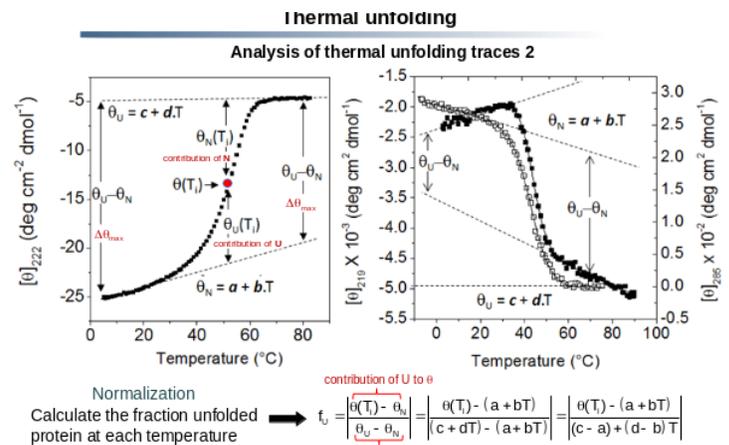
Thermal unfolding yields the melting temperature, the enthalpy and the heat capacity. In these experiments curves are made at different temperature. Then the wavelength is selected for which the difference between the curves for the low and high temperatures is largest. Changes are then monitored at this wavelength by continuously heating a protein with a constant rate.

## CD spectroscopy experiments

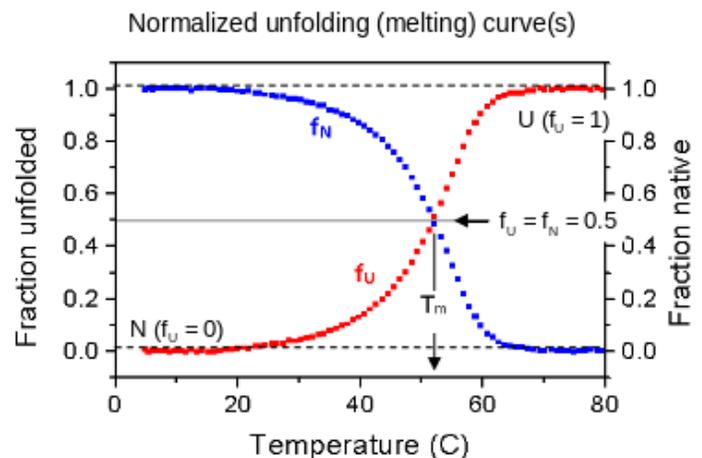


- 1) Select a wavelength ( $\lambda$ ) where the difference between N and U is maximal
- 2) Monitor the changes at  $\lambda$  by continuously heating at 0.5-2 °C/min

The measured signal at any moment is a linear combination of the signal of the folded state and the unfolded state, weighted by their corresponding fractions.  $S = f_U S_U + f_N S_N$  This means we can find these fractions at any point. From the image above we can plot the ellipticity at each temperature against the temperature and obtain a plot like the one below.

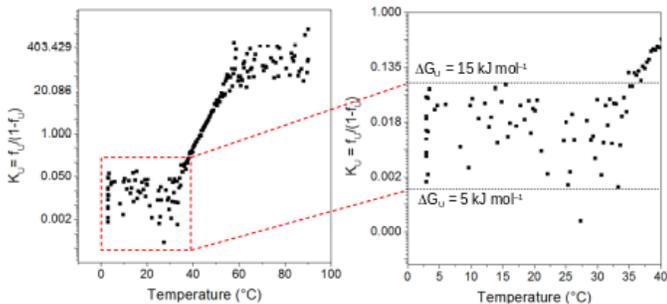


If we draw two lines tangent to the slope of the native and the unfolded state, the difference between the values of these lines at a certain temperature is  $\Delta\theta_{max}$ , the max difference in ellipticity at this particular temperature. We can also obtain the contribution of the folded and the unfolded state as shown in the figure. From this we can compute  $f_U$  at each temperature and plot this against the temperature. We obtain a **normalized** plot like the one below.



The exact same curves can be created with fluorescence spectra in the same way.

Since we know that  $K_U = \frac{f_U}{1-f_U}$ , we can make plots where we plot  $K_U$  against the temperature.



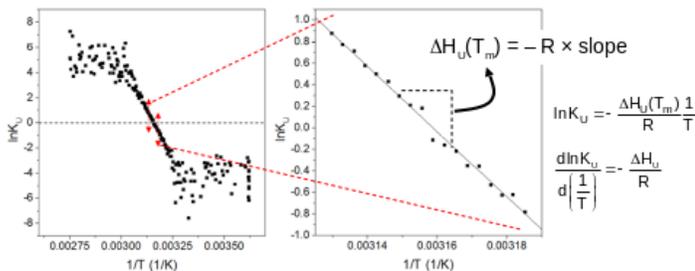
The typical stability of proteins is 10-40 kJ mol<sup>-1</sup>

We observe a large variation in the native and the unfolded state. Therefore, we cannot define  $K_U$  accurately in these regions. We only try to obtain information from the temperatures around the melting temperature,  $T_m$ .

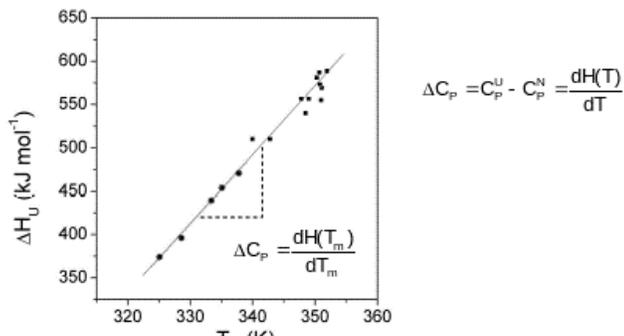
**Finding enthalpy** We can find the van 't Hoff enthalpy from the slope around  $T_m$  in the previous plot.

$$\Delta G_U = -RT \ln K_U = -RT \ln \frac{f_U N_U}{(1-f_U) N_N} = -RT \ln \frac{f_U}{1-f_U}$$

- ➡ Consider data points around  $T_m$   
 $T_m \pm 3^\circ\text{C}$     $0.4 < f_U < 0.6$
- ➡ Calculate  $K_U$  at each T
- ➡ Plot  $\ln K_U = f(1/T)$



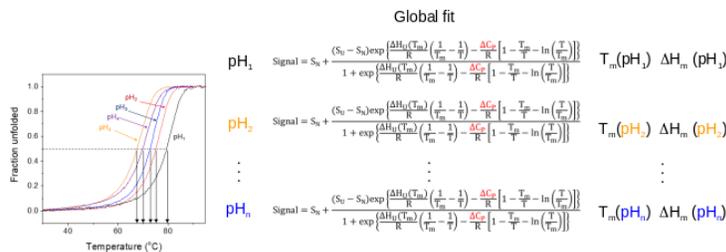
**Determining the heat capacity** Experiments are done at a different pH, salt concentration, etc. We obtain a  $T_m$  and an enthalpy pair for each of these pHs. A Kirchoff plot can be made in which the enthalpy is plotted against the melting temperature. The heat capacity is determined from the slope.



However, in order for this to work we need enough variation in the melting temperature and the enthalpy. If we don't have enough variation, the heat capacity cannot be determined accurately.

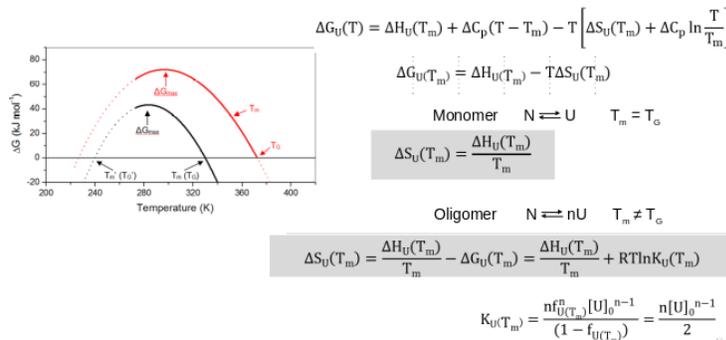
A different way to measure the heat capacity is by determining the exposure of molecular surface upon unfolding. The heat capacity changes depending on the water accessible surface. We can measure this water accessible surface and determine the heat capacity from this.

Another way to find the heat capacity is by using non-linear regression analysis. The heat capacity doesn't change, but the signal does. If you know the melting temperature and the enthalpy at different pHs, you can determine the heat capacity.



Shared fitting parameter:  $\Delta C_p$

**Determining entropy term** The entropy can be determined from the unfolding curves as well if the enthalpy and the melting temperature are known.

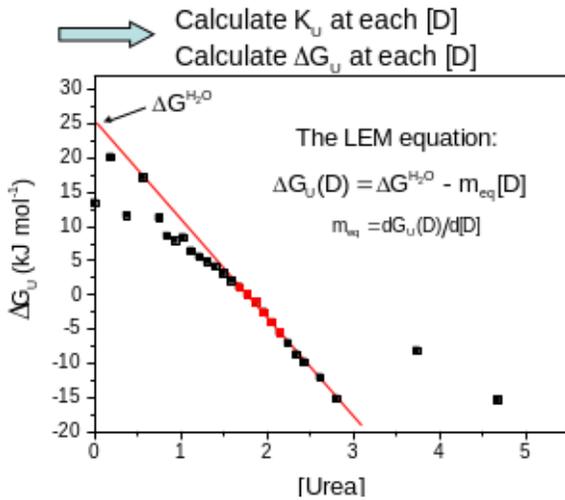


**Chemical Unfolding** In chemical unfolding experiments the temperature is fixed, while the concentration of denaturant varies. The concentration of denaturant required to achieve unfolding depends on its strength. The experiments consist of the following steps:

1. Fill different tubes with the same concentration of protein
2. Add increasing concentrations of denaturant
3. Adjust the pH in each tube such that it's the same (the denaturant can influence the pH)
4. Incubate until equilibrium is achieved. This step is very important.
5. Measure

Then we can measure the signal with CD or fluorescence spectroscopy, after which we can produce the same normalized curves in which we plot the fraction unfolded or folded

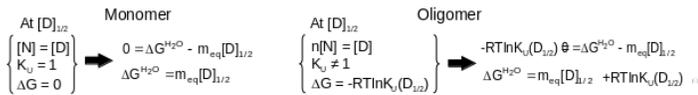
protein against the concentration of denaturant. Different parameters can be determined from those curves. For example, the free energy of unfolding can be determined at each concentration and the following curve can be constructed.



In this curve we can use the Linear Extrapolation Method (LEM) to find the Gibbs free energy at low concentrations of denaturant (we are usually interested in the native state of the protein), since we can't measure this directly because of the scattering at low (and high) concentrations. The LEM equation is:

$$\Delta G_U(D) = \Delta G^{H_2O} - m_{eq}[D] \quad (30)$$

We use  $[D] \approx [D]_{1/2}$ , the concentration at which  $f_U = f_N$  to determine  $m_{eq}$ . The resulting computation for the free energy at  $[D] = 0$  is different for monomers and oligomers.



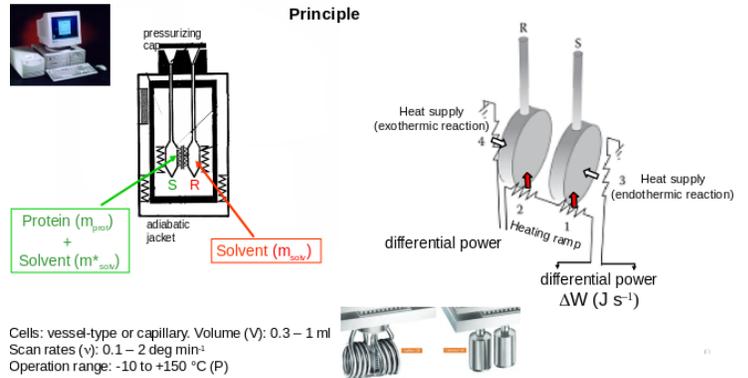
Using different denaturants shouldn't matter too much. They will give different curves, but they should intersect at the same point on the y-axis. In practice there often is a difference for different reasons. For example, some denaturants are salts and an increasing concentration of denaturant also means an increasing concentration of salts. These salts can stabilize the protein depending on the charge distribution of the protein.

**Creating unfolding curves at different temperatures** Each chemical unfolding experiment yields a free energy at a given temperature. However, this can only be done in a limited temperature range. Therefore, it is usually recommended to use combination of chemical and thermal unfolding. The free energy at high temperatures can be found with thermal experiments and at low temperatures with chemical experiments.

**Can intermediate states be determined from a spectroscopic unfolding experiment?** Intermediate states cannot be determined from an unfolding experiment. The reason is that the observed signal is a linear combination of all the

states weighed by the amount of each state. However, there is no thermodynamic linkage between the signal of a particular state and the free energy of that state. While we cannot detect particular intermediate states, we can sometimes detect that a protein is clearly not a two state protein. This can be seen by a non-smooth curve.

**Differential scanning calorimetry (DSC)** DSC allows for detection of intermediates. The reason is that certain parts of the protein may not unfold upon heating. DSC measures the heat capacity change upon unfolding, by which it also measures the change of hydration, which is an important addition.



If we do an unfolding experiment, both cells are heated at the same rate. In the sample cell an endothermic reaction takes place, which makes the temperature lower than the reference cell. A power is applied to make the temperatures in the sample and reference cells equal. This power is measured and is equal to the difference between the heat capacities.

$$\text{Differential power (J.s}^{-1}\text{)} \rightarrow \frac{\Delta W}{\nu} = C_p^{(\text{prot+solv})} - C_p^{\text{solv}} = \Delta C_p^{(\text{prot+solv})-\text{solv}} \quad (\text{J.K}^{-1})$$

Heating rate (K.S<sup>-1</sup>)      Heat capacity of the sample cell      Heat capacity of the reference cell

There are multiple definitions of the heat capacity. The **partial specific heat capacity** is defined as follows, with units  $JK^{-1}g^{-1}$ :

$$\Delta C_p^{\text{prot}} = C_p^{\text{prot}} m_{\text{prot}} - C_p^{\text{solv}} \Delta m_{\text{solv}} \quad (31)$$

$$\text{with } \Delta m_{\text{solv}} = m_{\text{prot}} \frac{\nu_{\text{prot}}}{\nu_{\text{solv}}} \quad (32)$$

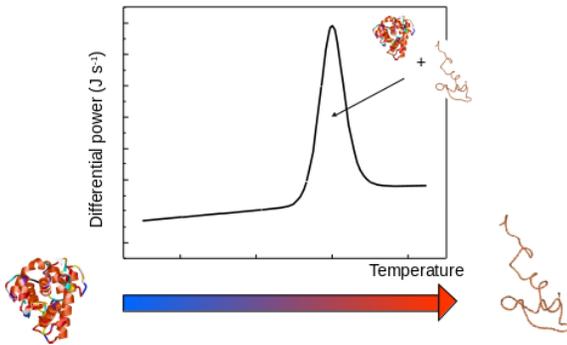
Here  $\nu$  is the partial specific volume of the protein and the solvent. We can express the heat capacity of the protein as

$$C_p^{\text{prot}} = -C_p^{\text{solv}} \frac{\nu_{\text{prot}}}{\nu_{\text{solv}}} + \frac{\Delta C_p^{\text{prot}}}{m_{\text{prot}}} \quad (33)$$

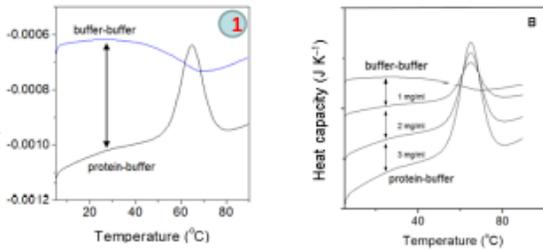
$$(34)$$

The **partial molar heat capacity** is defined as  $C_p^{\text{prot}} * \text{molar mass of the protein}$  with units  $JK^{-1}mol^{-1}$ .

The DSC experiment curve is shown below. Most of the unfolding happens at the peak. At this temperature there is approximately 50% of each state.

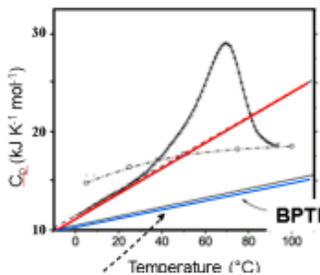


We eliminate the difference between the two cells by doing a blank experiment which gives a line. This line is subtracted from the curve, which yields a new curve.

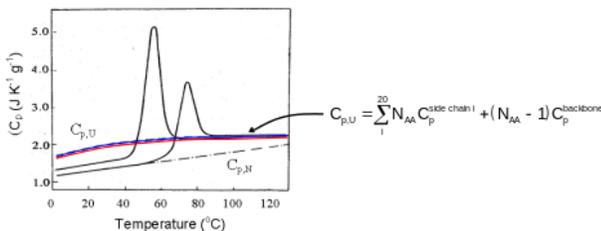


There are many other parameters that can be extracted from the DSC profile.

**The heat capacity of the native state** It can be found by the characteristic slope of the native region.



**The heat capacity of the unfolded state** The heat capacity of the unfolded state can be found by the slope of the unfolded region in the curve. It's found by summing over all the amino acids of the protein.



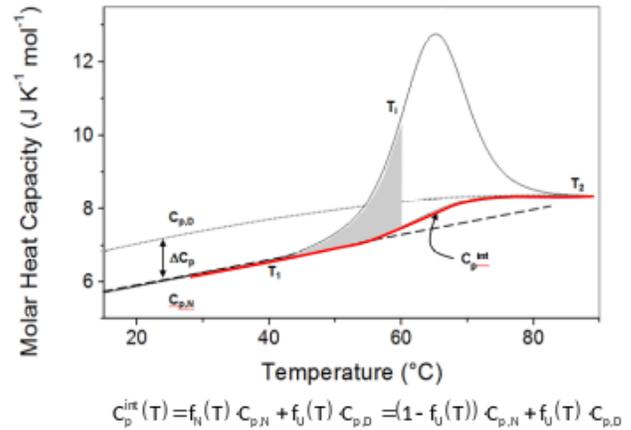
If there is aggregation the curve doesn't have a clear slope in the unfolded region.

**Unfolding heat capacity change** In reality the unfolding heat capacity change is temperature dependent. However, the dependence on temperature is so small it can be neglected.

$$\Delta C_p = C_{p,D} - C_{p,N} \quad (35)$$

**Enthalpy of unfolding** First the intrinsic heat capacity is determined, which is the red line in the curve below. The excess heat capacity is defined as:

$$C_p^{excess} = C_p - C_p^{intrinsic} \quad (36)$$



$$C_p^{int}(T) = f_N(T) C_{p,N} + f_U(T) C_{p,D} = (1 - f_U(T)) C_{p,N} + f_U(T) C_{p,D}$$

$$f_U(T) = 1 - f_N(T) = \frac{Q(T) - \int_{T_1}^{T_m} C_p^{excess}(T) dT}{Q_{tot} - \int_{T_1}^{T_m} C_p^{excess}(T) dT}$$

This gives us the melting temperature  $T_m$ , which is the temperature at which the excess heat capacity is maximal. We can also determine the calorimetric enthalpy change, which is:

$$\Delta H_u^{cal}(T_m) = \int_{T_1}^{T_2} C_p^{excess} dT \quad (37)$$

From this we can determine the free energy change at any temperature with:

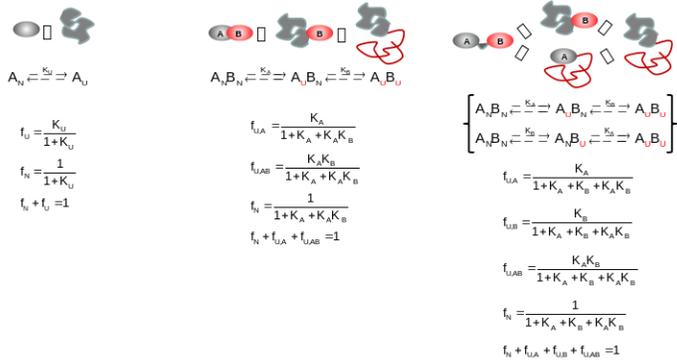
$$\Delta G(T) = \Delta H_U(T_m) \left(1 - \frac{T}{T_m}\right) + \Delta C_P \left[ T - T_m - T \ln \left( \frac{T}{T_m} \right) \right] \quad (38)$$

**Van't Hoff enthalpy change** Unlike the calorimetric enthalpy, the Van't Hoff enthalpy is model dependent (i.e. dependent of the amount of states that participate) and defined as

$$\Delta H_U^{vH}(T_m) = (2n + 2) RT_m^2 \frac{df_U}{dT} \quad (39)$$

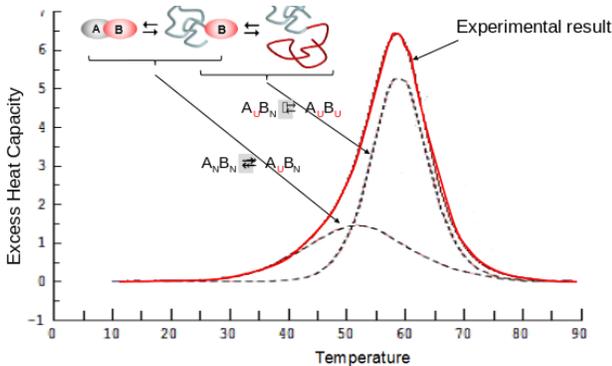
It's calculated by taking the slope around the melting temperature in the unfolding curve.

**Different models for unfolding** The model under which a protein unfolds needs to be known in order to find the equilibrium constant.



### Calorimetric versus Van't Hoff enthalpy

- $\frac{\Delta H_{vH}}{\Delta H_{cal}} = 1$  Indicates two state (un)folding (no intermediate).
- $\frac{\Delta H_{vH}}{\Delta H_{cal}} < 1$  Indicates significantly populated intermediates or that the protein concentration was underestimated, which can be an issue since calorimetric enthalpy depends on the protein concentration.



- $\frac{\Delta H_{vH}}{\Delta H_{cal}} > 1$  Indicates irreversible steps or that the protein concentration was overestimated.

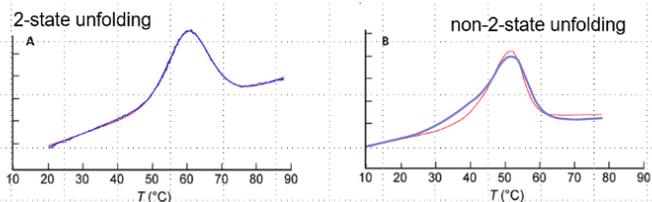
### Regression analysis of data

$$C_p = C_{p,N} + \frac{K_U(T)}{1+K_U(T)} \Delta C_p + \frac{\Delta H_U^2(T)}{RT^2} \frac{K_U(T)}{(1+K_U(T))^2}$$

Derivation in the Appendix to the

$$K_U(T) = \exp \left\{ -\frac{\Delta H_U(T_m)}{RT} \left( 1 - \frac{T}{T_m} \right) - \frac{\Delta C_p}{RT} \left[ T - T_m - T \ln \left( \frac{T}{T_m} \right) \right] \right\}$$

$$\Delta H_U^2(T) = \Delta H_U^2(T_m) + \Delta C_p (T - T_m)$$



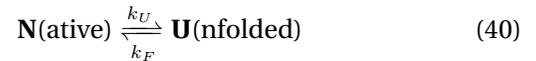
You should always do both spectroscopic and calorimetric experiments to determine if a protein can be assumed to follow the two state models.

## 9 Kinetics of Protein folding

Folding and unfolding of many small single domain proteins can be seen as an all or none process, without partly structured intermediates. The two state folding is modelled by a jump over a *free-energy* barrier of unstable conformations, at the top of which the *transition state* (TS) is represented.

### 9.1 One-step folding

Unfolding and refolding can be studied by rapidly transferring proteins into folding and unfolding conditions. There exists different varieties: rapid change of chemical denaturants (*stopped-flow method*), rapid change of temperature (*T-jump method*) or the rapid change of pressure (*P-jump method*). In the simplest case the folding/unfolding is a one-step reaction



with the equilibrium constant defined as:

$$K_U = \frac{[U]}{[N]} = \frac{k_U}{k_F} \quad \text{or} \quad K_F = \frac{[N]}{[U]} = \frac{k_F}{k_U} \quad (41)$$

with  $k_U$  and  $k_F$  are the microscopic rate constants for unfolding and refolding. The decay of  $N$  and  $U$  is given by

$$\frac{d[N]}{dt} = k_F[U] - k_U[N], \quad (42)$$

$$\frac{d[U]}{dt} = k_U[N] - k_F[U]. \quad (43)$$

The rate equation for the change of the folded state is therefore

$$\frac{d[N]}{dt} = -(k_F + k_U)[N] + k_F \cdot C \quad (44)$$

The term  $C$  is a constant given by  $C = [U] + [N]$ . The general solution is

$$[N]_t = A e^{-(k_U+k_F)t} + B = A e^{k_{obs}t} + B \quad (45)$$

with

$$A = [N]_0 - C \frac{k_F}{k_F + k_U}, \quad (46)$$

$$B = C \frac{k_F}{k_F + k_U}. \quad (47)$$

The concentration of  $U$  at time  $t$  is  $[U]_t = C - [N]_t$ . The concentrations of  $N$  and  $U$  change with time according to an exponential function with an observed rate constant  $k_{obs} = k_U + k_F$  often denoted with  $\lambda$ . The relaxation time is defined as  $\tau = 1/k_{obs}$ .

**Measuring [U] and [N]** is not possible in a direct way. However, changes of optical signals are proportional to the changes in concentrations. Let  $S_t$  be a function of the signal a time  $t$ .  $k_{obs}$  can be obtained from plots of  $\ln(S_t - S_0)$  versus time.  $S_t$  is given by

$$S_t = S_0 + A e^{-k_{obs}t} \quad (48)$$

$$S_t = S_0 + A[1 - e^{-k_{obs}t}]$$

$A$  is the amplitude and corresponds to the difference  $S_{eq} - S_0$  between the initial ( $t = 0$ ) and equilibrium ( $t \rightarrow \infty$ ) signals. The observation always yields  $k_{obs} = k_U + k_F$ .

- $[N] \gg [U]$ ,  $k_{obs} \approx k_F$  since  $k_F \gg k_U$
- $[U] \gg [N]$ ,  $k_{obs} \approx k_U$  since  $k_U \gg k_F$

Under strong folding or unfolding conditions the rate of the backwards reaction is neglected:

$$\frac{d[N]}{dt} = -k_F[U], \quad (49)$$

$$\frac{d[U]}{dt} = -k_U[N]. \quad (50)$$

The time dependent changes of the fractions of folded and unfolded protein are:

$$f_N(t) = (1 - f_U) = \left( \frac{[N]}{[N] + [U]} e^{-k_F t} \right), \quad (51)$$

$$f_U(t) = \left( \frac{[U]}{[N] + [U]} e^{-k_U t} \right). \quad (52)$$

Note that all above is valid for **monomeric** proteins only.

## 9.2 Chevron plot analysis

In many cases it is not possible to measure the rate of folding and unfolding under conditions where the protein is fully native or fully unfolded. This is often the case if the protein folds or unfolds very rapidly at the desired experimental conditions. The kinetics constants can be obtained by a series of experiments. The conditions are incrementally changed where the rate constants can be measured and then extrapolated for other conditions. This will obtain a reliable estimate for  $k_F$  and  $k_U$ .

The apparent kinetic constant at either condition is  $k_{obs} = k_F + k_U$ . This is equivalent to

$$\ln(k_{obs}) = \ln(k_F + k_U). \quad (53)$$

As often demonstrated,  $k_F$  and  $k_U$  depend linearly on the denaturing concentration  $[D]$  according to

$$\ln(k_F^d) = \ln(k_F^{H_2O}) - m_F[D] \quad (54)$$

$$\ln(k_U^d) = \ln(k_U^{H_2O}) + m_U[D] \quad (55)$$

where the constants indexed with  $D$  and  $H_2O$  refer to the reaction rates in denaturing and water (aqueous buffer). The multipliers  $m_f$ ,  $m_u$  describe the sensitivity of the refolding and unfolding rates on  $[D]$ :

$$m_F = \frac{d \ln(k_F)}{d[D]}, \quad m_U = \frac{d \ln(k_U)}{d[D]}. \quad (56)$$

Increasing  $[D]$  slows down refolding (minus sign) while speeding up unfolding (plus sign). Combining the above yields

$$\ln(k_{obs}) = \ln(k_F + k_U) \quad (57)$$

$$= \ln \left[ k_F^{H_2O} \exp(-m_F[D]) + k_U^{H_2O} \exp(+m_U[D]) \right] \quad (58)$$

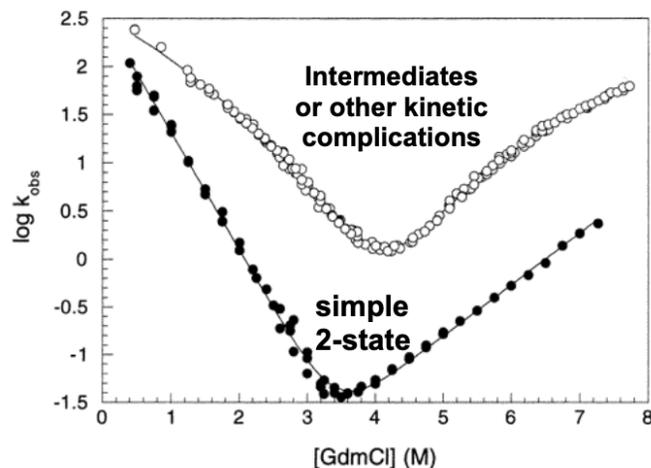
Plots of  $\ln(k_{obs})$  versus  $[D]$  are called **Chevron plots**. Given the data of an experiment  $k_F^{H_2O}$ ,  $k_U^{H_2O}$ ,  $m_F$ ,  $m_U$  can be obtained by least-square regression analysis using equation 57.

The parameters  $m_F$  and  $m_U$  provide information about the overall compactness and solvent accessibility of the transition state as it is thought that they reflect the surface exposure of the transition state ensemble. For a two state system:

$$RT(|m_F| + |m_U|) = m_{eq} \quad (59)$$

The units of  $m_F$ ,  $m_U$  are in  $M^{-1}$  while  $m_{eq}$  is in  $kJmol^{-1}M^{-1}$ . The ratio  $\beta_T = |1 - \frac{m_U}{m_U + m_F}|$  reflects the average compactness of the TS of unfolding relative to that of  $U$  from  $N$ . High  $\beta_T$  indicates that TS is similar to  $N$  in terms of compactness. In principle, the ratio  $|\frac{m_F}{m_F + m_U}|$  contains complementary information, since it characterizes the TS of refolding. In an ideal situation of 2SM folding/unfolding via identical TS for the refolding and unfolding reactions,  $\beta_U + \beta_F = 1$ .

The Chevron plot analysis is a powerful kinetic test for the 2SM as it tests for major intermediates. If intermediates are present,  $k_U$  and  $k_F$  are no longer linearly dependent on the denaturant concentration and the plot "rolls over" which means it is curved in either the folding or the refolding. This can be seen in figure below.



Intermediates can still exist even if there is no apparent roll over. For instance if the transition is very fast. The 2SM applies if:

- kinetic traces by different probes are mono-exponential,
- the same rate constants are derived from experiments using different probes,
- no kinks and roll-overs in the Chevron plot,
- $K_U^{eq} = \frac{f_U}{1-f_U}$  and  $K_U^{kin} = \frac{k_U}{k_F}$  are the same within error,
- $RT(|m_F| + |m_U|) = m_{eq}$  is fulfilled.

### 9.3 Transition state theory to protein folding

The *Arrhenius equation* applies to protein folding/unfolding, which can be formally treated as

$$k_{U,F} = A \cdot \exp\left(-\frac{E_A}{RT}\right) \quad (60)$$

$E_A$  is the activation energy of folding or unfolding and  $A$  is the reaction specific pre-exponential factor. In the framework of the transition state theory the equation recasts to:

$$k_{U,F} = \nu\kappa \exp\left(\frac{-\Delta H^\ddagger}{RT}\right) \exp\left(\frac{\Delta S^\ddagger}{R}\right) = \nu\kappa \exp\left(\frac{-\Delta G^\ddagger}{RT}\right) \quad (61)$$

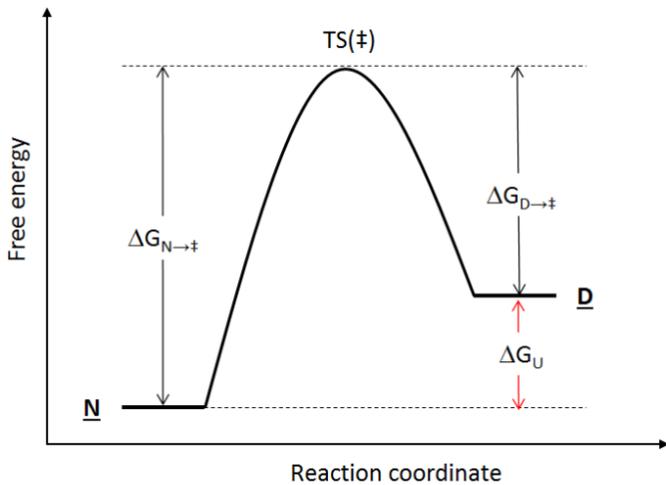
The factor  $\nu$  can be interpreted as the maximum possible rate of folding. In classical transition state theory,  $\nu = k_B/h$ , where  $k_B$ ,  $h$  are the Boltzmann constant and the Planck constant respectively.  $\kappa \leq 1$  is the transmission coefficient.  $\nu\kappa$  is difficult to estimate in protein folding.  $H^\ddagger$  is the activation enthalpy,  $\Delta S^\ddagger$  is the activation entropy and  $\Delta G^\ddagger = \Delta H^\ddagger - T \cdot \Delta S^\ddagger$  the activation energy. In unfolding *Eyring* plots of  $\ln(k_U/T)$  versus  $1/T$  are linear with a slope proportional to  $\Delta H^\ddagger$ . If the range of  $T$  can be shifted  $\Delta C_p^\ddagger$  can be estimated from

$$\Delta H^\ddagger(T_{med,2}) = \Delta H^\ddagger(T_{med,1}) - \Delta C_p^\ddagger(T_{med,2} - T_{med,1}) \quad (62)$$

with  $T_{med}$  being the median/mean temperature of each set of experiments.

The ratio  $\Delta C_p^\ddagger/\Delta C_{p,eq}$  should be a measure of the relative burial of hydrophobic surface in the transition state and can be compared with the  $\beta_T$  value (measures overall compactness of the TS). However, non Arrhenius behaviour is often observed in folding.

### 9.4 $\Phi$ value analysis



The saddle point of the graph above describes the energy of a TS as an energetic barrier separating the N and U states. Protein folding requires formation of high energy interactions which are proportional to  $\exp(-\Delta G_{U \rightarrow \ddagger}/RT)$  where

$\Delta G_{U \rightarrow \ddagger} = G_{\ddagger} - G_U$  is the free energy difference between the transition state and the ground state of the folding reaction. The same applies for the number of molecules in which critically important interactions are broken to pass the transition energy barrier that are proportional to  $\exp(-\Delta G_{N \rightarrow \ddagger}/RT)$  where  $\Delta G_{N \rightarrow \ddagger} = G_{\ddagger} - G_N$ .

The rate constants for the transition can be defined as

$$k_F = \nu\kappa \exp(-\Delta G_{U \rightarrow \ddagger}/RT) \quad (63)$$

$$k_U = \nu\kappa \exp(-\Delta G_{N \rightarrow \ddagger}/RT) \quad (64)$$

The free energy change of the overall folding is

$$\Delta G_U = \Delta G_{N \rightarrow \ddagger} - \Delta G_{U \rightarrow \ddagger} = G_U - G_N \quad (65)$$

$$= -RT \ln(k_U/k_F) = -RT \ln(K_U) \quad (66)$$

*Note: N is the reference state*

The value of  $\nu\kappa$  is not known and  $\Delta G_{N \rightarrow \ddagger}$  and  $\Delta G_{U \rightarrow \ddagger}$  cannot be measured but the value of  $\nu\kappa$  cancels out.

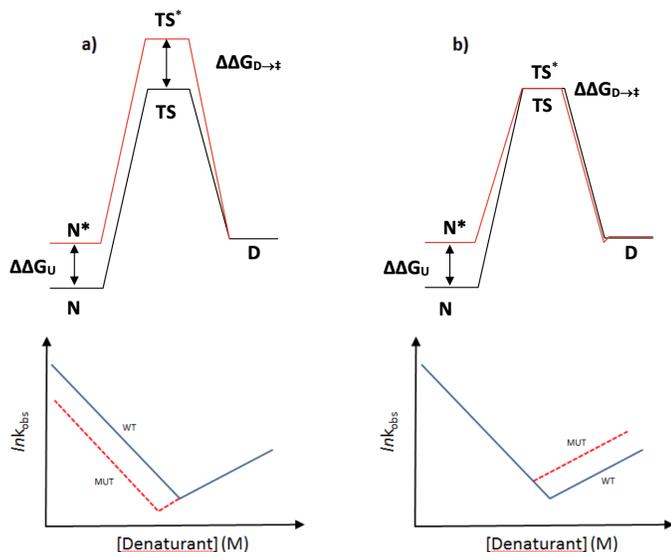
The *Leffler's* proportionality constants is defined as:

$$\alpha_x = \frac{\delta \Delta G^\ddagger / \delta X}{\delta \Delta G / \delta X} \quad (67)$$

In protein folding the perturbation can be introduced on structural level. If a mutation destabilizes the protein, both N and TS might be destabilized to the same extent. This is the case if the mutated residue is in a region that is already folded in the transition state. Alternatively, the mutated region is not yet folded and the TS will not be affected. For the activation reaction  $U \rightarrow \ddagger$ , the difference in the free energy between the mutant (denoted with \*) and wild type is defined as  $\Delta G_{U \rightarrow \ddagger}^* - \Delta G_{U \rightarrow \ddagger} = \Delta \Delta G_{U \rightarrow \ddagger}$ . The overall free energy difference between mutant and wild type is defined by  $\Delta G_U^* - \Delta G_U = \Delta \Delta G_U$ . The *Leffler's* proportionality constant, which in protein folding is usually called the  $\Phi$  value is defined as:

$$\alpha_x = \Phi_F = \frac{\Delta \Delta G_{U \rightarrow \ddagger}}{\Delta \Delta G_U} \quad (68)$$

$\Phi_F$  is the ratio of the folding activation free energy difference and the overall free energy difference between mutant and wild type protein. Note that here the ground state is the unfolded state, hence  $\Delta G_U = G_N - G_U$ . It follows that  $0 < \Phi_F < 1$ . If the mutation changes the stability of the TS as much as the native state  $\Delta \Delta G_{U \rightarrow \ddagger} = \Delta \Delta G_U \rightarrow \Phi_F = 1$ . In this case the site of the mutation is as much structured in the TS as in the native state, a) in figure below. On the other hand, if the mutation only affects the native state  $\Delta \Delta G_{U \rightarrow \ddagger} = 0 \rightarrow \Phi_F = 0$  which means the site of the mutation is not yet structured in the TS, b) in figure below.



**Experimental determination of the  $\Phi$  values** The ratio of the folding rate constants of the wild type and the mutant protein are related to  $\Delta\Delta G_{U \rightarrow \ddagger}$  by:

$$\Delta\Delta G_{U \rightarrow \ddagger} = RT \ln \left( \frac{k_F}{k_F^*} \right) \quad (69)$$

$$\Delta\Delta G_{N \rightarrow \ddagger} = RT \ln \left( \frac{k_U}{k_U^*} \right) \quad (70)$$

The rate constants can be obtained from the chevron plots.  $\Delta\Delta G_U$  is determined from isothermal denaturant-induced unfolding or thermal unfolding experiments under equilibrium conditions. If the limb of the chevron plot is not linear  $k_F$  and  $k_F^*$  are ambiguous, the  $\Phi$  values can be extracted from:

$$\Phi_U = \frac{\Delta\Delta G_{N \rightarrow \ddagger}}{\Delta\Delta G_U} \quad (71)$$

For a two state process it holds that  $\Phi_F = 1 - \Phi_U$ . All the parameters have to be collected at the same conditions. The assumption of the  $\Phi$  analysis is that the mutation causes localized packing defects without far reaching structural changes. The goal is to infer information of the approximate structure of the folding transition state. However, the results are not very clear as the values of  $\Phi$  are between 0 and 1 or exceed these theoretical limits. In these cases the interpretation is that the mutations affect the denatured state. The experimentally measured  $\Phi$  values can further be used in molecular dynamics simulations to obtain hypothetical models of the TS.

## 9.5 Folding rates of single domain two-state proteins

The folding rates within proteins differ to a great extent. There exists no correlation between the measured  $k_F$  and the overall structure. The following factors could possibly govern the folding rate:

1. the length of the polypeptide chain,
2. the sequence homology,

3. the overall (thermodynamic) stability,
4. the topology of the polypeptide chain in the native state.

## 9.6 Multi-step folding

Folding/unfolding can exhibit more than a single kinetic phase and can no longer be described by a single exponential function (equation 48). The definition of an intermediate is operational: under the given experimental conditions the time course of folding/unfolding is described by  $n$  exponentials indicating  $n$  intermediates. The observation of a signal change to the opposite sign is a clear indication of a multiple-step unfolding. Multi-step folding/unfolding can be formally described by:

$$S_t = S_{eq} + \sum_{i=1}^n A_i \exp(-k_{obs,i}t) \quad (72)$$

where  $S_t$  is the observed signal change,  $A_i$  the amplitude of phase  $n$  and  $n$  being the number of phases necessary to best describe the time course. Different experimental probes can lead to the observation of different time courses.

## 9.7 Types of folding pathways

Some proteins fold in an all-or-none manner, without detectable folding intermediates. However, it is clear that during the folding of most proteins, structural intermediates, containing stable secondary structure elements are formed rapidly. The role of these intermediates is not yet well established. Folding intermediates can either be on-path or off-path. An assumption is that the observed intermediates are critical in restricting the conformational space sampled by the polypeptide, thus allowing it to fold rapidly (on-pathway intermediates). In the simplest cases, the intermediates would form a direct linear pathway. Alternatively, the observed intermediates might assist folding but could be ordered in two or more distinct competing pathways. A third type of folding pathway is represented by transient accumulation of wrongly folded structures that are irrelevant or even detrimental to the formation of the final native state (off-pathway intermediates). They slow down folding, since time is needed for unfolding of such intermediates before the unfolded chain flow in the direct folding pathway. Slow folding of a globular small protein can indicate the occurrence of wrongly folded intermediates. Molecules may fall into a kinetic trap.

**Four models of protein folding** are conceptually established.

1. The *Framework model* considers the folding reaction as the sequential formation of native-like microdomains. The small secondary structure elements are formed locally during the initial stages of protein folding and come together by random diffusion and collision.
2. In the *Nucleation and nucleation-condensation model* a few key residues of the polypeptide chain form a local nucleus of secondary structure in the rate-limiting

step of folding. Around this nucleus, the whole native structure develops. The nucleation–condensation model in which a nucleus of local secondary structure has poor stability by itself, and its stabilization requires interactions between non-local residues and all the secondary structure and native-like tertiary contacts form in a concerted manner in a single rate-limiting step.

3. The *Hydrophobic collapse model* postulates that folding begins by an initial clustering of hydrophobic residue side chains which prefer to be excluded from an aqueous environment. The clustering of hydrophobic residues is expected to be non-specific and hence, to happen rapidly. The formation of an ensemble of collapsed structures would drastically reduce the available conformational search-space. Hydrophobic residues of the protein are clustered in the interiors of the collapsed forms. The formation of secondary structure and consolidation of specific tertiary contacts is promoted in these collapsed conformations with relatively fluid structures.

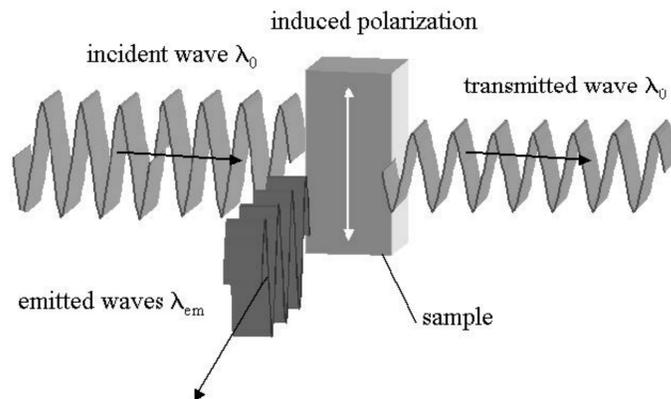
4. The *jigsaw model* postulates that each molecule folds by a different path, much as a jigsaw puzzle can be assembled in many different ways.

There is experimental evidence for each of the four models and none is universal. However, the hydrophobic collapse model and the jigsaw model seem less likely. The discussions so far have been embedded within the framework of the “classical view” about the folding mechanism in which the formation of a series of discrete intermediates along the reaction way from the denatured state to the native state. The “new view” describes protein folding in terms of statistical ensembles of states and focuses on the general features of folding on a complex multidimensional potential energy functional (energy landscape), the *funnel theory*.

## Part IV - Schuler

### 10 Classical Ensemble Spectroscopy

In typical spectroscopy experiments we measure many signals averaged across many molecules by irradiating a large volume inside a cuvette and measuring fluorescent output.



We are dealing with a sample volume of about 1 ml or 1 cm<sup>3</sup> and a protein concentration of 1 μM.

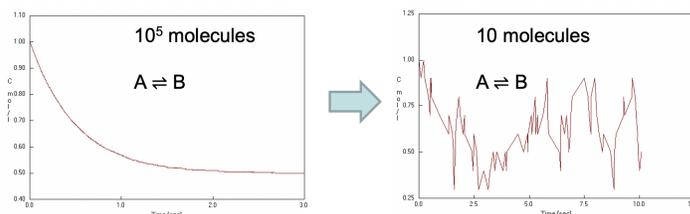
#### 10.1 Averaging and Heterogeneity

Measuring averages of a large number of molecules has multiple consequences.

- Distributions of molecular properties are averaged out → Information is lost.
- Molecules need to be synchronized for kinetic experiments.

These problems can be resolved if we setup single molecule experiments. In these experiments signals are recorded for every molecule.

- More heterogeneity can be resolved.
- Kinetics can be obtained from equilibrium measurements.



Reducing the number of molecules in a spectroscopy experiment results in the capture of more stochasticity.

#### 10.2 Single Channel Recording

The electric current across individual trans-membrane channels can be measured using patch-clamp methods (i.e. by

measuring the change in current by applying a micropipette to a single channel and measuring the voltage difference inside and outside the micropipette). This method can be used to quantify stochastic fluctuations between open and closed state.

## 11 Analysis of Single Molecule Kinetics

### 11.1 Ensemble and Single Molecule Kinetics



### 11.2 "Classical" Ensemble Kinetics

This is a continuous deterministic process. First we setup the reaction rate equations.

$$\begin{aligned} \frac{dN_A}{dt} &= -k_{AB}N_A + k_{BA}N_B \\ \frac{dN_B}{dt} &= -k_{BA}N_B + k_{AB}N_A \end{aligned}$$

Subsequently we find the solution of the differential equations using the initial conditions.  $N_A(0) = N$ ,  $N_B(0) = 0$

$$\begin{aligned} N_B(t) &= N \frac{k_{AB}}{k_{AB} + k_{BA}} (1 - e^{-(k_{AB} + k_{BA})t}) \\ N_A(t) &= N - N_B(t) \end{aligned}$$

### 11.3 Single Molecule Kinetics

We use **probabilities** instead of rates. We calculate the probability of being in state  $A$  or  $B$ .

$$p_A = \frac{N_A}{N} \quad p_B = \frac{N_B}{N}$$

Rewrite reaction rate equations in terms of probabilities.

$$\begin{aligned} \frac{dp_A}{dt} &= -k_{AB}p_A + k_{BA}p_B \\ \frac{dp_B}{dt} &= -k_{BA}p_B + k_{AB}p_A \end{aligned}$$

These are called the Master equations, solving for the time dependant probabilities yields.

$$p_A(t) = \frac{N_A(t)}{N} \quad p_B(t) = \frac{N_B(t)}{N}$$

### 11.4 Dwell Time

Time until a molecule in state  $A$  jumps to state  $B$ . This behaviour is equivalent to an irreversible reaction  $A \rightarrow B$ , thus we can neglect the backward flux.

$\frac{dp_A}{dt} = -k_{AB}p_A$  with  $p_A(0) = 1$  Thus the survival probability at time  $t$  is the following:

$$p_A(t) = e^{-k_{AB}t}$$

From there we compute the probability density at any given time  $t$ , which is equal to the reaction rate.

$$p_{AB}(t) = k_{AB}e^{-k_{AB}t}$$

The average dwell time is then computed as the integral from 0 to  $\infty$ .

$$\langle t \rangle = \int_0^{\infty} t \cdot p_{AB}(t) dt = \frac{1}{k_{AB}} = \tau_A$$

#### 11.4.1 Equilibrium

We have equilibrium when  $p_A$  and  $p_B$  do not change anymore. Thus we get the condition:  $\frac{dp_A}{dt} = \frac{dp_B}{dt} = 0$   
If we let  $t$  go to  $\infty$  we get for the fraction of  $p_A$  and  $p_B$  and consider single molecule kinetics to be a discontinuous and stochastic process we get the following:

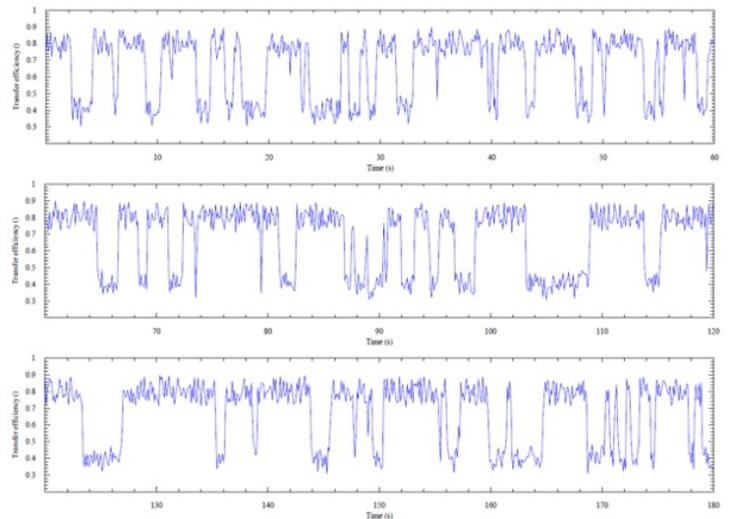
$$\frac{p_A(\infty)}{p_B(\infty)} = \frac{k_{BA}}{k_{AB}} = \frac{\tau_A}{\tau_B} = K_A$$

Which is equivalent to

$$\frac{p_U(\infty)}{p_F(\infty)} = \frac{k_U}{k_F} = \frac{\tau_U}{\tau_F} = K_U$$

### 11.5 Analysis of Single Molecule Trajectories

The following plot shows the FRET transfer efficiency simulations over time  $t$ . They describe a two-state process. Below, the process of folding-unfolding is displayed.



We measure the lifetimes of the folded and unfolded states. From these measurement we compute:

- the total number of events  $N$ : Count all folded/unfolded states.
- the total time in either of those states  $t_x$ : Sum of both state lifetimes.
- the probability of being in each state  $p_x$ : Divide lifetimes of either state by the total time
- the equilibrium constant  $K$  is given by the fraction of the dwell times ( $K_U = \tau_u/\tau_f$ ).
- the free energy difference  $\Delta G_U = -RT \ln K_U$
- the average lifetime  $\tau_x$ : is given by the total time in state  $x$  divided by the number of times the system is in state  $x$ .

- the rate coefficient  $k_{XY} = \frac{1}{\tau_X}$  (e.g.  $k_U = \frac{1}{\tau_F}$ )

## 12 Fluorescence Spectroscopy

Fluorescence is a process where light of a lower wavelength and thus higher energy level is absorbed, part of the energy is lost in an internal conversion process and the rest of the energy is released as light of a lower energy level.

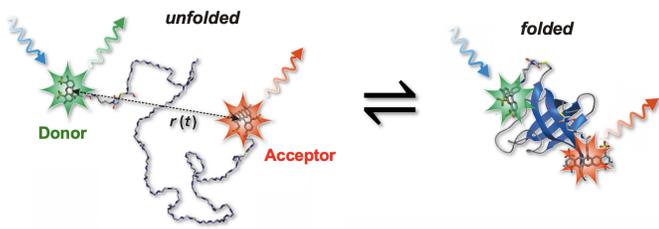
**Examples** Cyanine Dyes, Rhodamine Derivates or Fluorescent Proteins (e.g. GFP)

For single molecule experiments we have several requirements for our fluorophores:

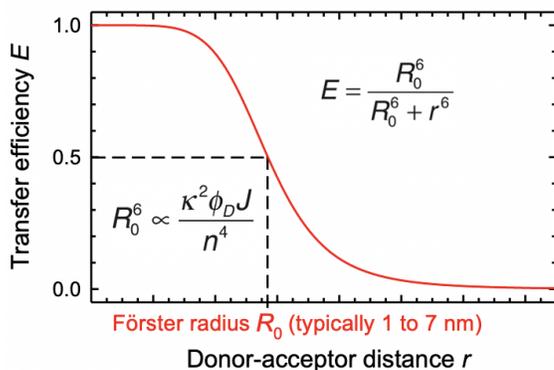
- high extinction coefficients
- high fluorescence quantum yields
- low triplet yield
- high photostability
- high solubility in water

### 12.1 Förster Resonance Energy Transfer

Methods used to measure if two components of proteins are within proximity. A donor fluorophore emits light and transfers energy to the acceptor by resonance. This process is highly distance dependant. In the following example the amount of green light is increased the higher the distance between donor and acceptor chromophore.



This allows us to quantify the distance between the two fluorophores. This relationship is displayed in the following plot.



For FRET experiments we can compute the spectral overlap which indicates the area of the donor fluorescence and the acceptor absorbance that overlap.

$$J = \int_0^{\infty} f_D(\lambda) \epsilon(\lambda) \lambda^4 d\lambda$$

We additionally have to consider that the resonance interaction strength depends on how well the fluorophores are aligned. In practice we use a  $\kappa^2$ -factor of  $2/3$  due to the consideration of an orientation average.

$$\kappa^2 = (\cos \omega_T - 3 \cos \omega_D \cos \omega_A)^2 = 2/3$$

We can then derive the Förster radius  $R_0$  (1-7 nm).

$$R_0^6 \propto \frac{\kappa^2 \omega_D J}{n^4}$$

### 12.2 Protein Labeling

We can label proteins for FRET by reduction using maleimide chemistry.

1. Reduction of disulfide bridges by DTT (Maleimide to Thioether)
2. Desalting of protein solution
3. Alexa 488 is added and binds to opened up HS-
4. Ion exchange chromatography is then used to select out marked proteins
5. The other chromophore (Alexa 594) is then added
6. In a last step the marked protein is again extracted using ion exchange chromatography

### 12.3 Dealing with Noise

One limiting difficulty of single molecule measurements is the large amount of noise from solvent molecules. These solvent molecules are in huge excess in comparison to the molecules of interest. The background mainly occurs due to the scattering effect of these solvent molecules.

#### Resolutions

1. Reduce the observation volume as the background is proportional to the number of illuminated molecules.  
→ **Spatial Selection**
2. Selection of a detection method with high selectivity for the molecule of interest.  
→ **Spectral Separation**

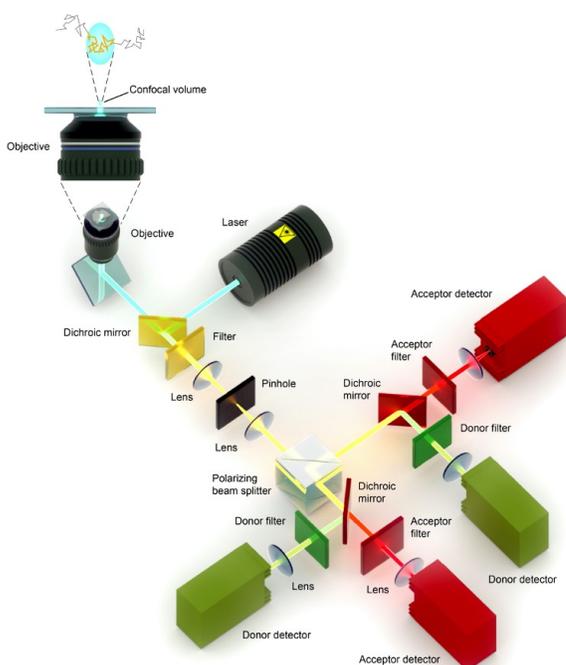
### 12.4 Total Internal Reflection Fluorescence

This is a method of microscopy where a phenomenon called total reflection is used in order probe into only a small area of the sample. This method has multiple advantages such as its simple implementation, the possibility of allowing to measure many molecules in parallel but it has only a moderate time resolution of 1 to 100 ms.

The good performance of this method is due to the spatial selection of the evanescent field and the spectral separation of the selective excitation and detection techniques.

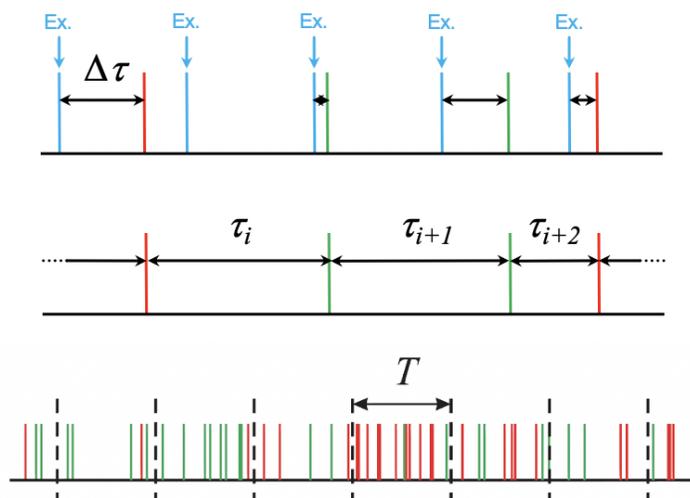
## 12.5 Confocal Single-Molecule Fluorescence Detection

Measuring single molecules by focusing a small laser beam on single molecules, the laser is targeted on different molecules one after another. This approach achieves spatial separation by focusing the laser (confocal observation volume:  $1 \text{ fl}$ , sample concentration:  $10 \text{ pM}$ ) and spectral separation using dichroic mirrors and filters. This method requires more sophisticated instrumentation compared to the TIRF method but allows for the measurement in a higher temporal resolution. Furthermore measurement can be taken on freely diffusing molecules or sequential measurements on immobilized molecules.



## 12.6 Basics of Photon Statistics

Photon statistics is the study of the statistical distributions produced in photon counting experiments. In these experiments, light incident on the photodetector generates photoelectrons and a counter registers electrical pulses generating a statistical distribution of photon counts.



The blue lines represent the exact time when a laser was fired in order to activate the fluorescent molecules. The time interval  $\Delta\tau$  represents the time it takes until a photon is emitted after activation.  $\tau_i + n$  represent the interphoton time, thus the time between the detection of two single photons. This time ranges from  $1 - 10 \mu\text{s}$  (i.e. only a fraction of pulses will yield a photon). The last plot shows all the arrival times of donor and acceptor photons. This gives us an indication about how many photons we detect in total during time  $T$ . We usually use a binning time  $T$  of around  $1 \text{ ms}$  and the count rate is usually around  $100 \text{ ms}^{-1}$ .

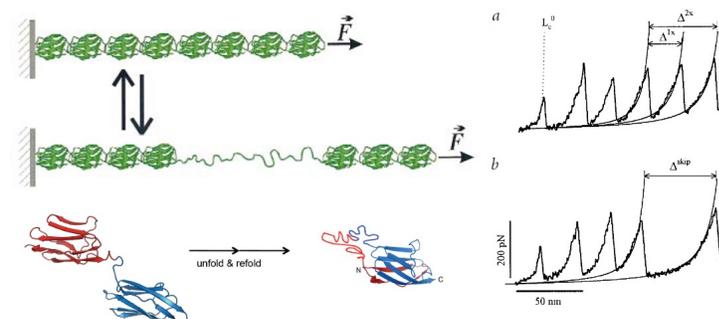
### 12.6.1 Photon Shot Noise

Shot noise is a type of noise which can be modeled using a Poisson process and occurs due to the particle like nature of light. We compute the probability of observing  $N_A$  acceptor photons in a fluorescence burst of  $N$  photons, given a fixed mean transfer efficiency  $\langle E \rangle$ .

$$P(N_A) = \binom{N}{N_A} \langle E \rangle^{N_A} (1 - \langle E \rangle)^{N - N_A}$$

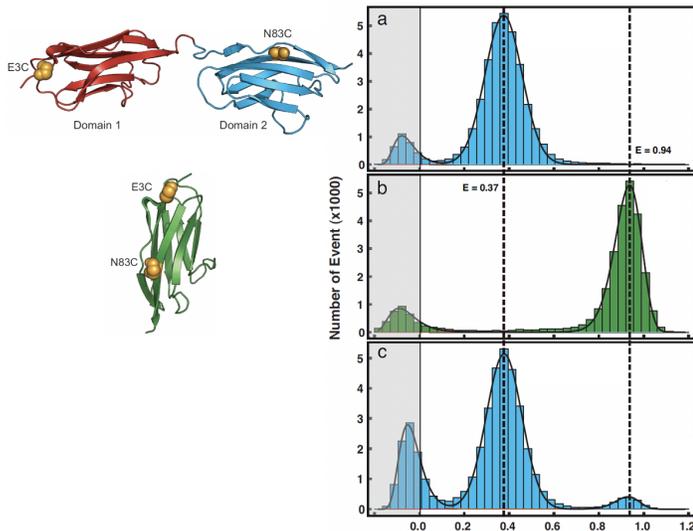
Shot noise leads to **broadening of transfer efficiency distributions**. Thus  $E$  distributions cannot be converted directly to distance distributions. But the underlying transfer efficiency distribution can be obtained by deconvolution of the shot noise contribution.

## 12.7 Revealing Misfolding in Multidomain Proteins



A force is applied to a chain of titins, this force is measured and recorded. We can see that titin complexes unfold one after another; this can be seen in the force-time diagram indicated by the tooth-like patterning. After the force is released the molecule refolds. If the unfolding-refolding cycle is repeated many times some of the edges will be missing upon unfolding, indicating missfolded titins.

The experiment is performed by attachment of a donor and receptor fluorophore to two neighbouring units of the titin chain.



The figure above shows a histogram of the transfer efficiencies between the fluorophores. We have the normal folded state in the top, the control state in the center (representing the missfolded state) and in the bottom section the experiment after many iterations showing that there is indeed a fraction of the molecules present in the missfolded state.

Results also indicate that sequence identity is an important factor in protein missfolding. It was shown that neighbouring segments with high sequence identity have a much higher chance of missfolding than low-identity neighbours as found in nature.

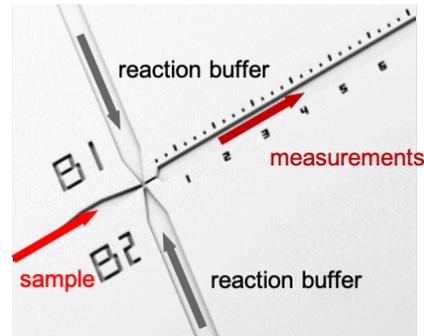
## 12.8 Moving Window Analysis

If we want to measure the kinetics of these folding-unfolding reactions then we need to track the molecules over time. This can be done by doing a moving window analysis where a measurement is taken approximately every 30 s. The donor and acceptor fluorophore intensity are measured at every time point. Using this approach it was found that the missfolded structure is thermodynamically less stable but kinetically stabilized. Thus missfolding is under kinetic control.

## 12.9 Microfluidic Mixing

Microfluidic mixing allows for the observation of the whole reaction process. The sample is mixed with the reaction buffers in a very small reaction chamber and is then passed through a small channel. We can observe different stages of the reaction by observing different locations within the chan-

nel. Due to the flexible nature of microfluidics the duration of this observation may be elongated by extending the length of the channel, even filters can be added to prevent inlet blocking.



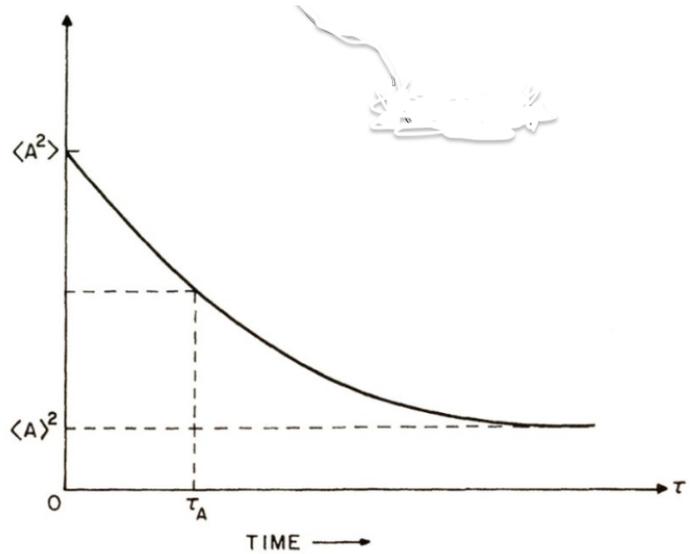
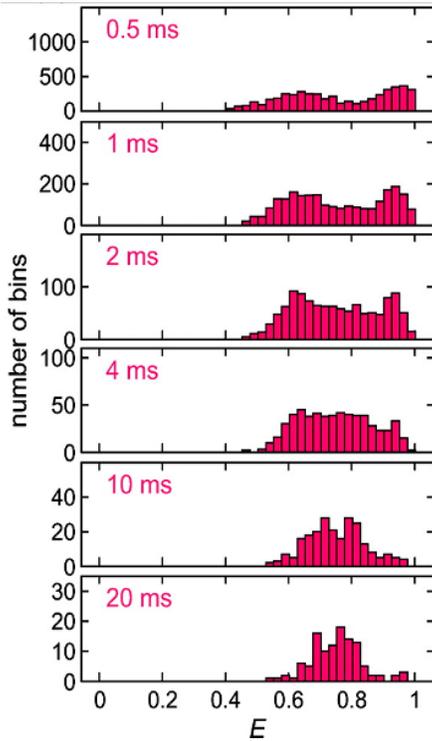
Using this technique we can find parameters such as the unfolding rate coefficient, by measuring the fraction of folded proteins at different timesteps after mixing and subsequently fitting a nonlinear function to the measurements.

## 12.10 Protein Kinetics at Equilibrium

A protein is fixed to a surface using biotin-streptavidin interactions. This protein binds to a binding partner, this interaction is measured using donor/acceptor fluorophores. We then follow the binding and unbinding events over time; this allows us to not only observe intramolecular changes such as conformational changes but also intermolecular interactions.

## 12.11 Line Shape Analysis

Analysis of transfer efficiency histograms can help detect and quantify kinetics of interconversion between populations based on the shape of transfer efficiency histograms. Slow interconversion results in a probability density histogram which shows two distinct peaks while very fast interconversion is characterized by one single peak. The following figure shows experimental results from  $\alpha_3D$ -folding experiments.



We can also normalize the auto correlation by dividing it by  $\langle A \rangle^2$ .

## 12.12 Correlation spectroscopy

### 12.12.1 Correlation

Calculate the signal intensity against itself and calculate the correlation coefficient:

$$R = \text{cor}(x, y) = \frac{\langle (x - \langle x \rangle)(y - \langle y \rangle) \rangle}{\sigma_x \times \sigma_y}$$

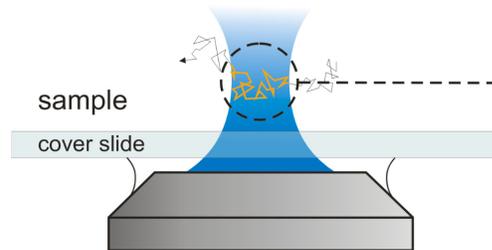
By calculating the correlation of a signal at time  $t$  and  $t + \delta t$ , we calculate how long-lasting the memory of the system is.

### 12.12.2 Auto-correlation

Auto-correlation can be calculated as follows: (T is the total time)

$$\langle A(t)A(t + \tau) \rangle = \frac{1}{T} \int_0^T A(t)A(t + \tau) dt$$

### 12.12.3 Fluorescence Correlation Spectroscopy (FCS)



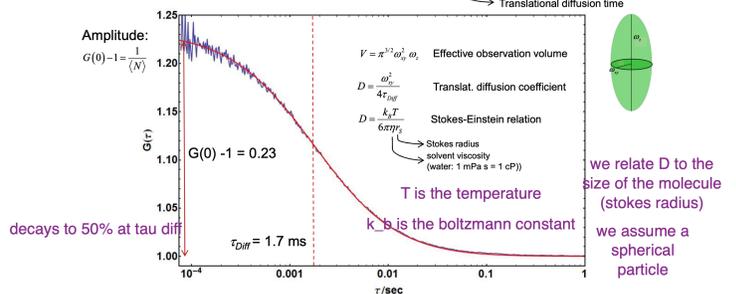
Concentrations of the fluorescent molecules typically in the  $nM$  to  $M$ . For the translational diffusion signal, we calculate the correlation function.

$\langle N \rangle$  is the average number of molecules in the confocal volume

#### FCS: Translational diffusion

we have to take into account the shape of the confocal volume

Correlation function for translational diffusion:  $G(\tau) = 1 + \frac{1}{\langle N \rangle} \frac{1}{1 + \tau/\tau_{diff}} \frac{1}{\sqrt{1 + (\tau/S)^2 \tau_{diff}^2}}$   $S = \frac{\omega_x}{\omega_y}$  Ratio of axial to lateral radii of the observation volume



Exercise: assume  $S = 5$ ,  $\omega_y = 350$  nm,  $T = 295$  K,  $\eta = 1$  mPa s  
→ calculate the Stokes radius and the concentration of the fluorescently labeled macromolecule

we have the volume and  $\langle N \rangle$ , so we can get the concentrations

## 12.13 Dynamics of biomolecular systems from FCS

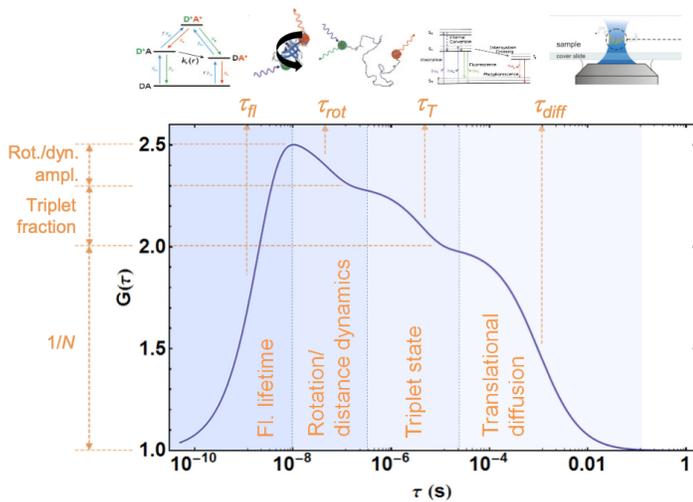
Any process that leads to fluctuations of fluorescence intensity on accessible time scale will contribute to the correlation function.

is around few **nano-meters**.

## 13 Force Spectroscopy I - D. Nettels

### 13.1 Atomic Force Spectroscopy

Atomic force spectroscopy (AFM) finds on the principle of partial unfolding of domains upon addition of force. An example for this is the sequential unfolding of immunoglobulin-like domains by AFM. This results in short spikes of force which drops again if the domain is unfolded until eventually the entire protein domain is unfolded. The set-up of an AFM experiment is described in figure 1.



Consider triplet-state as an example, if the intersystem crossing to the triplet state happens (electron is not going from  $s1$  or  $s2$  to  $s0$ ), there will be no light. This transition is in the order of  $10^{-6}$  and would be the transmission from the dark state to the emissive state. Bear in mind that the fluorescence life time, unlike the others, goes down in terms of correlation (if we look at smaller time-scales). If we want to look at the molecules on a time-scale greater than  $msec$ , we need to immobilize it.

We can look at the dynamics of conformational changes and interactions between the proteins in **nano-second** time-scale. For example, we can look at the interaction between ProT $\alpha$  and Histone H1, and figure out that they don't form a structure. Even NMR cannot explain this behaviour.

FCS could also be used in live cells for example for 2-D membrane protein diffusion. We can learn about the diffusivity of the molecules in the membrane and know about the membrane structure, for example extracting the lipid rafts.

We can also do single molecule in the time scale of hours, e.g. misfolding/oligomerization. (accessible time-scales slide 15/20, May 19th)

Other single-molecule fluorescence methods:

- In-cell single-molecule FRET: You would have good enough signal to detect the transfer efficiency. Shows that there are organelles in the cell which impede the freely-diffusing molecules. You can see if a protein which is unfolded outside of the cell, is folded inside the cell.
- Single-Molecule DNA Sequencing: SMRT technique. Uses the idea of having an evanescent field on top of the aluminium to avoid having more than one nucleotide fluorophore in the detection step. The nucleotides are in the scale of micro-molar in the solution
- Super-resolution Imaging: For an optical microscope, the resolution is around half of the wavelength of the light that is being used (around 200-300 nm resolution). If you have a molecule sitting at a position, you can determine its center much more accurately than its width. If we have sufficiently sparse distribution of molecules sitting on a surface, and we know that every one of the peaks comes from a single molecule, then we can determine the position of the single-molecule. The accuracy

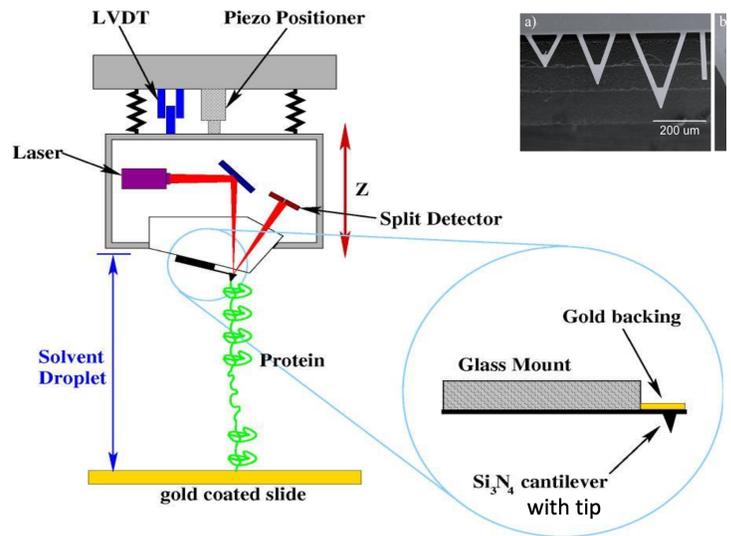
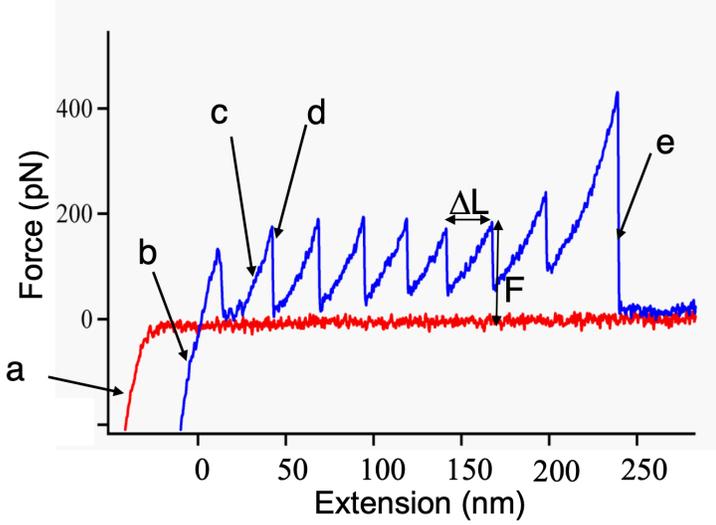


Figure 1: The sample protein is added to a gold coated slide on the one side and to a cantilever on the other. The cantilever is pulled by a piezo positioner and thereby the protein is unfolded. A laser is lead through a prism and records the changes in reflection in the sample.

In order to attach the protein to the anchor, cysteine residues are added and form thiol-gold interactions that function as an attachment. The other end that is attached to the cantilever does so by random picking of individual molecules at the surface. As this process is random, the length of the attached protein chain varies. The process is reversible and the unfolding forces range between 150 – 300 pN.

**Force extension curves** *Extension* is the distance between the two attachment points of a molecule (corresponding to the distance between surface and tip in AFM). This curve is visible in figure 13.1. In part a) the cantilever touches the surface without any force, constituting thereby the baseline. In b) the successful adhesion leads to a start of the actual experiment. In c) the protein is stretched and in d) one domain is unfolded. In e) the protein strand breaks so the force returns to the basal state. In this experiment, the protein was made of height domains.



We can understand the experiment in following ways:

- unfolding of domains makes the protein that is attached between cantilever and surface larger.
- the elasticity is mainly due to **entropy** - with no force we have a lot of conformation, whereas with stretching the number of conformations is reduced.

### 13.2 Entropic elasticity of ideal polymers

In experiments with DNA, as expected, the force rises steeply when the DNA becomes fully stretched. The data from the force-extension curves was fitted with a freely jointed chain (FJC) model which fit the data poorly, meaning that e.g. the onset of the rise is at lower forces than predicted by the fits and that at very small extensions that rise seems to be linearly. This fits better with a model called the worm-like chain model.

**Freely Jointed Chain** is a framework used to describe the behaviour of a polymer. Here, the polymer is modeled as a chain of  $N$  rigid vectors of length  $b$ . These are like chemical bonds between monomers and we assume no interaction between monomers far apart in the sequence. This is called an ideal chain. In the absence of force ( $F = 0$ ) the orientation of the vectors is random. With  $R$  we denote the end to end vector which is the averaged x-component so the averaged extension  $R_x$ . This leads to:

$$\begin{aligned} \langle R_x \rangle &= \left\langle \sum_{i=1}^N r_{x,i} \right\rangle = \sum_{i=1}^N \langle r_{x,i} \rangle = \sum_{i=1}^N \langle b \cos(\Theta_i) \rangle \quad (73) \\ &= b \sum_{i=1}^N \langle \cos(\Theta_i) \rangle = 0 \end{aligned}$$

That means on average the extension of a freely jointed chain is equal to zero. So simulations of FJC can be performed to find out the distribution and a gaussian with mean  $\mu = 0$  and standard deviation  $\sigma = 2.99 \approx 3$ . Like this we can calculate

the probability of a 1D gaussian distribution  $P_{1DG}$ .

$$P_{1DG}(R_x) = \left( \frac{1}{2\pi\sigma_x^2} \right)^{\frac{1}{2}} \exp \left( -\frac{R_x^2}{2\sigma_x^2} \right) \quad (74)$$

The 3D gaussian distribution should yield us the probability of the FJC chain  $P_{FJC}(R)$ .

$$\begin{aligned} P_{FJC}(R) &\stackrel{?}{=} P_{3DG}(R) = P_{1DG}(R_x)P_{1DG}(R_y)P_{1DG}(R_z) \\ &= \left( \frac{1}{2\pi\sigma_x^2} \right)^{\frac{3}{2}} \exp \left( -\frac{R_x^2 + R_y^2 + R_z^2}{2\sigma_x^2} \right) \quad (75) \end{aligned}$$

Now the question is whether we are able to calculate the variance of the FJC. (We assume that  $\sigma_x^2 = \sigma_y^2 = \sigma_z^2$ )

$$\begin{aligned} \sigma_x^2 &= \langle (R_x - \langle R_x \rangle)^2 \rangle = \langle R_x^2 \rangle = \langle R_x^2 + R_y^2 + R_z^2 \rangle / 3 \\ &= \langle R^2 \rangle / 3 \quad (76) \end{aligned}$$

We can now calculate the notion  $\langle R^2 \rangle$  in order to get the real value of the variance

$$\begin{aligned} \langle R^2 \rangle &= \left\langle \left( \sum_{i=1}^N r_i \right) \cdot \left( \sum_{i=1}^N r_i \right) \right\rangle = \sum_{i=1}^N \sum_{j=1}^N \langle r_i \cdot r_j \rangle \\ &= b^2 \sum_{i=1}^N \sum_{j=1}^N \langle \cos(\Theta_{i,j}) \rangle \quad (77) \end{aligned}$$

For a FJC:

$$\langle \cos(\Theta_{i,j}) \rangle = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases} \quad (78)$$

So we get for  $F = 0$ :

$$\langle R \rangle = 0 \quad (79)$$

$$\langle R^2 \rangle = Nb^2 \quad (80)$$

With the definition of the variance  $\sigma_x^2$  from above we get the following formulation of  $P_{FJC}(R)$

$$P_{FJC}(R) = \left( \frac{3}{2\pi\sigma_x^2} \right)^{\frac{3}{2}} \exp \left( -\frac{3(R_x^2 + R_y^2 + R_z^2)}{Nb^2} \right) \quad (81)$$

We can now look at this process in terms of thermodynamics namely the Helmholtz free energy:

$$G(R) = U(R) - T \cdot S(R) = -k_b T \ln \Omega(R) \quad (82)$$

$$P(R) = \Omega(R) / \int \Omega(R) dV \quad (83)$$

Rewriting this and plugging in from formula 81:

$$G(R) = -k_b T \ln P(R) - k_b T \ln \left( \int \Omega(R) dV \right) \quad (84)$$

$$= -k_b T \ln \left( \frac{3}{\sqrt{2\pi Nb^2}} \right) + k_b T \frac{3(R_x^2 + R_y^2 + R_z^2)}{2Nb^2} \quad (85)$$

$$-k_b T \ln \left( \int \Omega(R) dV \right) \quad (86)$$

We expect maximal entropy for  $R = 0$  and calculate now  $F = \frac{\partial G(R)}{\partial R_x}$

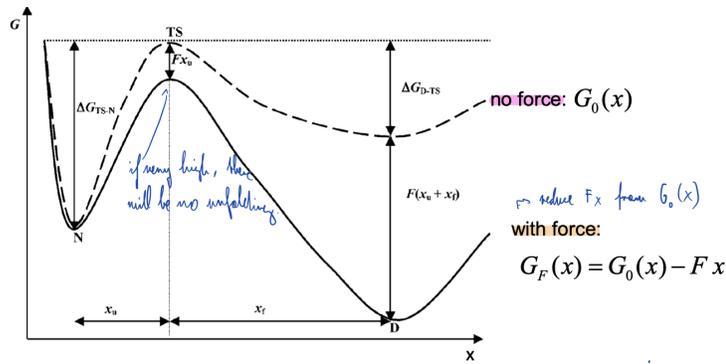
$$F = \frac{\partial G(R)}{\partial R_x} = \frac{3k_b T}{Nb^2} R_x \quad (87)$$

This now is a central result showing that for  $R_x \ll L$  the polymer is linearly elastic or Hookean as  $P_{FJC}(R) \approx P_{3DG}(R)$  so that  $F \propto R_x$ .

If we are however way beyond this regime so that  $R_x \ll L$  doesn't hold anymore, FJC is not a good approximation and we need to use the worm-like model.

### 13.3 Force induced Unfolding

When applying force, the free energy changes  $G_F(x) = G_0(x) - F(x)$



This leads to the following unfolding rate:

$$k_u(F) = \kappa \nu \exp\left(\frac{-(\Delta G_{TS-N} - Fx_u)}{k_b T}\right) \quad (88)$$

$$= k_u^0 \exp\left(\frac{Fx_u}{k_b T}\right)$$

### 13.4 Force Spectroscopy with Optical Tweezers

Very small particles with a high refractive index are attracted to intense regions of a tightly focused laser beam and can be trapped permanently slightly above the focal point. The restoring force of small displacements follow Hook's law - are trapped in a harmonic potential  $F_{trap} = -k_{trap} \cdot x$ .

Displacements from the trap center leads to a deflection of the focused laser beam which can be detected by a position detector. In order to do that, the apparatus needs to be carefully position and force calibrated. The trap can be moved by a beam steering unit and the bead position is recorded by a CCD camera

In order to understand the different situation we have to distinguish three different regimes.

- $d \ll \lambda$  : Rayleigh regime
- $d \gg \lambda$  : Ray optics
- $d \approx \lambda$  : Lorentz-Mie regime

First we will consider the Rayleigh regime. Here the object is a point object and the trapping light induces a dipole moment  $\mu(r, t) = \alpha E(r, t)$ . Like this we get the Energy.

$$U(r) = -\frac{1}{2} \mu \cdot E = -\frac{1}{2} \alpha |E|^2$$

Taking the gradient of this gives us the force:

$$F = -\nabla \cdot U(r) = \frac{1}{2} \nabla \cdot E^2$$

The gradient force is then (wherever that comes from)

$$F_{grad}(r) = \frac{2\pi n_0 a^3}{c} \left(\frac{m^2 - 1}{m^2 + 2}\right) \nabla I(r) \quad (89)$$

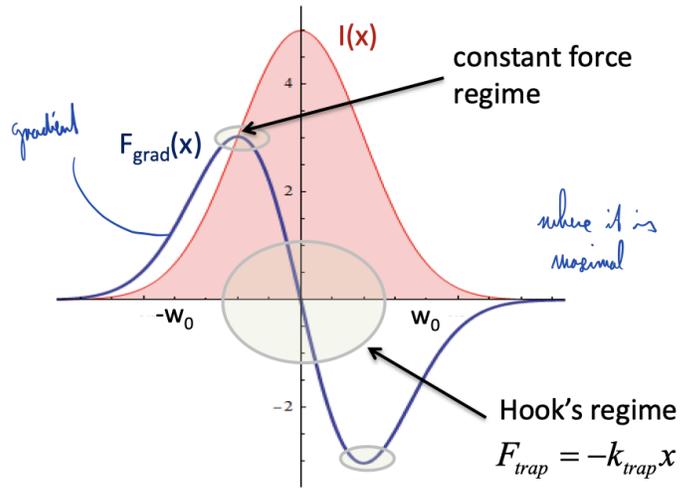
Next to the gradient force the scattering of the photons induces a momentum  $p = \hbar k$  with  $|k| = 2\pi/\lambda$ . This gives a scattering force which pushes the particle in direction of beam propagation

$$F_{scatter}(r) = \frac{8\pi n_0 k^4 a^6}{c} \left(\frac{m^2 - 1}{m^2 + 2}\right) I(r) \hat{z} \quad (90)$$

With these two forces we can describe the force that makes up the trap

$$F_{trap}(r) = F_{gradient}(r) + F_{scatter}(r)$$

At the very center the  $F_{gradient}$  equals to zero, whereas  $F_{scatter}$  is maximal. Therefore the trapping position will be slightly above the focal point since the two forces need to equal out and  $F_{grad}$  is maximal slightly away from the center.



The intensity  $I(x)$  and its gradient  $\nabla I \propto F_{grad}(x)$  can be observed in figure 13.4. We see that at the center,  $I(0)$  is maximum whereas  $\nabla I(0) = 0$ .

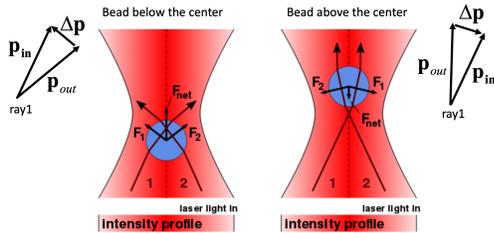
Next we will focus on ray optics and how light can trap an object. Incoming and outgoing photons have two momenta  $p_{in/pout}$  with identical magnitude.

$$|p_{in}| = |p_{out}| = h/\lambda$$

The difference of these two vectors  $\Delta p$  leads a a formulation of the Force on the bead

$$F = N\Delta p$$

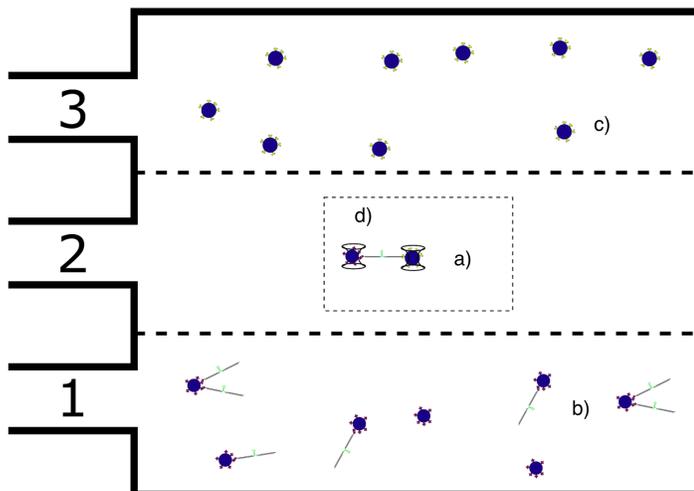
In a not focused laser beam (collimated beam) the restoring force pulls the bead back to the center and it is also pushed in beam direction. The latter thing doesn't happen in a focused beam as there we can trap as well in z-direction. Forces are balanced above the focal center.



## 14 Force Spectroscopy II

**Protein folding** can be studied using optical tweezers. Attached to the tweezers are two DNA handles that function as a linker between the beads and the protein of interest in the middle. The beads can then be moved independently via the optical traps and force extension curves can be recorded. In the case of **DNA** one can do the same thing by having a biotin handle on one side and a Dig handle on the other side. These can attach to their corresponding partners on the beads, streptavidin anti-Dig. The workflow for the PCR amplification is described in the slides.

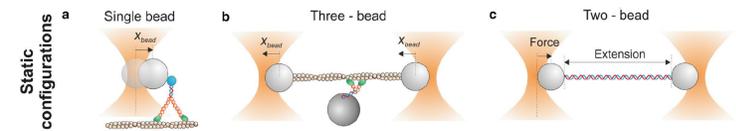
**Construct assembly** There is a microfluidic device with three laminar flows. In the first flow is the first bead with the entire construct attached. Firstly, the two traps are empty in the middle flow (a). Now one moves the beads into the first flow (b). If one trap has the construct attached, it is moved to the third stream where the second bead is located (c) and then finally one moves both traps with the attached construct to the middle flow (d).



Measuring hairpin folding/unfolding is possible with this technique by moving the beads apart. A hairpin unfolds usually at  $\approx 12 pN$ . If the folding and unfolding jumps occur at the exact same extension, the dynamics are said to be very fast. With a higher resolution one recognises individual folding/unfolding events. The dynamics are different for different trap-distances, meaning that at a certain extension you can have nearly only folded hairpins and at another distance only unfolded ones.

With these results one can calculate the average dwell time and determine whether the protein or DNA strand is a two-state folder or not.

### Examples



In a) the experiment is shown where kinesin is attached to a bead and moves along the microtubule. The force increases with the displacement until the motor stalls. Like this the force of an individual motor protein can be measured. In b) myosin is fixed by the traps and interacts with actins to move some  $nm$ . In c) is a set-up that allows for the study of mechanical properties of dsDNA.

Another approach is feedback systems that move along as e.g. a kinesin moves a long a microtubule. Like this the trap displacements measure the protein displacement. This works by keeping the force on the bead constant.

In a third type of experiments where the packing of DNA into  $\phi 9$  phage heads was measured. It was shown that the packaging occurred in steps of 10 bp with occasional dwell times in between that were dependent on the concentration of ATP.

### 14.1 Comparison of both Methods

	Optical tweezers	AFM
Spatial resolution (nm)	0.1–2	0.5–1
Temporal resolution (s)	$10^{-4}$	$10^{-3}$
Stiffness ( $pN\ nm^{-1}$ )	0.005–1	$10-10^5$
Force range (pN)	0.1–100	$10-10^4$
Displacement range (nm)	$0.1-10^5$	$0.5-10^4$
Probe size ( $\mu m$ )	0.25–5	100–250
Typical applications	3D manipulation	High-force pulling
Features	Low-noise and low-drift dumbbell geometry	High-resolution imaging
Limitations	Photodamage Sample heating	Large high-stiffness probe Large minimal force