Quantitative description of co-assembly and evolution of small heat-shock proteins

Dominik Saman

Balliol College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

September 2021

This thesis is dedicated to my family.

Acknowledgements

First, I would like to thank Professor Justin Benesch and Professor Carol Robinson for giving me the opportunity to do my DPhil with them, finding funding for me, and guiding me through the past four years. Both Carol and Justin have been extremely kind and generous to me and made sure I have always got the support I needed. Further, it was with Justin's guidance that I discovered the exciting world of small heat-shock proteins and their biophysics and evolution–an area of research I find both exciting and satisfying. I would also like to thank Justin for his invaluable feedback on this thesis.

I started my DPhil at the same time as (now Dr) Fabian Soltermann and (soon-to-be Dr) Zihao Wang whom I, throughout the years, became very good friends with, and who made my time here fun. Throughout the years, I have made very good friends with people Justin brought into the group–new DPhil students, Part IIs, and summer students–Angela, Christina, Ellie, Guodong, Josh, Luke, Marco, and Wiktoria–many of whom I have grown closer together over the years, and who I count as my closest friends. As the biophysics research groups within the Chemistry Department are a bit special in the way we all socialise together, I have made multiple friends outside the Benesch group, too–Anna, Daniel, Denis, Francesco, Siyun, and Virginia–among many others.

I would also like to thank the multiple people, many of whom I consider good friends, but who were also essential for the progress of this work–Prof. Andy Baldwin, Dr. Andy Dolan, Dr. Georg Hochberg, Dr. Jani Bolla, Dr. Mark Agasid, Dr. Miranda Collier, Dr. Vicky Cullen, Dr. Ted Mize, and Dr. Shane Chandler.

Outside of the lab, countless people have always injected excitement into my life and were there to support me when I most needed them. Thank you, Alex, André, Denisa, Florian, Greg, Kamila, Konrad, and many others...

Finally, I would like to thank, and dedicate this thesis, to my family, who were always there for me. Who supported me throughout my (embarrassingly rebellious) teenage years; who were there for me when I decided to come to the UK as a naive 17-year-old with the dream of studying here; who gave me the financial support I needed to live comfortably here.

Thank you mum for not only being my mum but also my best friend when I needed one. My life would not be complete without two Jaroslavs–one who I will always think about as dad even though he didn't have to be.

The second Jaroslav, my brother, makes me feel every day like someone's hero. You might not realise it now, but your curiosity and enthusiasm fill me with happiness every day.

Last but not least, I would like to thank my grandma and my cousin Klára for their support and friendship.

Thank you, everyone. Děkuji.

Abstract

Small heat-shock proteins (sHsps) are a family of molecular chaperones that are a part of the cellular proteostasis machinery. Many of them, such as human HspB1, HspB4 and HspB5 are polydisperse–they form a large range of stoichiometries. Some are also known to co-assemble, forming a large variety of sub-stoichiometries, with possible implications for their function.

In this work, we first develop theoretical models for studying the co-assembly between both monodisperse as well as polydisperse proteins using statistical mechanics and chemical kinetics, respectively. We then briefly apply the statistical-mechanical model on a system of two co-assembling monodisperse sHsps.

We then use two orthogonal biophysical techniques, native mass spectrometry (MS) and mass photometry (MP), to study the homo-oligomerisation of the two most abundant, and highly polydisperse, human sHsps–HspB1 and HspB5. Next, we use MS and MP to study highly heterogeneous mixtures of these two proteins on both the level of stoichiometries as well as the sub-stoichiometry level.

We next use the model we have developed and apply it to our mixing data to obtain both the complete description of these co-assemblies as well as to extract any energy biases in these mixtures, which we then probe experimentally using MS, MP and molecular dynamics modelling.

Finally, we study the co-assembly of HspB1, HspB5 and HspB6 between different vertebrate species and mention that some HspB1 and hspB6 orthologs either completely lost, or are in the process of losing, their polydispersity. Curiously, HspB5 orthologs still seem to be polydisperse, which might be due to the functional role of polydispersity in the eye lens, where HspB5 is highly expressed.

Statement of Originality

This thesis covers most of my work done in the Department of Chemistry at the University of Oxford between October 2017 and July 2021. The individual chapters are tied to the central theme of quaternary structure polydispersity.

I have been helped by numerous people throughout my years at the department, but the thesis itself is a product of my own work, unless stated otherwise in the Acknowledgement section of each chapter. Again, as I have been helped by numerous people, either through indirect help or directly through them making their data available to me, and I believe that research is fundamentally collaborative, I use the 'we' pronoun in this work.

Figures can be assumed to contain data or schematics produced by me, unless stated otherwise either directly in the caption or in the Acknowledgement section.

This work has not been submitted for a degree, diploma, or any other academic certification at any other institution, in whole or in part.

Dominik Saman
September 2021

Contents

List of Figures

List of Tables

List of Abbreviations

3D 3-dimensional

AA(s) . . . . . amino acid(s)

aCD . . . . . alpha-crystallin domain

APTES . . . . (3-Aminopropyl)triethoxysilane

BME . . . . . β-mercaptoethanol

CD . . . . . circular dichroism

c.i. . . . . . confidence interval

CID . . . . . collision-induced dissociation

CTD . . . . . C-terminal domain

Ctp . . . . . C-terminal peptide

DTT . . . . . dithiothreitol

FWHM . . . . full width at half maximum

GHCl . . . . . guanidine hydrochloride

HCD . . . . . high-energy collision dissociation

IEX . . . . . ion-exchange chromatography

IM (MS) . . . ion-mobility (mass spectrometry)

IPTG . . . . . Isopropyl β-D-1-thiogalactopyranoside

iSCAT . . . . . interferometric scattering

KDE . . . . . kernel density estimation

LB . . . . . Lauria broth

LCA . . . . . last common ancestor

OD . . . . . optical density

MCMC . . . . Markov-chain Monte Carlo

MD . . . . . molecular dynamics

MP . . . . . mass photometry

MS . . . . . mass spectrometry

MS/MS . . . . tandem mass spectrometry

NTD . . . . . N-terminal domain

ODE(s) . . . . ordinary differential equation(s)

PBS . . . . . phosphate-buffered saline

RCF . . . . . relative centrifugal force

SEC . . . . . size-exclusion chromatography

sHsp(s) . . . . small Heat-shock protein(s)

TMAO . . . . trimethylamine N-oxide

UV . . . . . ultraviolet

w.r.t . . . . . with respect to

1

Introduction

Contents

1.1 Introduction

The work presented in this thesis focuses on a quantitative description of co-assembly and evolution of small heat-shock proteins (sHsps). In this chapter, we introduce the family of sHsps, and discuss their quaternary structure as well as evolution. We then move onto the experimental techniques that we use in this work, and finally introduce the two theoretical frameworks we use for the description of sHsp co-assembly: statistical mechanics and chemical kinetics.

1.2 Small heat-shock proteins

sHsps are a family of molecular chaperones found in almost all living organisms, from bacteria to mammals ([1]), with a number of paralogs ranging from two in *Saccharomyces cerevisiae* ([2]) to tens in plants ([3]). Plants have evolved a robust network of sHsps as a part of their protein homeostasis machinery, possibly due to their immobility, which necessitates that they withstand a wide range of temperatures and other stressors. In humans, there are ten known sHsp paralogs with different expression levels in different tissues and varying substrate specificity ([4]) (Table 1.1). The known sHsps have a highly variable quaternary structure with some only forming monomers, while a host of others are known to form well-defined higher oligomers, and many others still form a wide range of oligomeric sizes.

Many sHsps also interact with each other and form co-assemblies, both *in vitro* ([5], [6]) and *in vivo* ([7]), with implications to chaperoning capacity.

1.2.1 The structure of small heat-shock proteins

General protein structure

The structure of proteins can broadly be classified into four separate levels of organisation. The so-called primary structure is simply a linear sequence of amino

Name	Also known as	MW (kDa)	Oligomeric distribution	Promiscuous chaperone?
HspB1	Hsp27	22.8	Polydisperse	Yes
HspB2	MKBP	20.2	Small oligomers	No
HspB3	HspL27	17.0	Dimers or trimers	No
HspB4	alphaA-crystallin	19.9	Polydisperse	Yes
HspB5	alphaB-crystallin	20.1	Polydisperse	Yes
HspB6	Hsp20	17.1	Dimers	No
HspB7	cvHsp	18.6	Monomers	No
HspB8	Hsp22	21.6	Monomers/dimers	No
HspB9	CT51	17.5	Not known	No
HspB10	ODF1	28.4	Not known	No

Table 1.1: The ten human sHsp with their monomer masses, modes of oligomerisation, and their activity as 'promiscuous' chaperones, defined as having universal activity to all substrates in Mymrikov's work ([4]).

acids (AAs) that form the backbone of peptides or proteins. The primary structure of proteins can be obtained either by, for example, sequencing the protein, or by inferring it from a region of DNA that codes for this protein. Even though the sequence does not provide any information about the 3-dimensional (3D) arrangement of these AAs per se, recent advances in machine learning have made it possible to predict the 3D structures of proteins with high confidence only based on their primary structure ([8], [9]).

The higher level of organisation of these AAs into local structural elements is called the secondary structure. The most common secondary structure elements are alpha helices, beta sheets (such as in Fig. 1.1), and turns ([10], Chapter 1). The secondary structure content of proteins can be inferred from the primary structure to some extent as some AAs are known to preferentially form alpha helices or beta sheets. Experimentally, the secondary structure content of proteins can be easily estimated using circular dichroism (CD) (discussed in more detail further in this Chapter, Section 1.6.1).

The 3D shape of the protein forms the basis of the tertiary structural level. In

general, the tertiary structure consists of interlinked secondary structure elements. Experimentally, the tertiary structure can be studied using for example X-ray crystallography ([11]), cryo-EM ([12]) or NMR ([13]). On this level of organisation, we can also find different domains within the protein. Formally, protein domains are defined as conserved parts of protein sequence or shape that can evolve and fold independently of each other ([14]).

Multiple protein subunits can also assemble into higher structures that form the functional basis of these proteins. This high-level organisation is called the quaternary structure. These assemblies range from simple dimers consisting of two subunits up to many tens or hundreds of subunits in viral capsids ([15]). Most known proteins indeed form oligomers ([16]), with many forming homo-oligomers, consisting of two or more identical protein subunits ([17], [18]). Similarly, many proteins also co-assemble to form hetero-oligomers ([19], [20]). The proposed role of oligomerisation ranges from functional gain ([21]) through allosteric regulation ([22]) to genetic economy ([23]), among others.

Interestingly, the quaternary structure of some sHsps, such as human HspB1, HspB4 and HspB5, is often polydisperse - they assemble into species ranging from monomers and dimers up to oligomers more than 40 subunits in size. Highly detailed information about the quaternary structure of proteins can be obtained with, for example, native mass spectrometry (MS) and mass photometry (MP), as we discuss further in this chapter, and later apply these experimental methods to study highly heterogenous protein systems in later chapters.

The effect of the quaternary structure on the function of sHsps

Both HspB4 and HspB5 are highly concentrated in the eye lens, where the total protein concentration can vary anywhere from about 400 mg/mL in humans ([24])

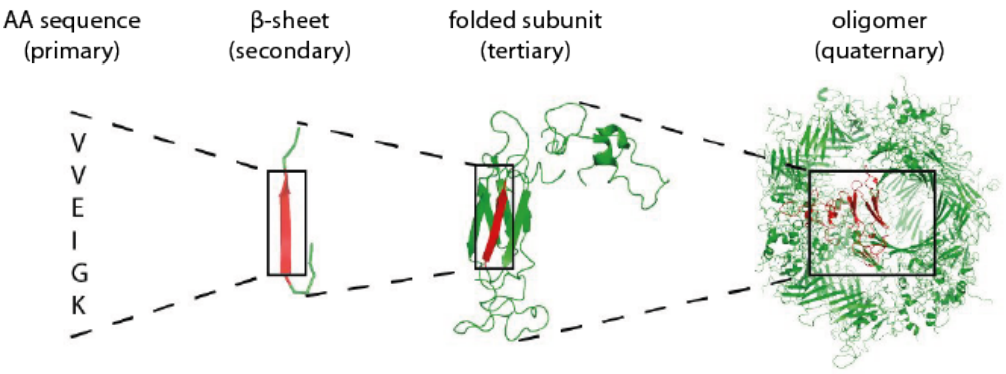


Figure 1.1: The four levels of protein complexity. From left to right: 1) the primary structure is the AA sequence of a given protein that might 2) fold into secondary structure elements. 3) The tertiary structure is defined as the fold of the whole protein monomer. Finally, 4) the quaternary structure is broadly the oligomerisation state of monomeric subunits.

to well over a gram per millilitre in blue-eyed trevally fish ([25]). Here, both the unusual structural properties and their chaperone activity play a functional role. The highly polydisperse, variable quaternary structure of these proteins presumably prevents their aggregation even at the high protein concentrations mentioned above ([26]). Indeed, this is one of the reasons that so far, it has not been possible to obtain a full-length X-ray crystal structure of a polydisperse sHsp. However, even truncated HspB4 and HspB5 crystal structures allude to the multiple structural mechanisms that are important for quaternary structure polydispersity: namely, the palindromic sequence in the C-terminus; the flexible hinge loop preceding it; and the multiple possible registers (AP_I-AP_{III}) for the dimer interface ([26]).

Additionally, HspB4 and HspB5 function as molecular chaperones to prevent protein aggregation in the eye lens. Aggregation often seen with ageing as the eye lens proteins accumulate damage caused by UV radiation, oxidation, post-translational modification, and truncations ([28], [29]) that may lead to cataract formation.

Yet we know of at least one human sHsp, HspB1, that is polydisperse even though its prevalence in the eye lens is negligible compared to HspB5 ([30]), which

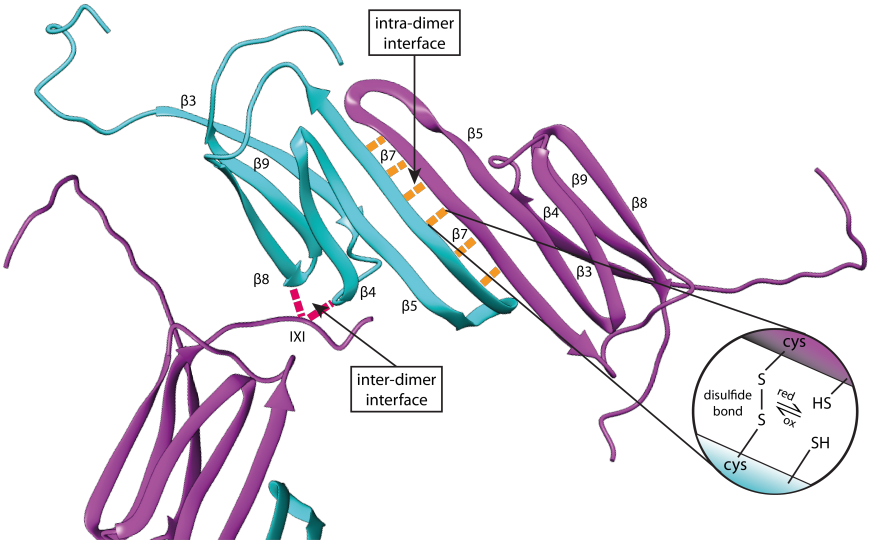


Figure 1.2: The inter- and intra-subunit interactions. Dimer association is mediated by the anti-parallel pairing of the $\beta 7$ strand on each subunit. The IXI motif in the CTD can patch the hydrophobic groove between the $\beta 4$ and $\beta 8$ strands, and is often responsible for inter-dimer binding. Figure adapted from [27].

might prompt the question: How does the quaternary structure of these proteins affect their chaperone-like function?

Even though some sHsps, like human HspB3, HspB6, HspB7 ([4]) do not form any higher stoichiometries to an appreciable extent–while still being chaperone-active–a wide range of sHsps can form higher oligomers. To date, no systematic experimental study on the effect of these higher oligomers, and their potential polydispersity, on the chaperone-like function has been conducted, and there seems to be evidence for both sides of the argument about the structure-function relationship.

Jacobs et al. ([31]) suggest that the optimal chaperone-substrate binding results in the chaperones themselves self-assembling. Using lattice modelling, they found that chaperone-chaperone interactions do not affect the stability of a protein solution–based on a model with aggregation-prone protein substrates assembling through strong, directional interactions. They also showed that under fine tuning to ensure the chaperone-substrate interaction is both reversible, but at the same time still

protects the proteosome, the larger oligomers are a direct consequence of this tuning.

Another interesting case is HSP18.1, a plant sHsp from *Pisum sativum* (pea), which is mostly a dodecamer under normal conditions below approximately 30 °C. However, with increasing temperature, there seems to be an increase in the average size of these higher oligomers and consequently their polydispersity. This change seems gradual, not corresponding to any secondary or tertiary structure changes, and stems from the varying thermal stabilities of the higher oligomers. In a further experiment, upon incubation with luciferase from *Photinus pyralis* (eastern firefly) at 42 °C, they form vastly polydisperse and heterogeneous luciferase:HSP18.1 co-assemblies, where the number of HSP18.1 subunits seems to increase with the number of luciferase molecules bound within the luciferase-HSP18.1 complexes. Since virtually all luciferase:HSP18.1 co-assemblies had more than twelve HSP18.1 subunits, Stengel et al. ([32]) then hypothesise that the dodecamers could function as a reservoir for the higher, active states of HSP18.1 with its substrates.

In 2008, Benesch et al. ([33]) found that upon phosphorylation, HspB5 becomes more chaperone-active. This comes with changes in the sub-structure of its higher oligomers from that based on both monomers and dimers to predominantly monomer sub-structure. The changes in chaperone activity and sub-structure happen seemingly without affecting the overall range of stoichiometries formed. In 2011, Baldwin et al. showed that this peculiar behaviour might be inherent for polydisperse sHsps ([34]). In their model, even oligomers are built of dimers, while odd oligomers are built of dimers with one odd monomer with varying dimer and edge interface strengths. In this context, the dimer interface is the intra-dimer interface and 'edge' is it inter-dimer interface. Most importantly, the quaternary dynamics is completely unbiased in these systems, which gives rise to the broad range of stoichiometries, and can be explained in terms of monomer exchange.

Monomer exchange was seen in other sHsps ([35]), but some others are known to be able to exchange dimers ([36]).

All these studies, and many more, seem to suggest that the larger oligomers might be reservoirs for the smaller, active species; whether they originated as a consequence for fine tuning for chaperone-substrate reversibility, or whether they form oligomers to sequester themselves from the cellular environment to not unnecessarily promote protein aggregation; or both. The ability to form a wide range of oligomeric sizes may also allow for a broader range of how many and what size of these substrates they are able to sequester ([37]). However, no direct evolutionary proof of this behaviour is available at the moment.

The question of how the co-assembly between two or more sHsps may affect their function is even more complicated. We know that human sHsps co-assemble in vitro ([5], [6]), and also possibly in-vivo ([7]), but a detailed description of these co-assemblies, let alone how they interact with possible substrates, has been challenging.

1.3 Evolution and variety of sHsps

The sHsps are present in almost all living organisms (with the notable exception of a couple of pathogenic bacteria ([38]) that possibly depend on the regulation machinery of the host organism for protection from hostile environments). This makes them an attractive target for evolutionary studies. Starting at the domain level, Krierbuber et al. ([39]) were able to show that the three domains: N-terminal domain (NTD), C-terminal domain (CTD) and α-crystallin domain (aCD) evolve independently; they show that the aCD phylogeny often reflects speciation events, whereas the other two domains evolve independently of these events, and often faster than the aCD. Notably, though, there were changes to the aCD, too. For

example, bacterial sHsps would have several evolutionary branches taking only the aCD sequences into account, possibly reflecting functional diversification.

Another notable evolutionary sHsp-related feat happened in plants, which are the only known eukaryotes, other than *Drosophila* ([40]) and *Toxoplasmosa gondii* ([41]), to have sHsps in various membrane-bound compartments rather than just in the cytosol. This might reflect the sophistication of their protein homeostasis machinery that comes with their reduced ability to move, which means they have to withstand harsher conditions compared to organisms able to move. This is supported by the fact that there are no organelle-targeted sHsps in algae ([42], [3]).

Most of the work in this thesis focuses on vertebrate sHsps. Franck et al. ([43]) surveyed the evolutionary diversity of vertebrate sHsps and found two additional HspB homologs (in addition to the 10 human sHsps) in bird and amphibians, which they named HspB11 and HspB12, and a further four (HspB12-HspB15) in teleost fish. Most of these paralogs originated during a period of extensive gene duplications early in vertebrate evolution.

Focusing in more detail on the different vertebrate lineages, they found that two further surveyed mammalian genomes (mouse and rat) only have the same 10 HspB homologs as humans. Meanwhile, *G.gallus* (red junglefowl) as a representative of the bird-reptile clade, shares seven HspB mammalian orthologs (HspB1-5, HspB7 and HspB8), but has three additional HspB homologs not found in mammals–HspB11a, HspB11b and HspB12. No other paralogs were found in other sequenced avian and reptilian genomes. At the time of the publication, amphibian genomes were not fully sequenced, but the authors still found five orthologs of human HspBs (HspB1-2, HspB4, HspB7-8) and multiple genes for HspB11-like homologs in *X. laevis* (African clawed frog). They further found a HspB5-like homolog in *R. catesbeiana* (American bullfrog). As for the teleost fish, a combined 12 HspB homologs were found in

T. rubripes (the Japanese puffer) and *D. rerio* (zebrafish). Seven of those were paralogous to the mammalian HspBs.

1.4 Native mass spectrometry

The history of mass spectrometry (MS) starts in the early 20th century when the physicist J. J. Thomson successfully developed what he called a 'mass spectrograph' to measure the atomic weight of different elements ([44]). His students then went on to develop more improved versions of this instrument and proved the existence of two stable isotopes of neon, ^{20}Ne and ^{22}Ne. The field continued to develop during the 1920s and 1930s with the discoveries of, among others, the carbon isotope ^{13}C widely used in metabolic studies or uranium-235, the fissionable isotope of uranium ([45]), as well as the studies of small organic molecules. It was only in the 1960s that mass spectrometry found a biological application, initially to study peptides. Any wide biological applications were however quite hindered by the method used to ionise the peptides from solution–electron impact, which required volatile derivates ([46]). With gentler ionisation methods (discussed in the next section, 1.4.1), it has become increasingly possible to study biological structures under more 'native' conditions.

Nowadays, it is possible to preserve large protein complexes in the gas phase, which has opened mass spectrometry to the fields of biophysics and structural biology ([47], [48], [49], [50], [51]).

Commonly termed native mass spectrometry, the method encompasses a wide range of different techniques for preserving quaternary-level interaction from solution, through ionisation ([52], [53]) and transmission ([54], [55]) through the mass spectrometer, all the way to the mass-to-charge detector ([56], [57], [58]).

In the past two decades, native MS has been used to answer increasingly complex and diverse questions. It moved from native monomers to protein complexes ([52]).

It was used to study interface strengths and their evolution ([59]), to study ribosomes ([60]) and to look at the topology of different oligomerisation modes ([61], [62]). Or even to describe highly polydisperse systems like sHsps ([63]). Nowadays, it is even possible to study membrane proteins both embedded in a lipid vesicle ([64]) as well as coming directly from mitochondrial membranes ([65]).

1.4.1 Nanoelectrospray ionisation

Since the early day of mass spectrometry, one of the key issues has been how to transfer the analyte molecules into the gas phase. Some earlier designs included fast atom bombardment, plasma desorption, and thermospray ionisation ([66]). These were, however, not suitable for the transfer of either higher-mass species or species directly from solution. In the 1980s, it became possible to ionise species with mass largely exceeding the then-limit of a couple kDa: electrospray ionisation (ESI), which was later further developed into nano-electrospray ionisation (nESI). In this method, a so-called 'Taylor cone' is formed at the tip of the spray emitter (a needle) above a certain voltage. The small droplets are formed which undergo multiple Rayleigh divisions and finally ion desorption (Fig. 1.3).

There are a couple of models for how the droplets lead to charged proteins that can be later transmitted and analysed in the mass spectrometer: the ion evaporation, charged residue, and chain ejection models ([51]).

In the ion evaporation model, species with low molecular weight (generally hundreds of Daltons) are expelled from the charged droplets in a solvation shell, which is then lost at the interface of the mass spectrometer.

In the charged residue model (CRM in Fig. 1.3 b), which best describes the charging or bigger globular proteins, the solvation shell is slowly lost with the charge being transferred to the sample species that was inside the charged droplet.

Finally, the chain ejection model (CEM in Fig. 1.3 b) is the most applicable for unfolded proteins. Due to having a large hydrophobic area exposed to the solvent inside the water droplet, they are not stable inside the droplet. A part of the unfolded species can migrate to the surface of the droplet where it might get charged which causes an ejection of the analyte molecule out of the droplet.

Generally, for bigger species, nESI results in a range of m/z values for the same species (which are often gaussian-distributed)–this allows for less ambiguous mass assignment (Fig. 1.3 c))

nESI is recognised as one of the most efficient methods of introducing liquid samples into the mass spectrometer ([67]).

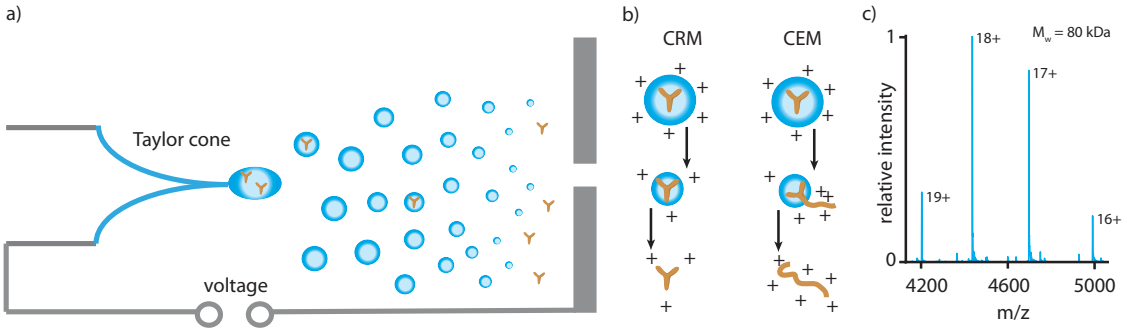


Figure 1.3: Description of the nESI process. a) The Taylor cone is formed at the tip of the spray emitter above a certain voltage. Small droplets are formed which undergo multiple Rayleigh divisions and evaporation events, eventually leaving just the charged analyte. b) Graphical description of the two different charging mechanisms: the charge reduction model (CRM), and the charge ejection model (CEM). c) Mass spectra resulting from nESI often have multiple charge states corresponding to one mass.

1.4.2 Transmission and m/z selection

After entering the instrument, the ions are guided through a quadrupole, which consists of four parallel rods onto which direct current (DC) and radiofrequency (RF) potentials (effectively alternating current(AC)). Only certain m/z will have stable ion trajectory and reach the next chamber within the instrument depending on the ratio between the DC and AC applied, while others with unstable trajectories

will undergo progressively larger oscillation within the quadrupole and collide with the quadrupole rods. The RF of the quadrupole is often reduced to allow the transmission of heavier species. For example, at RF of 832 kHz the upper m/z limit is just 4190 m/z, whereas decreasing the RF down to about 300 kHz or below allows transmission of ions in the tens of thousands m/z range ([51]). Originally, quadrupoles were used as mass analysers, but due to their relatively low mass resolution, they are nowadays mostly used as mass filters. By tuning the RF, we can select a particular m/z that will have a stable trajectory within the quadrupole. All other m/z will have unstable trajectories, and will not enter the next chamber. The pressure in the quadrupole is generally also increased compared to traditional mass spectrometers, which, again, facilitates the transmission of higher masses due to collisional cooling with background molecules.

1.4.3 Mass analysers

As mentioned above, the initial mass analyser of choice was the quadrupole, with which one can scan through the m/z space by selecting appropriate RF and detecting ions with stable trajectories under the RF.

Nowadays, there are other mass analysers, such as the time-of-flight mass analyser, magnetic sector analyser, ion trap analyser or Fourier-transfer ion cyclotron mass analyser ([68], [56]).

In this work, we exclusively use the Orbitrap mass analyser. In the Orbitrap, ions rotate around a spindle. On top of this motion, the ions also undergo oscillations parallel to the spindle. The image current induced by their motions is then Fourier-transformed into a mass spectrum. Since this is independent of any spatial dispersion of the ions, the Orbitrap offers a high resolution compared to some of the older mass analysers. The Orbitrap mass analyser has been adapted only in the past

couple of years for native MS ([69]) and its much bigger m/z range compared to the initial use of the Orbitraps–proteomics.

1.4.4 Native MS in the context of structural biology

Native MS has become an indispensable technique in structural biology as explained above, with an ever-increasing range of biologically relevant applications. It is nowadays possible to use it to study whole viral capsids, often in their DNA-filled forms, that reach up to 4 MDa in mass. For example, in 2008 Morton et al. used native MS to study hepatitis B viral capsids and were able to assign two different morphologies of said capsids with different masses ([70]). Further developments have allowed the study of exchange dynamics ([71]), stability ([72]) or ligand binding ([73]) of these capsids.

Further, it is now possible to study membrane proteins using MS, where the proteins are released from membrane-mimicking detergent micelles by collision-induced dissociation (CID, discussed in the next section) in the gas phase. Since the energy needed to release them is often quite high, it is possible for the proteins to unfold or dissociate. This can be partially corrected by screening a wider range of detergents used to solubilise the membrane proteins. The initial milestone in the development of this application includes the study of *Thermus thermophilus* ATP synthase. With the development of the application of membrane proteins, it is now possible to readily study ATP synthase stoichiometries or ligand binding ([74]). MS can also reveal the stabilising effects of lipid binding on membrane protein stability ([75]). Recently, it has also become possible to study membrane proteins ejected directly from their native environments, as shown with the respiratory chain complexes in mitochondrial membranes of *E. Coli* ([65]).

Native MS can also be used to study the shape and stability of protein assemblies using so-called ion-mobility MS (IM MS). Ion mobility is a separation technique orthogonal to the MS m/z separation. Rather than separating ions based on their mass (and corresponding m/z), it separates ions based on their shape, which additionally allows us to study their size and stability.

There are multiple different IM methods, with the perhaps simplest and most traditional being drift mobility. In this method, ions pass through a drift tube containing an inert gas under an electric field gradient. The drift time (the time it takes for the ion to pass through the drift cell) is related to their shape as ions with a bigger collision cross-section will experience more collision events resulting in a higher drift time. Molecular modelling can then be used to correlate the drift time to a proposed oligomer geometry.

In this work, we focus on the description of complex, polydisperse co-assemblies. To study these, we use tandem MS.

1.4.5 Tandem MS and collision-induced dissociation

Collision-induced dissociation (CID) is a technique that allows us to dissociate subunits from the parent proteins complex, which can both help to elucidate the internal sub-structure of a protein complex, as well as charge-reduce the resulting 'stripped' species, which often helps with further analysis.

Once the ions reach the collision cell filled with inert gas (generally noble gasses, N_2 or SF_6), they interact with the gas molecules both elastically and inelastically. In elastic collisions, the partners mainly exchange kinetic energy, whereas inelastic collisions transfer the kinetic energy also into vibrational or internal energy ([47]). Heavier gasses often favour more inelastic collisions. After a certain number of collisions, the analyte ions may accumulate enough internal energy to expel one or

more of its (non-covalently bound) subunits. The departing subunit is considered to be unfolded as it often leaves with proportionally quite high charge, suggestive of high surface area. This leaves the 'stripped' oligomer with less charge.

The order in which subunits leave might be suggestive of their relative arrangement and size. Generally, peripherally located subunits that unfold more easily and have smaller interaction area leave first ([47]).

In a tandem experiment, a particular m/z range can be selected in the quadrupole before the selected ions undergo CID. On some instruments, this can be done multiple times which may help to, as an example, reveal the subunit arrangement in the first couple of cycles followed by isolating the subunits that interact with a ligand, and finally characterising the ligand itself.

In this work, we use a single tandem MS 'cycle': one m/z selection followed by CID. The mechanism is further explained in Chapter 3, but briefly: m/z selection is used to narrow the number of species of interest followed by CID to charge-reduce the selected oligomers and obtain a baseline-resolved, deconvolvable spectrum, similar to that shown in Fig. 1.4 c).

1.5 Mass photometry

Mass photometry (MP) is an optical, in-solution method based on interferometric scattering (iSCAT) that allows the quantification of single molecules through their mass, as the mass and the iSCAT contrast (as explained in more detail below) are directly correlated. This allows us to indirectly determine the mass distribution of any biomolecular solution.

Simplified, in an MP experiment, the analyte solution is dropped onto a suitable transparent substrate, usually a glass coverslip, that is positioned on top of, and in contact with, the MP setup. The setup comprises of a laser source that is

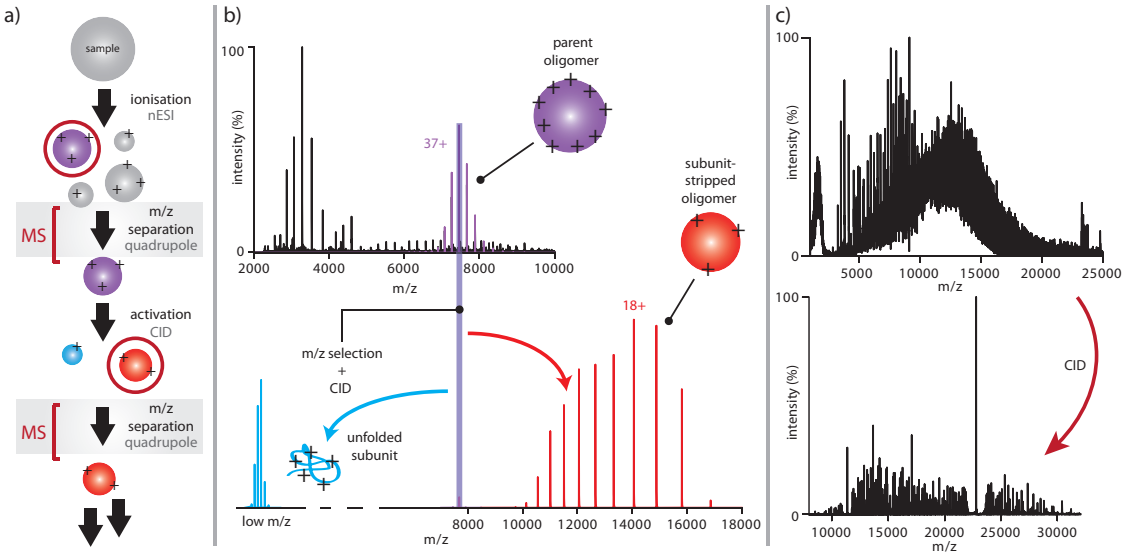


Figure 1.4: Overview of an MS/MS experiment. a) In tandem MS experiments, multiple sequential m/z selections and CID events are done to better elucidate the quaternary structure of protein complexes. For this work, we only used one selection and CID cycle followed by m/z detection. b) An example of one MS/MS cycle. In the top spectrum, we see charge series corresponding to a dimer (black) and an octamer (purple). One of the octamer m/z peaks is selected and CID-ed and loses a highly charged monomer (blue). The resulting heptamer (red) is charge-reduced as it lost, on average, roughly 20 elementary charges. c) CID applied to the whole unresolvable 'hump' typical of highly polydisperse systems results in charge reduction (move towards higher m/z). This spectrum is now baseline-resolved, as there is less overlap between the charge series. Figure adapted from [27].

focused onto the formed glass-analyte solution interface through a set of optical lenses, acoustic-optical deflectors, quarter-wave plates, polarising beam splitter, and other components, as described in detail below. The reflected and scattered light from the interface are then collected by a suitable image sensor, usually a CMOS sensor, and used for analysis.

1.5.1 Interference-based microscopy

In an interferometric experiment, as the transparent substrate is illuminated (I_i), the majority of the light passes through the substrate. However, a small portion is reflected (I_r) and an even smaller portion of the light is scattered (I_s) from the interface–these are then collected by the objective lens. The detected intensity

(I_d) at a given point in the image is given by:

$$I_d = I_i[r^2 + |s^2| + 2r|s|cos(\Phi)] \tag{1.1}$$

where r^2 is the reflectivity of the interface, s is the scattering amplitude of the object, and Φ is the phase difference between the reflected and scattered fields.

Here, the scattering amplitude is directly proportional to particle volume ([112]).

Assuming the particle volume is sufficiently small, we can then neglect the $|s^2|$ term, and define the interferometric scattering contrast, c, as the ratio of light intensity in the presence of a scatterer (I_d) and the absence of a scatterer (essentially the background, I_b):

$$c = \frac{I_d}{I_b} = \frac{r^2 + 2r|s|cos(\Phi)}{r^2} = 1 + \frac{2|s|cos(\Phi)}{r} \tag{1.2}$$

There are multiple ways of increasing c; such as reducing the reflectivity of the interface by changing the solvent of choice or coating the glass surface, but changing the solvent would practically make the measurements biologically incompatible, and coating is often impractical and costly.

Another solution is placing a partial reflector in the beam path; the reflector selectively attenuates the reflected light travelling in the centre of the beam while leaving the scattered light that travels around the edges mostly unaffected.

The full iSCAT setup can be seen in Fig. 1.5.

1.5.2 Data analysis

After the 'native' MP movie is recorded, the final step is background subtraction for visualising the analyte molecules interacting with the glass surface, which can then be further analysed to obtain the masses of the analyte species.

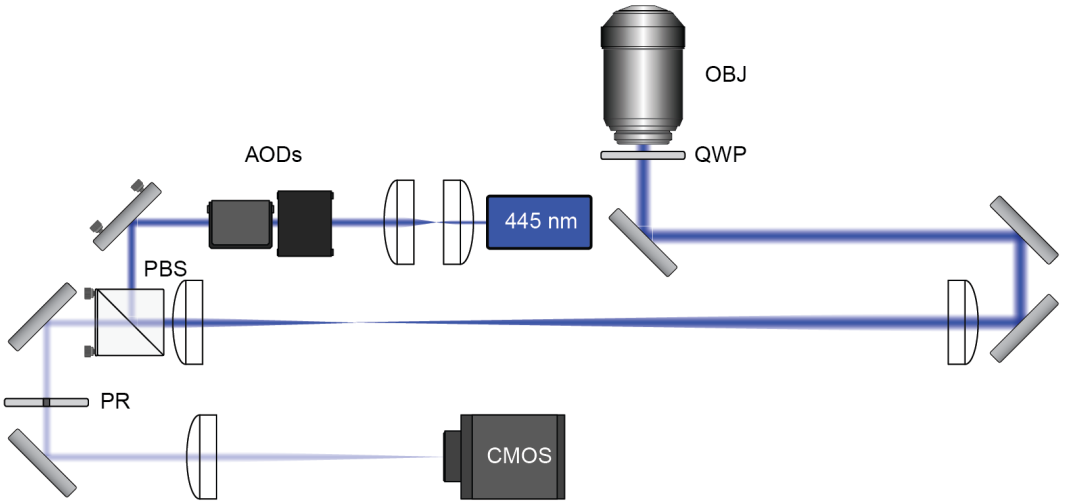


Figure 1.5: The iSCAT setup used for mass photometry. The setup compromises of an objective (OBJ), a quarter-wave plate (QWP), a polarizing beam splitter (PBS), two orthogonally oriented acoustic-optical deflectors (AODs), a partial reflector (PF) and a CMOS image sensor. Figure adapted from [76] with permission.

Background removal is done by subtracting blocks of frames from each other as shown in Fig. 1.6 a) (left of the arrow). This allows for the visualisation of particles binding to the glass surface (Fig. 1.6 a, right of the arrow).

This can be applied to the whole native movie to convert it into a 'ratiometric' movie. In the ratiometric movie, assuming the molecule binds to the glass surface, the intensity of the event initially increases, reaches a maximum, and starts decreasing again as shown in Fig. 1.6 b) and c). This is an artefact of the background subtraction and not the molecule nearing the interface and leaving again; in fact, the landing events are quite fast on the timescale of the refresh rate of the detector, and usually happen between two frames.

1.6 Circular dichroism

Circular dichroism (CD) is an experimental technique that uses polarised light to study the secondary structure and folding of proteins ([79]). CD is defined as

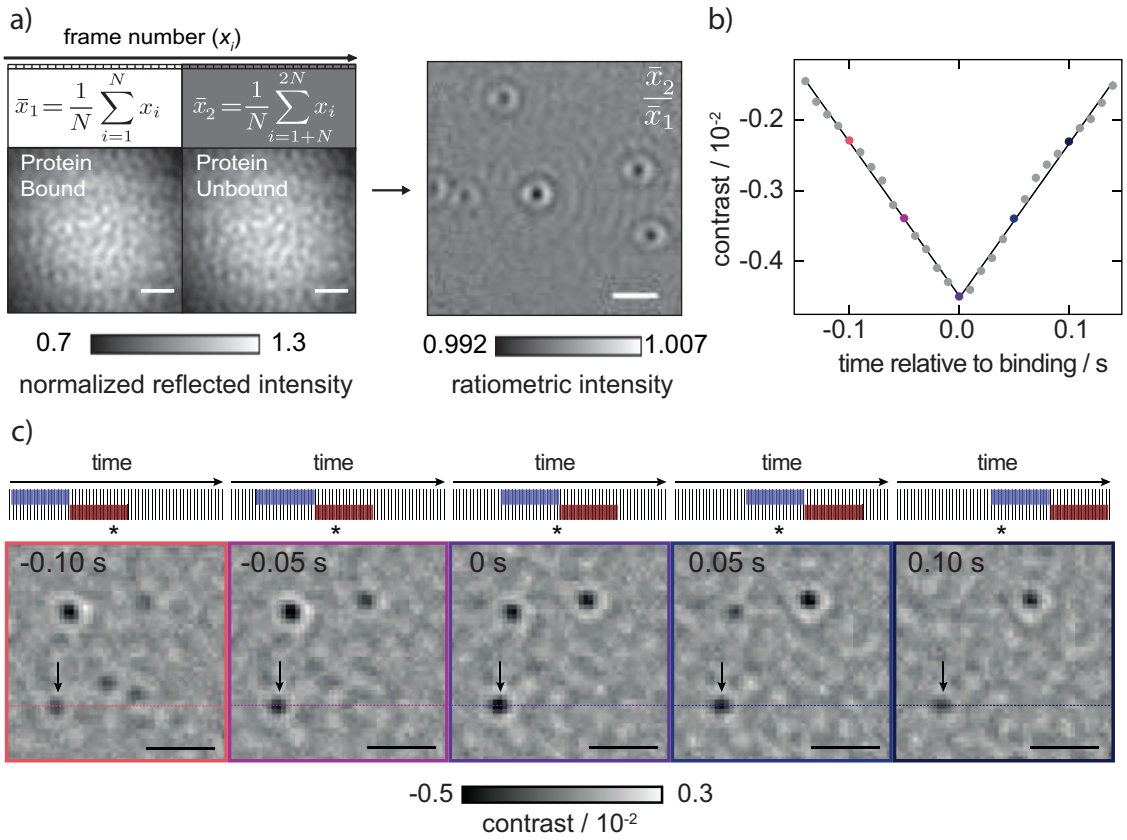


Figure 1.6: The conversion to a ratiometric movie. a) shows the background removal done by subtracting blocks of frames from each other. Once this is applied to the whole movie, the intensity of any landing event initially increases, reaches a maximum, and starts decreasing again (b) and c)), Here, a) is adapted from [77] and b) and c) are adapted from [78] with permission.

the unequal absorption of left- and right-handed circularly polarised light. This unequal absorption happens readily when circularly polarised light interacts with asymmetric elements within a molecule as these elements might absorb the left- and right-handed circularly polarised light to different extents, which might result in the radiation having an elliptical polarisation (Fig. 1.7 a)). Kelly et al. ([80]) mention three such 'asymmetric elements':

1. intrinsically chiral structure, such as a carbon atom with 4 different substituents

2. element covalently linked to a chiral centre in a molecule

3. element placed in asymmetric environments by virtue of the 3D structure adopted by the molecule

The difference in absorbance between the left- and right-handed polarised light is most commonly reported in terms of ellipticity, θ, with units of degrees (deg), where $\theta = tan^{-1}(b/a)$ where a and b are the major and minor axis of the resulting ellipse in Fig. 1.7 a). CD spectrum can then be obtained as a function of wavelength. In general, the ellipticities of biological samples are quite small, on the order of tens of millidegrees.

1.6.1 Secondary structure information available from CD

In proteins, there are multiple chromophore regions of interest; by increasing wavelength:

1. the peptide bond (< 240 nm, far UV)

2. aromatic AA side chain and disulfide bonds (260-320 nm)

3. non-protein cofactors including pyroxidal-5'-phosphate (330 nm), flavins (300-500 nm), heam groups (410 nm), and others...

Usually, the region below 240 nm is the most informative about the secondary structure of a protein due to the absorbance of the peptide bond in this region.

Different secondary structures have characteristic spectra in the far UV region as shown in Fig. 1.7.

The individual characteristic spectra differ due to long-range order in the amide chromophore (N-C=O), but generally show these transitions:

- for the alpha helix, exiton coupling of the p->p* transitions leads to positive (corresponding to p->p*) perpendicular at 192 nm and negative (corresponding to p->p*) parallel at 209 nm negative at 222 nm is red shifted (corresponding to n->p*)

- for beta sheets, there are two transitions: negative at 218 nm (corresponding to p->p*) and positive at 196 nm (corresponding to n->p*)

- finally, even the random coil shows these transitions: positive at 212 nm (corresponding to p->p*) and negative at 195 nm (corresponding to n->p*)

where the n->p* transition comes from the lone pair on oxygen to a anti-bonding orbital, and the p->p* transition correspond to the non-bonding p orbital to the anti-bonding p* orbital within the delocalised N-C=O chromophore.

By assuming that the measured spectrum is a linear combination of these, it is possible to deconvolve the secondary structure of the protein of interest.

It is further possible to follow the CD as a function of increasing temperature in a so-called 'melting' experiment, where we effectively see a loss of ordered secondary structure elements and an increase of the contribution from a random coil to the CD spectrum ([79]) as the protein unfolds (Fig. 1.7 c)).

1.7 Statistical mechanics

Statistical mechanics is a framework that applies statistical and probabilistic descriptions to microscopic ensembles to obtain macroscopic properties of these ensembles.

The field itself was first established in the second half of the 19th century by the work of, most notably, Ludwig Boltzmann, Josiah W. Gibbs (who coined the name 'statistical mechanics' in 1884 ([81])), and James C. Maxwell, who themselves

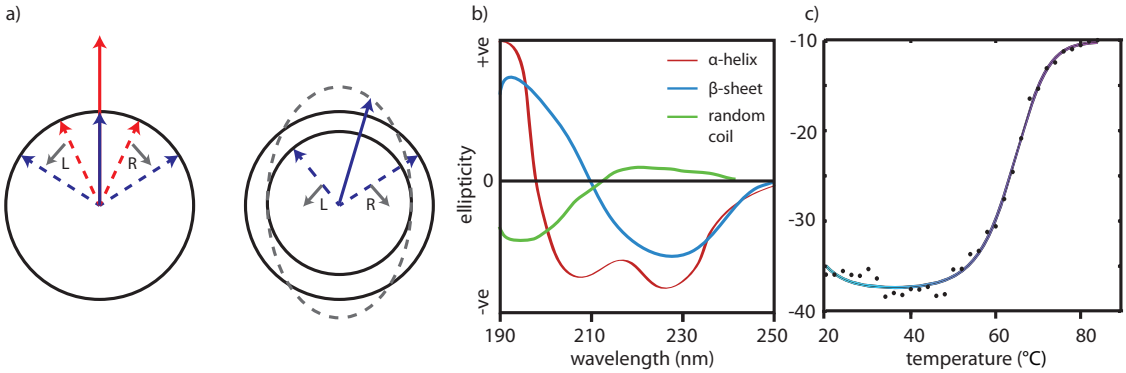


Figure 1.7: Overview of circular dichroism. a) shows the origins of the CD effect. When a plane-polarised light is shone at an achiral molecule, the two circular components will still have the same amplitude and orientation. If the light interacts with asymmetric elements within a molecule that absorbs the left- and right-handed circularly polarised light, the resulting vector of the two contributions is elliptically polarised, giving rise to the CD effect. b) shows roughly the CD spectrum for different secondary structure elements–these can be combined to deconvolve the secondary structure of a protein. There is a contribution to the CD spectrum from the random coil due to the random coil Finally, if we follow one specific wavelength, it is possible to measure the CD at the given wavelength as a function of temperature to extract a melting temperature of the protein.

built their models on the growing belief in the atomic theory; even though some attempts at a statistical-mechanical description of matter can be traced to even earlier, such as D. Bernoulli's work on the kinetic theory of gasses.

Although the first statistical-mechanical descriptions were initially not systematic, most of the framework was formalised and generalised by the early 20th century with the publishing of Gibbs' *Elementary Principles in Statistical Mechanics*.

1.7.1 The fundamental postulate & ensembles

In this work, we only use equilibrium statistical mechanics; that is: we assume that the oligomeric (sub-)stoichiometries we study in the next chapters are in an equilibrium state.

The fundamental postulate of equilibrium statistical mechanics is sometimes known as the equal *a priori* probability postulate.

It states that for given 3D systems with known particle positions and momenta,

in the resulting Γ-space (also called an ensemble) that contains all the systems that satisfy the given conditions, the number of such systems in a given volume element $d^{3N}rd^{3N}p$ is simply $\rho(r,p)d^{3N}rd^{3N}p$ where $\rho(r,p)$ is a suitable probability density function for the given ensemble.

Therefore, the number of systems within a given volume is proportional to the phase-space volume itself. It follows that the *a priori* probability of randomly picking any system from the ensemble is the same; this ensemble is known as the microcanonical ensemble.

Depending on the 'conditions' mentioned above, there are different ensembles. The three most used ensembles (with their 'fixed variables' as explained below the list) are:

- Microcanonical ensemble (E, N, V fixed)
- Canonical ensemble (T, N, V fixed)
- Grand canonical ensemble (T, μ, V fixed)

where E is the system's energy, V is its volume, T is temperature, N is the number of particles, and μ is chemical potential.

The microcanonical ensemble then describes a system with fixed energy, number of particles and volume. In this ensemble, each microstate has the same probability.

The canonical ensemble, with fixed absolute temperature, a number of particles, and volume, represents closed systems in thermal equilibrium with a heat bath at a fixed temperature.

In this ensemble, the energy E of the individual systems is allowed to vary, and so the probabilities associated with individual ones are not generally the same but can be calculated as:

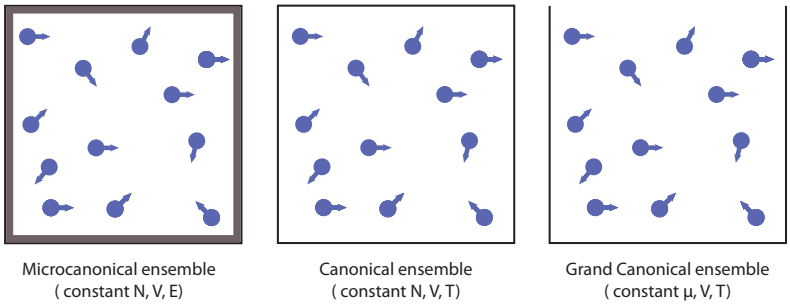


Figure 1.8: Visual representation of the three Canonical ensembles. The microcanonical ensemble is an isolated closed system, meaning that the number of particles, volume, and the energy of the ensemble stays the same. The Canonical ensemble is in a contact with a heat bath and closed, which results in the number of particles, volume, and temperature being the same. Finally, the Grand Canonical ensemble is open and can exchange both energy and particles. In this ensemble, the chemical potential, volume and temperature are kept constant.

$$p = e^{(F-E)/(k_B T)} \tag{1.3}$$

where F is the Helmholtz free energy, E is the system's energy, k_B is the Boltzmann's constant, and T is fixed temperature.

Finally, the grand canonical ensemble is an open system that can exchange both particles and energy with a reservoir, while its volume, chemical potential and temperature are kept constant.

Here, the probability associated with each microstate is

$$p = e^{(\Omega+\mu N-E)/(k_B T)} \tag{1.4}$$

where Ω is the grand potential and is constant for the ensemble, μ is the system's chemical potential, E is the microstate's energy, k_B is the Boltzmann's constant, and T is fixed temperature.

1.7.2 Application to protein co-assembly

Further in this thesis (Chapter 2), we apply statistical mechanics to study the co-assembly of monodisperse proteins.

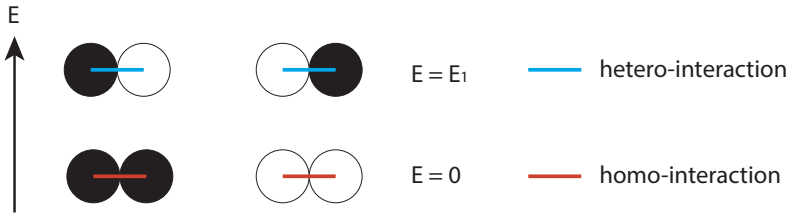


Figure 1.9: The hetero- and homo-interactions can have different energy. Here, we set E(A-A)=E(B-B)=E(homo)=0 and E(A-B) = E_1 for simplicity.

To study the co-assembly, we assume the different interaction energies: hetero-interaction and homo-interaction energy. Since we study the co-assembly between two species, A and B, we can in general assume two different homo-interaction energies, E(A-A) and E(B-B), and one hetero-interaction energy, E(A-B). As a simplification in this work, we set E(A-A)=E(B-B)=E(homo)=0, similarly to what Hochberg et al. used ([82]). We can then study the system purely in terms of the hetero-interaction energy.

As the simplest example, we can use statistical mechanics to study the abundances of various hetero- and homo-dimers depending on the hetero-interaction energy, as shown in Fig. 1.9.

We can then calculate the relative abundance, x_i of each of these species as:

$$x_i = \frac{e^{-E_i/kT}}{Z} \tag{1.5}$$

where Z is in the partition function. For a canonical ensemble that is classical and has discrete energy levels:

$$Z = \sum_i e^{-E_i/kT} \tag{1.6}$$

We can then use Eq. 1.5 to calculate the relative abundance of any of the hetero- or homo-dimers as a function of energy difference, E_1, as shown in Fig. 1.10. Here, with a negative E_1, the two hetero-dimers are highly abundant. As

the energy difference decreases, the two homo-dimers become progressively more occupied. When there is no energy difference, the four different species are equally abundant meaning that the ratio of AA:(AB+BA):BB is binomial–1:2:1. Finally, as E_1 becomes large and positive, only the two homo-dimers are occupied to an appreciable extent.

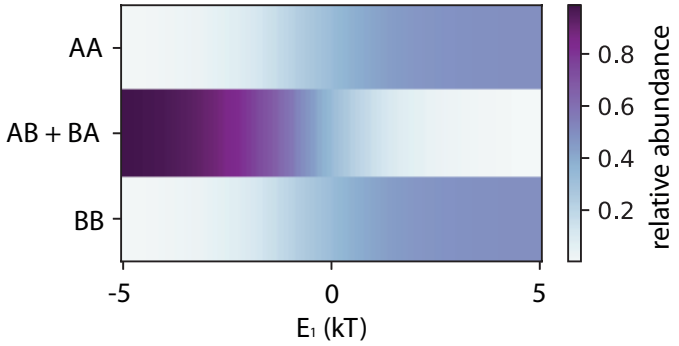


Figure 1.10: The relative abundance of the two hetero- and two homo-dimers as a function of energy difference.

We can use the same reasoning to build up and calculate the relative abundances of higher oligomers as we show in Chapter 2.

1.8 Chemical kinetics

Chemical kinetics is a branch of chemistry that focuses on the quantitative description of reaction rates and their mechanism through mathematical modelling.

The field itself was established in the late 19th century through the seminal work on the law of mass action, which states that the rate of a chemical reaction is proportional to the concentration (or activity) of reactants. At around the same time, Jacobus Henricus van 't Hoff published his work on chemical dynamics, "Études de dynamique chimique", for which he won the first Nobel prize in chemistry in 1901 ([83]). In the 20th century, the field continued developing into the broad

field as we know it now through a focus on reaction mechanisms and complex chemical reaction networks.

1.8.1 Reaction rates & rate equations

The rate of the following general reaction:

aA + bB $\longrightarrow$ cC

can be calculated with respect to any reactant or product as:

$$\frac{d[A]}{adt} = \frac{d[B]}{bdt} = \frac{-d[C]}{cdt} \tag{1.7}$$

In general, the reaction rates depends on the concentration of reactants and products. Assuming the reaction above is not reversible for simplicity, the reaction rate can be expected to be of the form

$$\frac{-d[C]}{cdt} = k[A]^{\alpha}[B]^{\beta} \tag{1.8}$$

where α and β often need to be determined experimentally for reactions with a more complicated, multi-step, reaction mechanisms, and k is the rate constant which, according to the Arrhenius equation only depends on temperature ([84]):

$$k = Ae^{\frac{-E_a}{RT}} \tag{1.9}$$

where A is the pre-exponential factor, constant for each reaction, E_a is the activation energy for the reaction, R is the universal gas constant, and T is the absolute temperature.

While the Arrhenius equation is semi-empirical, there is a an approximate relation between it and the Eyring equation which describes reaction rates using transition theory. Here, the rate can be calculated as:

$$k = \frac{k_B T}{h} e^{\frac{-\Delta G^*}{RT}} \tag{1.10}$$

where k_B is the Boltzmann constant, T is the absolute temperature, h is the Planck's constant, R is the universal gas constant, and ΔG^* is the Gibbs energy of activation to reach the transition state.

Assuming a unimolecular, one-step reaction, and

$$\Delta G^* = \Delta H^* - T\Delta S^* \tag{1.11}$$

where ΔH^* and ΔS^* are the enthalpy and entropy of activation, respectively, the approximate relationships for A and E_a are:

$$A = \frac{k_B T}{h} e^{1+\Delta S^*/R} \tag{1.12}$$

and

$$E_a = \Delta H^* + RT \tag{1.13}$$

where k_B is the Boltzmann constant, T is the absolute temperature, h is the Planck's constant, and R is the universal gas constant.

In the case of subsequent additions or losses of monomers, which seems to be the case for polydisperse sHsp co-assembly ([34]), the formation of an N-mer proceeds through the addition of an sHsp monomer, A, to a (N-1)-mer:

$(\mathrm{N}-1)_{\mathrm{mer}} + \mathrm{A} \longrightarrow \mathrm{N}_{\mathrm{mer}}$

Similarly, the general N-mer can lose one of its N subunits:

$\mathrm{N}_{\mathrm{mer}} \longrightarrow (\mathrm{N}-1)_{\mathrm{mer}} + \mathrm{A}$

Both of these are assumed to be elementary reactions.

Assuming the rate of the reaction is proportional to the concentrations of the (N-1)-mer, N-mer, and the single monomer A; that is, the forward reaction is bi-molecular, and the backwards one is uni-molecular:

$$\frac{d[N_{mer}]}{dt} = k^{on}[(N-1)_{mer}][A] - k^{off}[N_{mer}] \tag{1.14}$$

At equilibrium, the concentration of reactants and products does not change. Thus:

$$\frac{d[N_{mer}]}{dt} = 0 \tag{1.15}$$

and so:

$$k^{on}[(N-1)_{mer}][A] - k^{off}[N_{mer}] = 0 \tag{1.16}$$

from which, it is possible to calculate the equilibrium concentration of any given N-mer. We expand on this idea later on in Chapter 2 to model the equilibrium distribution and kinetics of sHsp self- and co-assembly.

1.8.2 The Euler method

To model dynamic processes when the system is not at equilibrium, we need to solve numerous simultaneous ordinary differential equations (ODEs). Oftentimes, in complex systems, this becomes practically very difficult, and the solution has to be found numerically. In this work, we used the Euler method to find the equilibrium solution and simulate the equilibration process. The Euler method is the most simple explicit method for numerical integration of ODEs in which the error per step is the square root of the step size, and the global error is proportional to the step size ([85]).

In simple terms, given a function $y(t)$ with a known value for a given t, the value of the function, $y(t + nh)$, at step, n, can be calculated as:

$$y(t + nh) = y(t + (n - 1)h) + hf(t + (n - 1)h, y(t + (n - 1)h)) \tag{1.17}$$

from the known value at $y(t)$ in n successive step of size h, where $f(t, y(t)) = y'(t)$. The solution is explicit, meaning that the solution $y(t + nh)$ is an explicit function of $y(t + ih)$ where $i \leq n$.

The accuracy of the Euler method solution depends on the value of h as with larger h, the error generally increases.

To show how this can be applied to find a solution to a chemical kinetics problem, we can imagine a simple reaction

A $\xrightarrow{\mathrm{k}^a}$ B $\xrightarrow{\mathrm{k}^b}$ C

in this context

$$\frac{d[A]}{dt} = -k^a[A] \tag{1.18}$$

$$\frac{d[B]}{dt} = k^a[A] - k^b[B] \tag{1.19}$$

$$\frac{d[C]}{dt} = k^b[B] \tag{1.20}$$

Instead of solving the ODEs analytically, we can find a numeric solution using the Euler method as we know that:

$$\Delta[A] = k^a[A]\Delta t \tag{1.21}$$

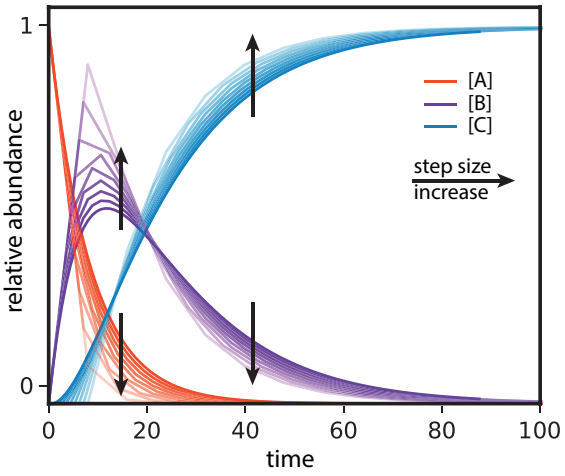


Figure 1.11: Divergence of the Euler method. Example of how the Euler method solution diverges from the exact solution as the step size, Δt increases. The equilibrium concentrations of species A, B, and C, as a function of time (number of steps multiplied by Δt) in red, purple, and blue, respectively. The increase in chosen step size is shown as reduced opacity of the corresponding curve, and from 1 to 9 in increments of 1.

$$\Delta[B] = (k^a[A] - k^b[B])\Delta t \tag{1.22}$$

$$\Delta[C] = k^b[B]\Delta t \tag{1.23}$$

This solution is easy to solve numerically. To explore how the solution varies with Δt, we set $[A]_{init} = 1$, $[B]_{init} = [C]_{init} = 0$, $k^a = 10^{-3}$ and $k^b = 5 \times 10^{-4}$ and varied Δt from 1 to 9 in increments of 1. The resulting solutions can be found in Fig. 1.11. Indeed, as we increase Δt, the approximate solution deviates from the exact solution, and for time steps $\Delta t > 20$, the solution started reaching negative concentrations (not shown).

2

Physico-chemical description of protein co-assembly

Contents

2.1 Motivation

Quantitative solutions for the distribution of oligomerisation states of both polydisperse sHSPs ([34]), as well as other polydisperse homo- ([86]) and hetero- ([87]) -oligomeric systems have been described in the past decade. Recently, Hochberg et al. ([82]) published a quantitative description of non-mixing monodisperse

sHsps. However, a quantitative description of heterogeneous, polydisperse sHsp co-assemblies has been elusive; arguably because these systems have so far evaded a quantitative experimental description. With recent advances in mass-measurement techniques, this has become possible, as we show in the following chapters. As we show in Chapter 3, these advances allow us to study polydisperse co-assembly experimentally. This necessitated an expansion in protein assembly modelling that can describe these newly acquired results. In this chapter, we introduce models that can describe the assembly of protein molecules, and expand on those models, to build a robust model of co-assembly in both monodisperse and polydisperse systems. We also show that in particular limits, the models simplify to the models that have already been described in the literature ([34], [82]).

In general, there are several modes of interactions between two monodisperse proteins (Fig. 2.1). One extreme example is two (monodisperse) proteins that don't interact at all with each other (upper left). The opposite extreme would be the formation of obligate $n:n$ hetero-mers (upper right). The last example we point out here would be unbiased co-assembly, in which case the two proteins interact indiscriminately with each other (upper middle). These same arguments can be made for the co-assembly between polydisperse proteins (lower half of Fig. 2.1).

Here, 'specific self-assembly' means homo-oligomeric assembly and 'specific co-assembly' means hetero-oligomeric, $n:n$ oligomerisation.

At this point, we don't have enough data to reliably model a 'cross'-co-assembly between polydisperse and monodisperse proteins, so we only explore the two extreme cases we have experimental data for–the co-assembly of two monodisperse proteins and the co-assembly of two highly polydisperse proteins. We use statistical mechanics to describe the former and chemical kinetic to describe the latter.

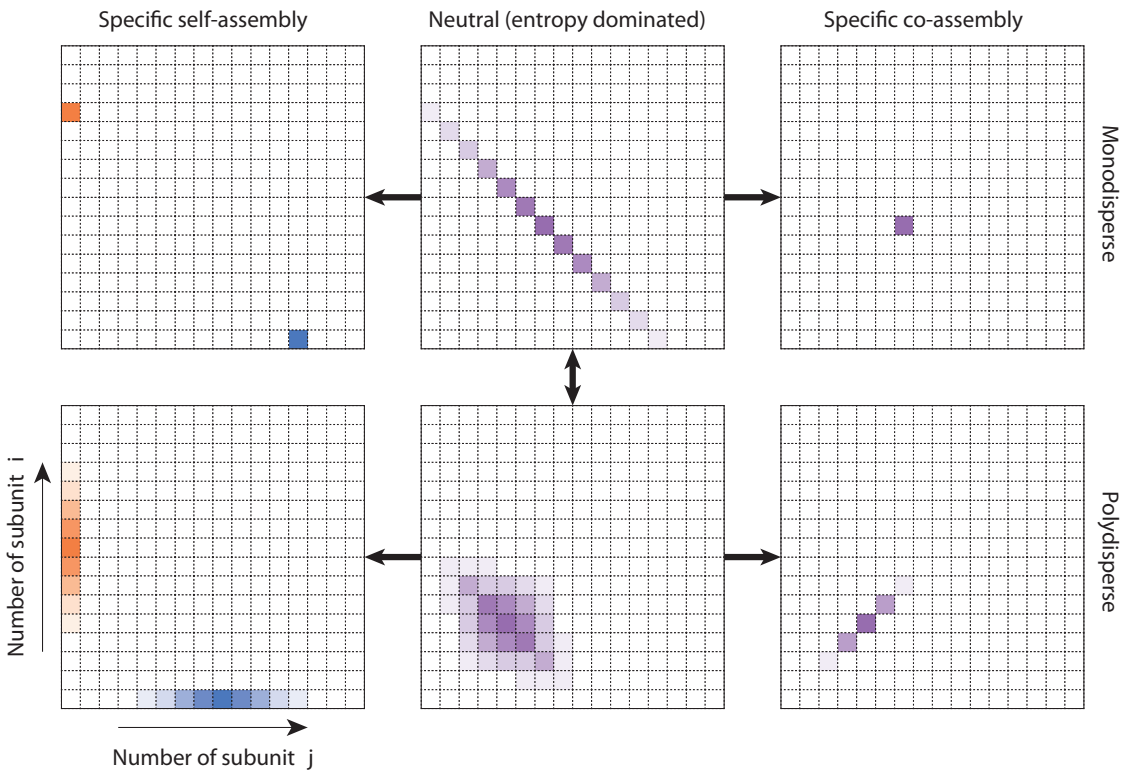


Figure 2.1: Possible modes of protein self- and co-assembly, both in proteins with well-defined quaternary structure, and in polydisperse proteins. The left plots show specific self-assembly where there is no inter-protein co-assembly. Right plots show an obligate n:n co-assembly, where the proteins prefer to strictly co-assemble. The plots in the middle show 'unbiased' co-assembly, where there is no energy preference between hetero- and homo-oligomerisation. Orange and blue colours are used for the case of strict self-assembly. Any co-assembled species are in purple. We use opacity for the same colour to show the relative abundance. I would like to thank Prof. Justin Benesch for creating this figure.

2.2 Introduction

2.2.1 Co-assembly between two monodisperse proteins

Hochberg et al. ([82]) developed a statistical-mechanical description of interaction between two monodisperse proteins (such as shown in Fig. 2.1, top left) to show the energies needed to keep the oligomers self-assembled with only very minor co-assembly (on the scale of the experimental error of MS measurements, roughly 2 percent of species are allowed to hetero-oligomerise). Here, we expand their model to describe the complete spectrum of co-assembly, from pure self-assembly

to obligate, n:n co-assembly.

Boltzmann description of co-assembly of equal mixtures

For two distinguishable types of subunits, A and B, there is a favourable mixing entropy that reflects the number of ways the subunits can co-assemble: $\frac{(a+b)!}{a!b!}$, where a and b are the number of A and B subunits within a particular sub-stoichiometry, respectively. Opposite to this is the fact that (AB)-hetero-interaction may be either more or less favourable than (AA)- and (BB)-homo-interaction, with any associated energy penalty depending on the number of unfavourable contacts formed within a particular hetero-mer. With this model, we need to assume a co-assembly geometry with oligomers bigger than a 12-mer. This is because for species bigger than 12-mers, the oligomers can either co-assemble in a circular arrangement still, or co-assemble into tetrahedron-based structures. Hochberg et al. assumed a ring geometry for these assemblies, with CTD-aCD interactions running along top and bottom, forming pseudovertices. However, some sHSPs are also known to have a tetrahedron-based structure ([88], [3], [89]). Compared to a tetrahedron, where there are four vertices, the ring geometry allows for subunits to be arranged with a smaller number of incompatible interactions in a heteromer. The ring geometry will therefore lead to a conservative estimate for how unfavourable heteromers are compared to homomers for oligomers with more than 12 subunits.

The resulting relative populations of these co-assemblies can then be calculated using the Boltzmann distribution:

$$x_i = \frac{\sum_{k=1}^{\binom{N}{i}} e^{-\Delta\Delta G_k^i/RT}}{\sum_{j=0}^{N}\sum_{j=0}^{\binom{N}{j}} e^{-\Delta\Delta G_k^j/RT}} \tag{2.1}$$

where x_i is the mole fraction of N-mers containing i A subunits, k gives all possible subunit arrangements in an N-mer containing i number of A subunits,

and $\Delta\Delta G_k^i$ is the energy difference between homo- and hetero-interactions with i number of A subunits of the size N in an arrangement k.

For an unbiased co-assembly, where $\Delta\Delta G_k^i = 0$, the resulting distribution is binomial:

$$x_i = \frac{\sum_{k=1}^{\binom{N}{i}} e^{-\Delta\Delta G_k^i/RT}}{\sum_{j=0}^{N}\sum_{j=0}^{\binom{N}{j}} e^{-\Delta\Delta G_k^j/RT}} = \frac{1}{2^N}\binom{N}{i} \tag{2.2}$$

Boltzmann description of co-assembly of unequal mixtures

Often, the two species can't be mixed in an exactly 1:1 ratio. Constraining the conservation of energy, and the total number of each A and B subunits, Hochberg et al. ([82]) show that it is possible to calculate the fractional population x_i as defined above, and where p is the molar fraction of A subunits, as:

$$x_i = a^i b^{N-i} \sum_{k=1}^{\binom{N}{i}} e^{-\Delta\Delta G_k^i/RT} \tag{2.3}$$

where a and b can be found by solving the following two formulas:

$$\sum_{i=0}^{N} i a^i b^{N-i} \sum_{k=1}^{\binom{N}{i}} e^{-\Delta\Delta G_k^i/RT} = Np \tag{2.4}$$

and

$$\sum_{i=0}^{N} (N-1) i a^i b^{N-i} \sum_{k=1}^{\binom{N}{i}} e^{-\Delta\Delta G_k^i/RT} = N(1-p) \tag{2.5}$$

The authors, however, only used this formula to estimate the energy bias needed so that only 2 percent of all oligomers are mixed (i.e. have at least one subunit of either of the two co-assembling species) as a function of size. This description is inadequate to describe the complete range of co-assembly from pure self-assembly to obligate n:n co-assembly, as we show later. To extend their model, we devise a

counting method for the multiplicities of different circular subunit arrangements and briefly apply this description to a system of two plant sHsps that partially co-assemble, and extract the energy bias of this co-assembly.

2.2.2 Model of homo-oligomerisation of polydisperse proteins

Since the model of co-assembly is built on principles of homo-oligomerisation developed by Baldwin et al. ([34]), it seems fitting to introduce these principles first. Baldwin et al. compared multiple different models of oligomerisation (helical and splinter models, as well as their modifications) and found that the so-called 'modified helical polymerisation' fits the oligomeric distribution of HspB5 under different conditions (pH change that results in changes in even:odd oligomeric ratio) the best.

In this model, oligomerisation happens by subsequent gains or losses of monomers:

$\mathrm{P}_1 + \mathrm{P}_1 \underset{\mathrm{k}_2^-}{\overset{\mathrm{k}_2^+}{\rightleftharpoons}} \mathrm{P}_2$

$\mathrm{P}_2 + \mathrm{P}_1 \underset{\mathrm{k}_3^-}{\overset{\mathrm{k}_3^+}{\rightleftharpoons}} \mathrm{P}_3$

...

$\mathrm{P}_1 + \mathrm{P}_{i-1} \underset{\mathrm{k}_i^-}{\overset{\mathrm{k}_i^+}{\rightleftharpoons}} \mathrm{P}_i$

where P_1 is a free monomer in solution and P_i is an oligomer of size i.

It follows that the general rate equation for any oligomer of size i ($i \neq 0$) is then:

$$\frac{d[P_i]}{dt} = -k_i^-[P_i] + k_i^+[P_{i-1}][P_1] - k_{i+1}^+[P_i][P_1] + k_{i+1}^-[P_{i+1}] \tag{2.6}$$

and that of a free monomer, P_1:

$$\frac{d[P_1]}{dt} = \sum_2^n k_i^-[P_i] + k_2^-[P_2] - \sum_2^n k_i^+[P_{i-1}][P_1] + k_2^+[P_1]^2 \tag{2.7}$$

To satisfy both equations at equilibrium:

$$[P_i] = \frac{k_i^+}{k_i^-}[P_1][P_{i-1}] \tag{2.8}$$

The model uses a single on-rate, k^+ for the addition of monomers, independent of the oligomeric size, and assumes that all monomers in an oligomer prefer to pair up. To describe the fact that the odd-numbered oligomers are less stable than even-numbered ones, there are two off-rates: one for the loss of a monomer from a dimer, k_d^-, and one for the loss of a single monomer from an odd-numbered oligomer, k_m^- (Fig. 2.2).

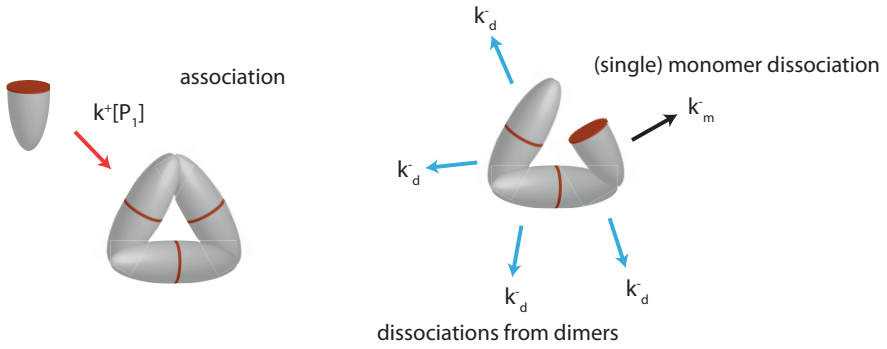


Figure 2.2: Explanation of on- and off-rates in the kinetic model of oligomerisation. In the kinetic model of protein assembly, there's only one association rate constant, k^+. Having two dissociation rates, k_d^- and k_m^-, allows to model the differential stability of even- and odd-numbered oligomers.

Substituting these new rate constants into Eq. 2.8, it is then possible to calculate the relative abundance of any even-numbered oligomer as:

$$[P_i]_{even} = \frac{k^+}{ik_d^-}[P_{i-1}][P_1] \tag{2.9}$$

and that of an odd-numbered oligomer as:

$$[P_i]_{odd} = \frac{k^+}{(i-1)k_d^- + k^- m}[P_{i-1}][P_1] \tag{2.10}$$

Assuming that the loss of a monomer from a dimer involves breaking both intra-dimer (dimer interface) and inter-dimer ('edge') interactions, whereas the loss of a single, odd monomer only involves breaking the 'edge' interaction, one can also calculate the associated Gibbs free energies of these processes as:

$$\Delta G_{edge+dimer} = G_{i,edge+dimer} - G_{i-1,edge-dimer} = -RTln\frac{k^+[P_1]}{k_d^-} \tag{2.11}$$

and

$$\Delta G_{edge} = G_{i,edge} - G_{i-1,edge} = -RTln\frac{k^+[P_1]}{k_m^-} \tag{2.12}$$

ΔG_{dimer} can then be calculated as:

$$\Delta G_{dimer} = \Delta G_{edge+dimer} - \Delta G_{edge} = -RTln\frac{k_m^-}{k_d^-} \tag{2.13}$$

These energies assume addition to and loss from a single, specific site within the oligomer, meaning they are independent of the oligomeric size.

2.3 Results & discussion

2.3.1 Description of co-assembly between two monodisperse proteins using statistical mechanics

Since $\Delta\Delta G_i$ is inherently dependent on both the subunit composition and their arrangement, and the number of possible arrangements increases exponentially, we decided to first use a mean-field model for the subunit arrangement for simplicity:

$$\Delta\Delta G_k^i = \frac{i(i-1)\Delta G_{AA} + (N-i)(N-i-1)\Delta G_{BB} + 2\Delta G_{AB} i(N-i)}{N-1} \tag{2.14}$$

where N is the oligomeric size, and i is the number of i subunits in the N-mer. Note that this function is independent of subunit arrangement, k.

Setting $\Delta G_{AB} = 0$, and $\Delta G_{AA} = \Delta G_{BB} = \Delta G_{homo}$, the above equation simplifies to

$$\Delta\Delta G_k^i = \Delta G_{homo}\frac{i(i-1) + (N-i)(N-i-1)}{N-1} \tag{2.15}$$

And equation 2.3 simplifies to:

$$x_i = a^i b^{N-i} \binom{N}{i} e^{-\Delta G_{homo}(i(i-1)+(N-i)(N-i-1))/RT(N-1)} \tag{2.16}$$

We then calculated the equilibrium distributions for preferential co- and self-assemblies, as well as the case of no energy bias between co- and self-assembly for two 20-mers (Fig. 2.3) to show how this arrangement-independent energy function as well as the mixing ratio affect the final equilibrium distributions.

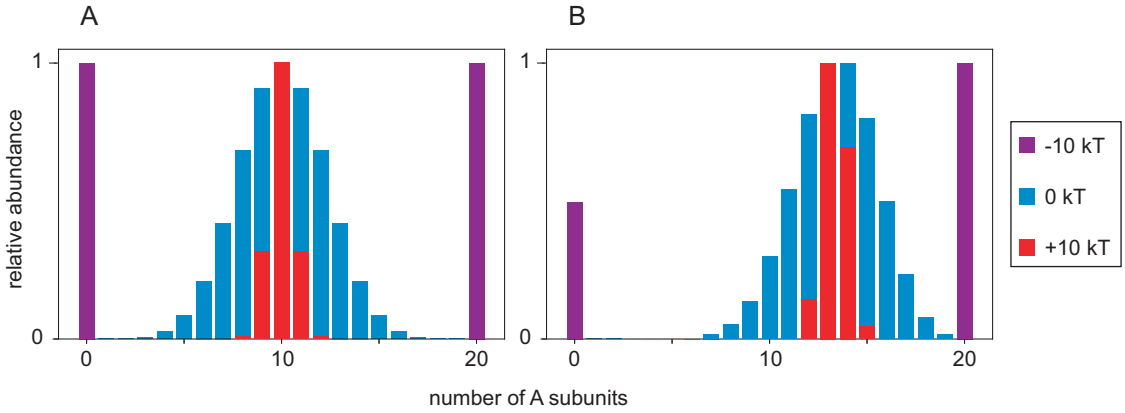


Figure 2.3: Comparison between assembly modes in an equal (A) and unequal (B) mixture in a 20-mer. Purple distributions correspond to a 10 kT bias for self-assembly, there is no bias for either self- or co-assembly in blue distributions, and the red distributions correspond to a 10 kT bias for co-assembly. For (A), $a = b = 1$ for all three energy differences. For (B), to keep the 1:2 mixing ratio constant, a and b had to be optimised, with a:b ratios of 1.036:1, 1:2, and 2200:1 for purple, blue and red distributions, respectively.

Curiously, it seems that, in terms of the energy difference between hetero- and homo-interactions, the same absolute bias seems to completely disrupt hetero-oligomerisation in the case where homo-interactions are stronger. In the case where hetero-interactions are stronger, this energy difference is not enough to cause a complete, obligate n:n co-assembly in the case of equal mixing, nor its equivalent for the unequal mixture. To further explore how the assembly space varies with the hetero- and homo-interaction energy difference, we have scanned over the whole range of energies between -10 kT and 10 kT (Fig. 2.4).

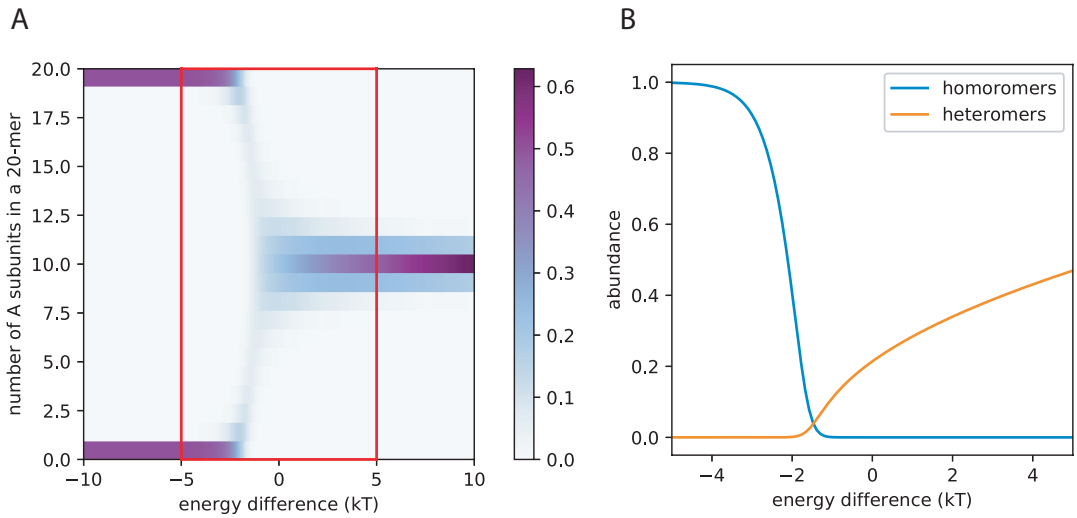


Figure 2.4: 20-mer co-assembly in the mean field approximation. a) The assembly space as a function of energy bias. b) The fraction of pure homo-oligomers (blue) and obligate n:n hetero-oligomers (orange) as a function of energy difference between -5 kT and 5 kT.

It can be seen that below approximately -4 kT energy difference, the 20-mers seem to be completely self-assembled, whereas in the case of co-assembly, the same (positive) energy difference of about 4 kT is not enough to make even half of the co-assemblies complete, obligate n:n heteromers. Indeed, for an (almost) complete n:n co-assembly, defined as 98 percent oligomers being n:n obligate hetero-oligomers, the energy difference is roughly 78 kT. This is quite large, but it should be noted that these energies, even though they are the energy difference per interaction, also depend on the size of the oligomer in question. The energies scale roughly logarithmically with the oligomeric size, as shown in Fig. 2.6.

The high energy difference needed for n:n co-assembly comes from the arrangement-independent energy model. Since we assume a mean-field contribution to the hetero- and homo-interactions, this model works well in the extreme cases of self-assembly and assembly without any preference for self- or co-assembly. But in the case of co-assembly, a mean-field approximation to the contributions for hetero- and homo-interactions does not work well, since it is reasonable to assume

that in the extreme case of obligate, n:n co-assembly with an approximately ring geometry, the only significantly populated arrangement would be of the form of $...A-B-A-B-A-B-A...$, where the number of hetero-interactions is maximised.

To further investigate what the ideal energy function should be, we look at the two extreme cases: self-assembly only, and obligate, $n:n$ co-assembly.

As shown in Fig. 2.5, the (absolute) energy difference should be the same for both cases, at least for an even number of subunits. That is, swapping one black subunit for a white subunit in both the case of purely self-assembled and n:n co-assembled subunit results in the swap of two homo- and two hetero-interactions, and vice versa, respectively. This contradicts the predictions above.

Since the pure self-assemblies have a 'multiplicity' of 1 (no matter how many steps we rotate the ring-like oligomer, we still end up with the same arrangement) and even n:n hetero-oligomers have a 'multiplicity' of two (for an even N-mer, we can rotate by $2\pi/N$ to arrive at a different arrangement), the energy comparison is not trivial. Hochberg et al. defined the 'demixing' energy as the energy needed to keep 98 percent of the oligomeric species as pure homo-oligomers, with 2 percent mixed oligomers. Depending on how we define the oligomeric composition for the equivalent case for obligate, $n:n$ co-assembly, we find that the energy difference is either the same, or slightly lower for an even $n:n$ co-assembly compared to self-assembly (Fig. 2.5).

We thus need to find a model that satisfies the following three conditions:

1. For no energy difference between homo- and hetero-oligomerisation, the resulting sub-stoichiometry distribution should be the binomial distribution.

2. For the cases of extreme self- and co-assembly, the energies should correspond to those we arrived at in Fig. 2.6

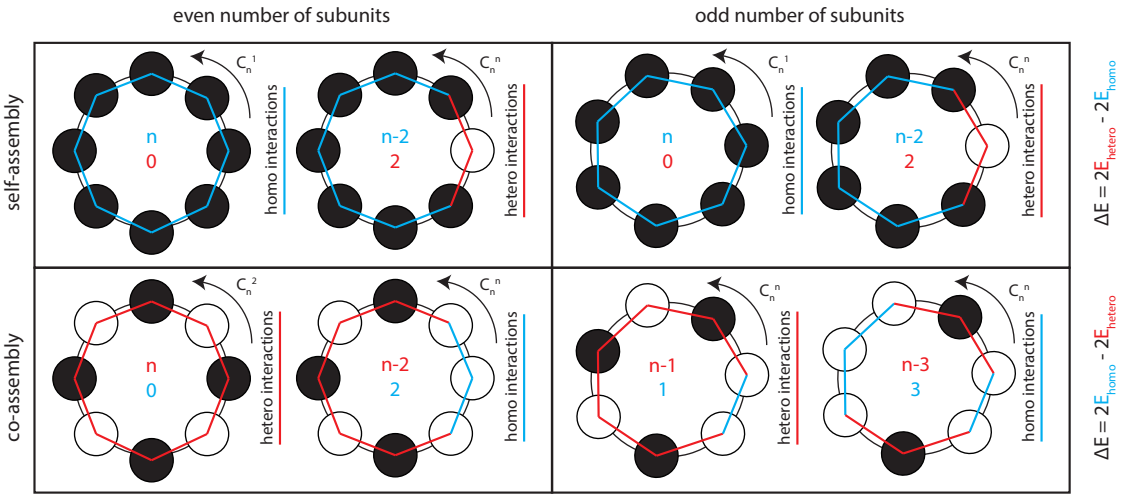


Figure 2.5: Changes of 'multiplicity' upon substitutions. Substitution of a single black subunit for a white subunit results in changes in symmetry, which also changes the 'multiplicity' of the given configuration. In absolute terms, the energy associated with substituting one subunit in an even, purely self-assembled oligomer is the same as doing the same substitution in an even obligate $n : n$-mer. The situation is more complex for odd oligomers where the energy changes are the same, but the contribution from symmetry is generally different.

3. In the extreme case of pure self-assembly, we should be able to reproduce Hochberg's results.

To begin, we decided we needed a better counting method for the different subunit arrangements, and their 'multiplicities'. To do this, we use the following algorithm in which we treat the ring arrangements of an N-mer as a string of the form -$AABBA$- (as an example) that wraps around itself, so the first and last monomers interact with each other. Each such string can be treated as an integer composition of N. For example: for $N = 5$, the following string -$AABBA$- can be expressed as 2+2+1. Note that -$BBAAB$- would be represented by the same expression. The algorithm itself does the following:

1. Start with a monomer, $\{1\}$. $\{1\}$ in this case represents both single subunits, A and B.

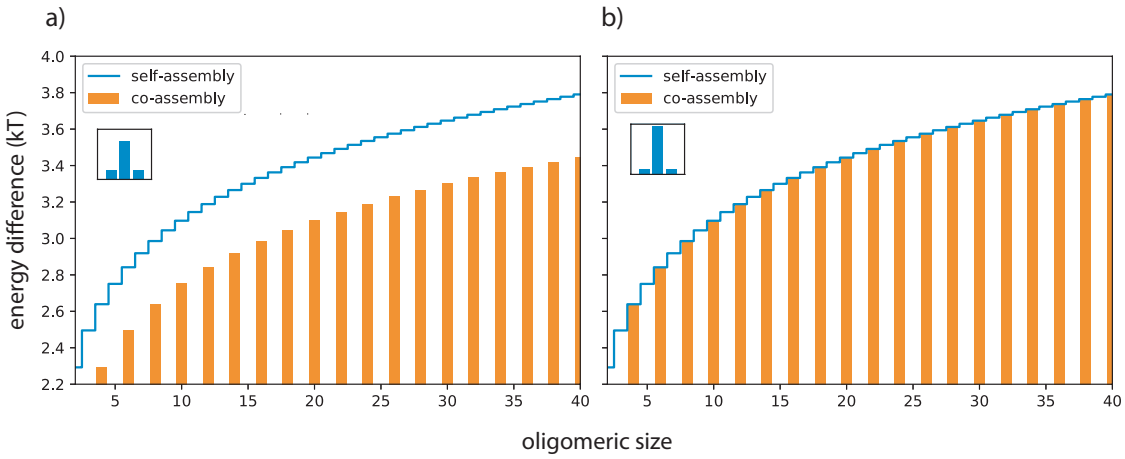


Figure 2.6: Using multiplicities to calculate the energy biases needed for complete self-assembly or obligate, n:n-co-assembly. With the reasoning established in Fig. 2.5, it is possible to calculate the energy needed so that only 2 percent of species form (n-1):1-mer (blue). It is also possible to calculate the corresponding energy for even (n-1):(n+1)-mers relative to n:n-mers. Depending on how we define the exact fraction, the energy biases needed are either the same or slightly smaller (in absolute terms). In a) the ratios are 2:96:2 and in b) the ratios are 1:98:1 of (n-1):(n+1)-mer, n:n-mer, and (n+1):(n-1)-mer. This is because in the case of co-assembly, the definition of abundance used for self-assembly is ambiguous.

2. To find a composition of all possible dimers, take the composition of the monomer, {1} and either add 1 to the first term or prepend {1}. This results in {2} and {1,1}. Which represent AA (and BB), and AB (and BA) respectively.

3. Similarly, for trimers, we have: {3}, {1,2}, {2,1}, and {1,1,1}. These represent AAA (BBB), ABB (BAA), AAB (BBA) and ABA (BAB), respectively.

4. The same arguments can then be used to find all compositions of an N-mer from the known compositions of an $(N-1)$-mer.

Once we find the multiplicities using the algorithm above, finding the corresponding energy term is quite trivial. For example, for a {3, 1, 2, 1} composition, we have $\Delta\Delta G_k^i = 2\Delta G_{AA} + \Delta G_{AB} + \Delta G_{AB} + \Delta G_{AA} + \Delta G_{AB} + \Delta G_{AB} = 3\Delta G_{AA} + 4\Delta G_{AB}$, and associated multiplicity of 7.

To test that this model satisfies the three conditions set above, we first simulated an 'unbiased' co-assembly in 12-mers. The resulting distribution was indeed binomial (Fig. 2.7 a). We then scanned through the energy bias space from -5 to 5 kT (2.7 b and c). In b), we can see the normalised relative abundance of pure 12:0-homo- and pure 6:6-hetero-oligomers (blue and orange bars, respectively) as a function of energy bias, similar to Fig. 2.4 b). We can, indeed, see that this distribution is roughly symmetric around the 0 kT point, with proteins practically completely self-assembled or n:n co-assembled below -4 kT and above 4 kT energy difference respectively. Fig. 2.7 c) shows the normalised abundance of each of the sub-stoichiometries as a function of energy.

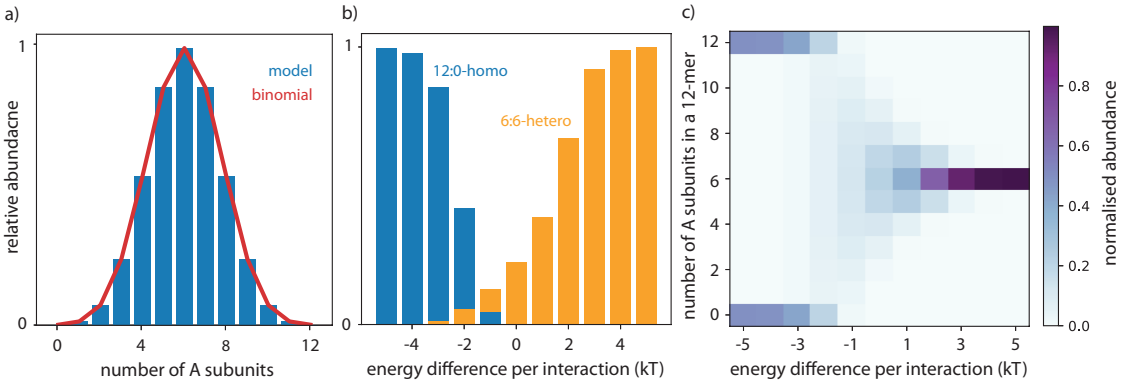


Figure 2.7: 12-mer co-assembly predicted using the 'counting method'. a) The 'counting method' recovers a binomial distribution of non-preferential co-assembly. b) The relative abundance of pure homo-oligomers (orange) and pure 6:6-hetero-oligomer as a function of energy bias is roughly symmetric. 3) The whole assembly space for a 12-mer (showing all sub-stoichiometries) as a function of energy bias.

There is a slight energy bias towards self-assembly in Hsp21-Hsp26.5 mixtures

To put the model to test, we mixed two monodisperse small heat-shock protein 12-mers from *Arabidopsis thaliana*.

Since the model allows us to simulate unequal mixtures, we mixed the two proteins at three different ratios, 2:1, 1:1, and 1:2, and recorded native MS spectra

which we were able to deconvolve and assign sub-stoichiometries to the deconvolved masses as shown in Fig. 2.8. We first assigned the charge series coming from each individual mass as shown on the left. The assigned masses can be seen on the right. Since we know the mass of both sub-units, it is then trivial to assign the sub-stoichiometries. The spectra shown are for the 1:1 mixing ratio.

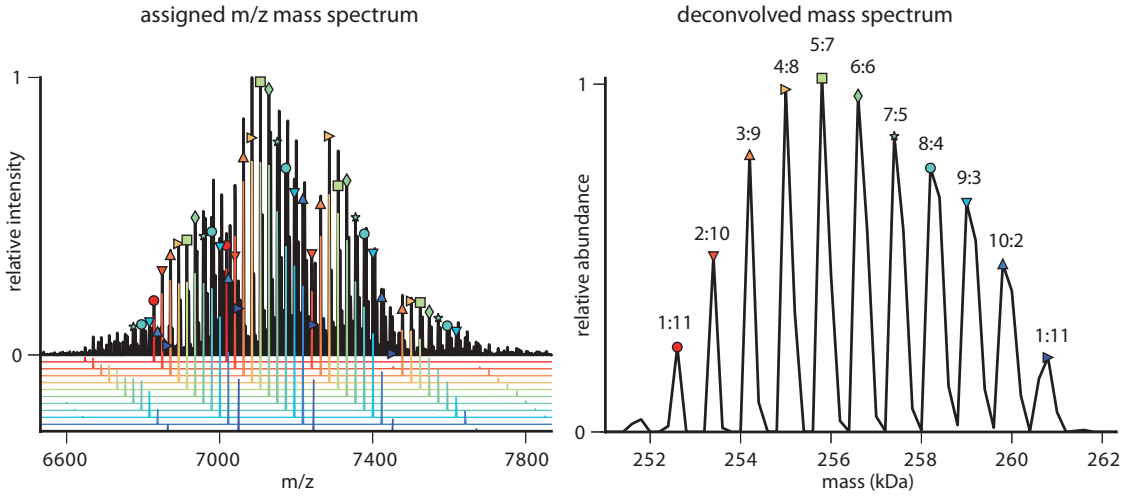


Figure 2.8: The sub-stoichiometry assignment workflow. (Left) Charge series coming from different masses (shown in colour where different colours correspond to different masses) are assigned in the native mass spectrum (black). We can then use the assigned masses to calculate the stoichiometry that results in any particular mass (right), and extract their abundances.

We then used the model to extract the energy bias between hetero- and homo-interactions by fitting the energy difference to all three mixing ratios simultaneously. We allowed a and b to vary for the different mixing ratios. The result can be seen in Fig. 2.9. For simplicity, we assumed that $\Delta G_{AA} = \Delta G_{BB} = 0$, which means that the energy difference is then simply ΔG_{AB}.

The extracted energy difference between the hetero- and homo-interactions was 1.98 kJ/mol of interactions. This means that the hetero-interactions are slightly disfavoured in these mixtures compared to homo-interactions. This energy is on the scale of a weak hydrogen bond ([90]). It is quite curious that even though the two proteins are separated in the cell–one is in the mitochondria while the other

is in the chloroplast–the two proteins still co-assemble without a large preference for neither self- nor co-assembly.

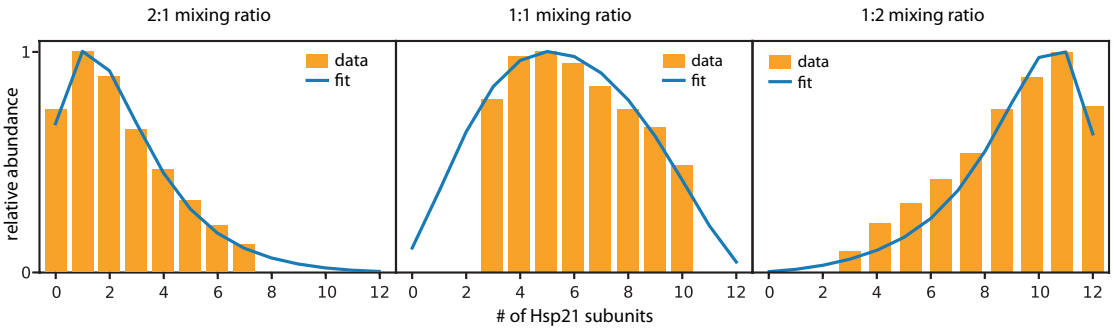


Figure 2.9: The experimental (orange) and fitted (blue) oligomeric distribution for the Hsp21-Hsp26.5 mixtures for the three mixing ratios. The extracted energy bias was 1.98 kJ/mol of interactions meaning that the homo-interactions are more stable compared to the hetero-interactions, but the difference is still only on the order of the strength of a weak hydrogen bond.

2.3.2 Description of co-assembly between two highly polydisperse proteins using chemical kinetics

Building on the previous chemical-kinetical model of homo-oligomerisation (Section 2.2.2, [34]), to describe hetero-oligomerisation, we need to introduce additional rate constants. Assuming, again, that the monomer addition is size-independent, we introduce two on-rates. The fact that the on-rates are size-independent can be explained by the fact that physically, the initial binding of a free monomer to an oligomer is the rate-limiting step, followed by rapid internal rearrangement. This size independence of the initial binding steps could also be explained by there only being one particular binding 'pocket' within the oligomer. The result of the on-rate being independent off size of the oligomer is the roughly Gaussian distribution of sizes rather than an infinite growth. HspB1 and HspB5 hetero-oligomers, the primary experimental target of this model, are composed of homo- and hetero-dimers ([91] and our data as shown in the next chapter). To account for this, we introduce four dimer-off-rates. Finally, since our experimental data suggest

that odd and even oligomeric sizes are not equally populated, we introduce two off-rate constants for single subunits (Fig. 2.10).

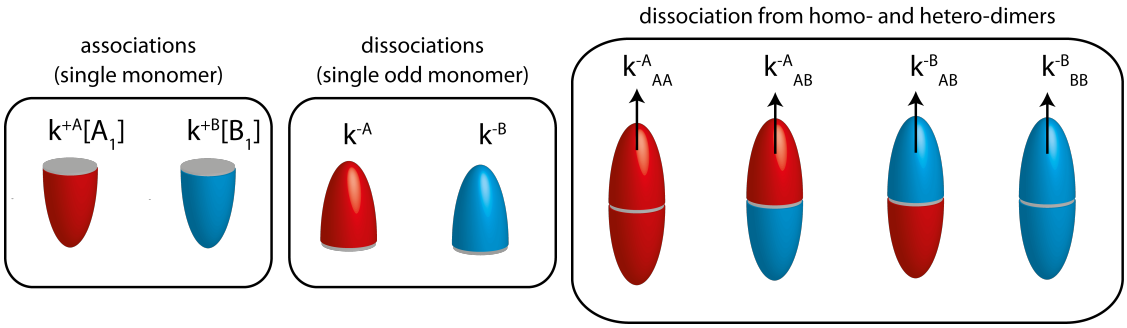


Figure 2.10: On- and off-rates in the kinetic model of sHsp oligomerisation. In the simplest case of the kinetic model of co-assembly, there are two monomer on-rates, two off-rates for single, unpaired monomers, are four off-rates for the losses of A or B monomers from homo- and hetero-dimers.

In Fig. 2.10, k^{+A} and k^{+B} are the on-rates for species A and B respectively, k^{-A} and k^{-B} are the off-rates of a single, odd monomer for species A and B respectively, k_{AA}^{-A} corresponds to a loss of A from an (AA)-homo-dimer, and k_{BB}^{-B} corresponds to a loss of B from an (BB)-homo-dimer. k_{AB}^{-A} and k_{AB}^{-B} are the off-rates from hetero-dimers for species A and B respectively.

2.3.3 Rate equations and the 'kinetic solution'

To ensure that the eight rate constants are able to completely describe hetero-co-assembly, as well as to build the general rate equations for the time dependence of the concentration of any arrangement, we built a kinetic 'map' (Fig. 2.11) that shows all the possible substitutions for a generic $(AA)_{aa}(BB)_{bb}(AB)_{ab}(X)$ arrangement, where X can either be 'empty' in the case of even oligomers, or a single A or B subunit in odd oligomeric sizes.

From this, it is possible to obtain the general rate equation for a $(AA)_{aa}(BB)_{bb}(AB)_{ab}$ arrangement:

$$v = -\frac{d[(AA)_{aa}(BB)_{bb}(AB)_{ab}]}{dt} =$$

$$[(ab)(k_{AB}^{-B} + k_{AB}^{-A}) + 2(aa)k_{AA}^{-A} + 2(bb)k_{BB}^{-B} + k^{+A} + k^{+B}][(AA)_{aa}(BB)_{bb}(AB)_{ab}]$$

$$- k^{+B}([(AA)_{aa}(BB)_{bb}(AB)_{ab-1}(A)] + [(AA)_{aa}(BB)_{bb-1}(AB)_{ab}(B)])$$

$$- k^{+A}([(AA)_{aa}(BB)_{bb}(AB)_{ab-1}(B)] + [(AA)_{aa-1}(BB)_{bb}(AB)_{ab}(A)])$$

$$- [2(bb)k_{BB}^{-B} + (ab)k_{AB}^{-B} + k^{-B}][(AA)_{aa}(BB)_{bb}(AB)_{ab}(B)]$$

$$- [2(aa)k_{AA}^{-A} + (ab)k_{AB}^{-A} + k^{-A}][(AA)_{aa}(BB)_{bb}(AB)_{ab}(A)]$$

The rate equations for the case of odd oligomers can be found in a similar fashion.

We can use the Euler method to find the equilibrium distribution by substituting dt for Δt (Eq. 2.17), and iterate through all possible dimer arrangements until the solution reaches an equilibrium which we defined as the point at which the change in relative abundances of all sub-stoichiometries separated by 100 steps in the simulation was smaller than $10^{-5}\%$. Δt was chosen small enough so that the method would not 'overshoot' the equilibrium solution, but also large enough so that the simulation finished in a reasonable time period. For the rate constant used, the size of Δt was generally $10^{-3} < \Delta t < 10^{-4}$.

$$[(AA)_{aa}(BB)_{bb}(AB)_{ab}]_{(t=\Delta t)} = [(AA)_{aa}(BB)_{bb}(AB)_{ab}]_{(t=0)} - v_{(t=0)}\Delta t \qquad (2.17)$$

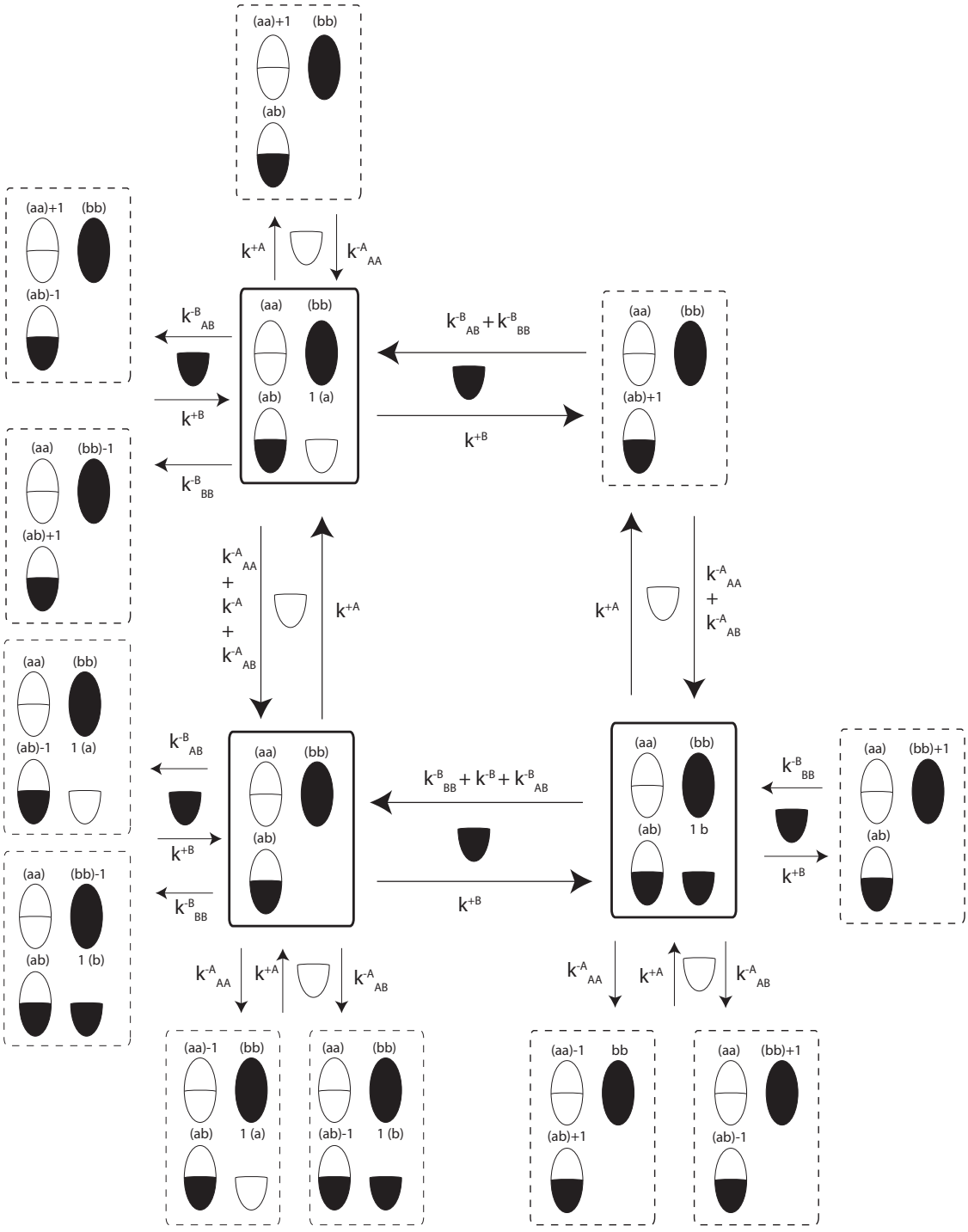


Figure 2.11: Any $(AA)_{aa}(BB)_{bb}(AB)_{ab}(X)$ combination can be built by subsequent additions and/or loses of (A) and (B) monomers. For simplicity, we have omitted any multipliers in front of the rate constants. (X) can either be 'empty' in the case of an even oligomer, or a single (A)or (B) in the case of odd oligomers.

Counting up the arrangements that contribute to a single sub-stoichiometry

Since the model is developed in terms of homo- and hetero-dimer arrangement, rather than in terms of sub-stoichiometries themselves, we needed to devise an algorithm to count all the arrangements that contribute to each specific sub-stoichiometry.

For even sub-stoichiometries (A_aB_b-mer where $a + b$ is even):

Step 1: Compare a and b and find the smaller of the pair.

Step 2: If $a < b$:

(a) The starting arrangement is $(BB)_{(b-a)/2}(AB)_a$

(b) Iterate through all possible arrangements by substituting $2(AB)$ for $1(AA) + 1(BB)$ until the number of (AB) dimers in the arrangement is either one or there is none left. Add the abundances together.

Step 3: If $b < a$:

(a) The starting arrangement is $(AA)_{(a-b)/2}(AB)_a$

(b) Iterate through all possible arrangements by substituting $2(AB)$ for $1(AA) + 1(BB)$ until the number of (AB) dimers in the arrangement is either one or there is none left. Add the abundances together.

Step 4: Return the sum of abundances of the arrangements iterated through.

The situation is slightly more complicated for odd oligomers, where there could be both either a single A subunit, or a single B subunit.

In the case where there is an odd number of A subunits, and an even number of B subunits, and assuming $a < b$:

Step 1: Assuming $b > 0$, start with arrangements with a single (B) subunit.

(a) The starting arrangement is $(BB)_{(b-a-1)/2}(AB)_a(B)$ - note that this arrangement has a single B subunit.

(b) Iterate through all possible arrangements by substituting $2(AB)$ for $1(AA) + 1(BB)$ until the number of (AB) dimers in the arrangement is either one or there is none left. Add the abundances together.

Step 2: Then iterate through all arrangements with a single (A) subunit.

(a) The starting arrangement is $(BB)_{(b-a+1)/2}(AB)_{(a-1)}(A)$ - note that this arrangement has a single A subunit.

(b) Iterate through all possible arrangements by substituting $2(AB)$ for $1(AA) + 1(BB)$ until the number of (AB) dimers in the arrangement is either one or there is none left. Add the abundances together.

The other three combinations, where b is odd and/or $b > a$ follow the same reasoning. This is why we, perhaps confusingly, check that $b > 0$ and $a > 0$ in the first and second step, respectively, as the same algorithm is applied to oligomers where there are only A or only B subunits, in which case the respective steps are skipped.

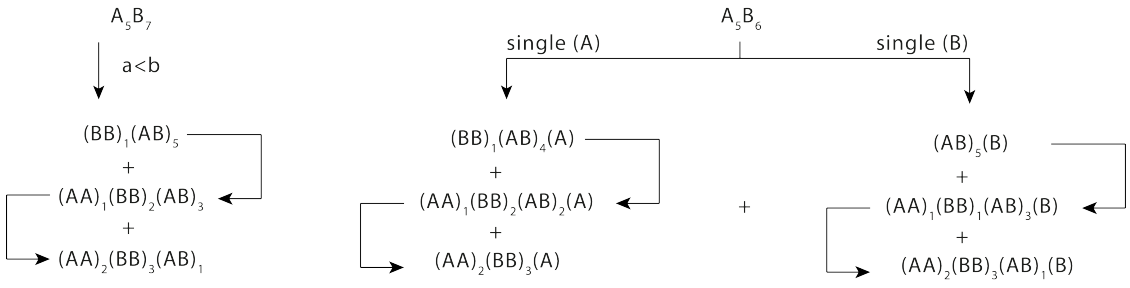


Figure 2.12: The algorithm for finding all arrangements for a particular sub-stoichiometry. As examples, we show an even stoichiometry on the left and an odd one on the right.

While we are not yet able to experimentally probe dimeric arrangements, the modelling allows us to obtain information about these arrangements as shown in Fig. 2.13.

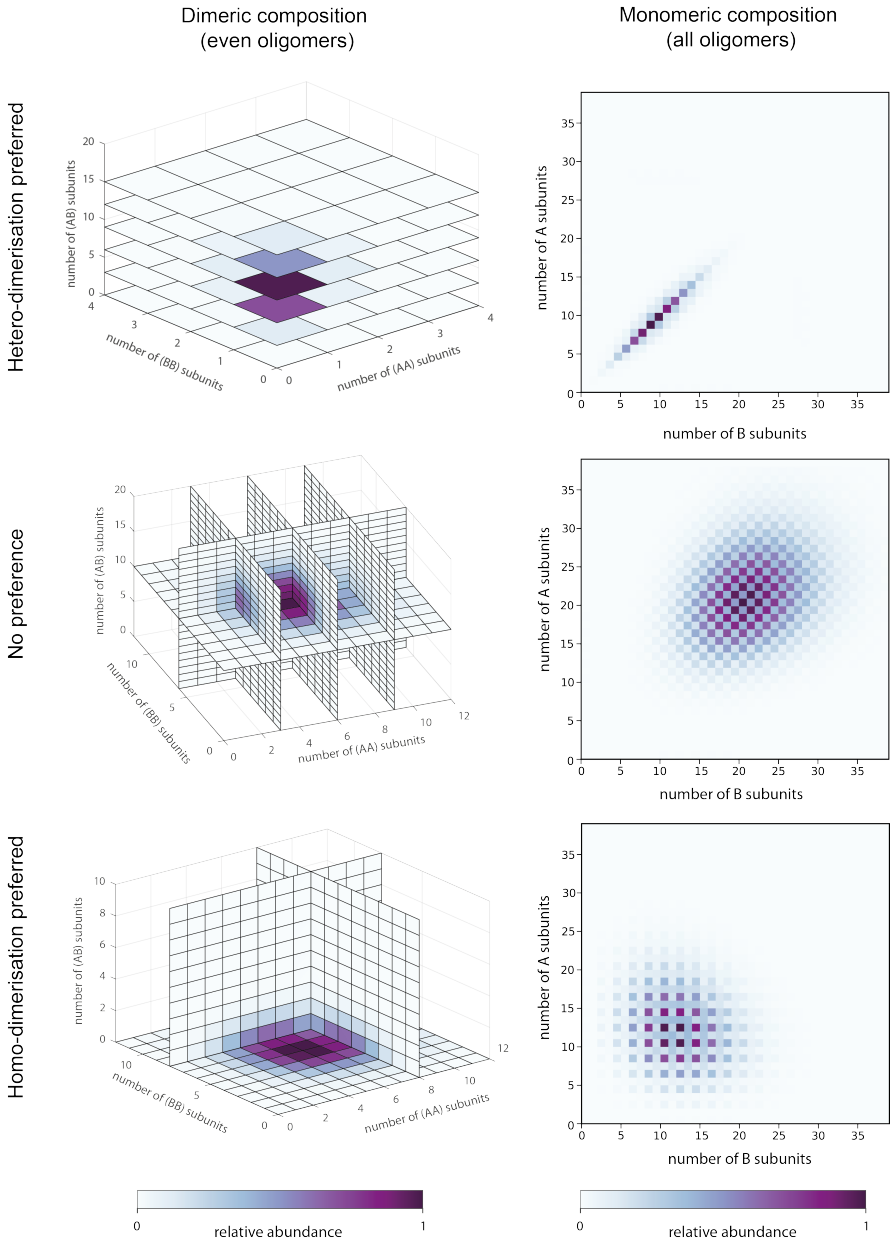


Figure 2.13: Dimeric arrangements and their corresponding '2D histograms'. The dimeric arrangements (shown as slices through a 3-dimensional distribution) of obligate n:n co-assembly (top) show the oligomers are composed almost exclusively of (AB)-hetero-dimers. For 'unbiased', non-preferential co-assembly (middle), there is a contribution from both hetero- and homo-dimers. Finally, preferential homo-dimerisation (bottom) shows only (AA)- or (BB)-homo-dimers occupied to a significant extent. The homo-dimers themselves can still co-assemble into higher oligomers as shown in the 2D histogram (bottom right).

The kinetic solution allows us to follow the co-assembly process

Assuming that the purely homo-meric on- (k^{+A} and k^{+B}) and off- (k^{-A}_{AA} and k^{-B}_{BB})-rates are the same before and after the proteins are mixed, we can build up homo-oligomeric distributions and let them 'kinetically' co-assemble (dependent on k^{-A}_{AB} and k^{-B}_{AB}) using the Euler method as described in Chapter 1 and above.

As a proof of concept, in Fig. 2.14 we show a system with two different species with the same oligomeric distributions (shown along the x- and y-axes at step 0) and follow their co-assembly process. In the lower plot, we show the normalised abundance of the different 20-mer sub-stoichiometries as a function of simulation steps. 20-mers were chosen because it is the most abundant stoichiometry in the simulation.

Since the co-assembly process is symmetric along the minor diagonal, we only show the relative abundance of half of the sub-stoichiometries, as e.g. the abundances of the 12 : 8 and 8 : 12 sub-stoichiometries are identical.

Initially, only the homo-oligomers are present, corresponding to the 0 : 20 sub-stoichiometry for 20-mers. As the co-assembly process progresses, more 'co-assembled' stoichiometries get transiently occupied until the simulation reaches a relatively stable sub-stoichiometry distribution, very close to the equilibrium distribution. As expected for unbiased co-assembly, the final sub-stoichiometry distribution is binomial.

2.3.4 The equilibrium solution

Since the Gibbs energy of the formation of an oligomer consisting of any combination of (AA)- and (BB)-homo-dimers, and (AB)-hetero-dimers should be path-independent, we can start building any oligomeric arrangement first with

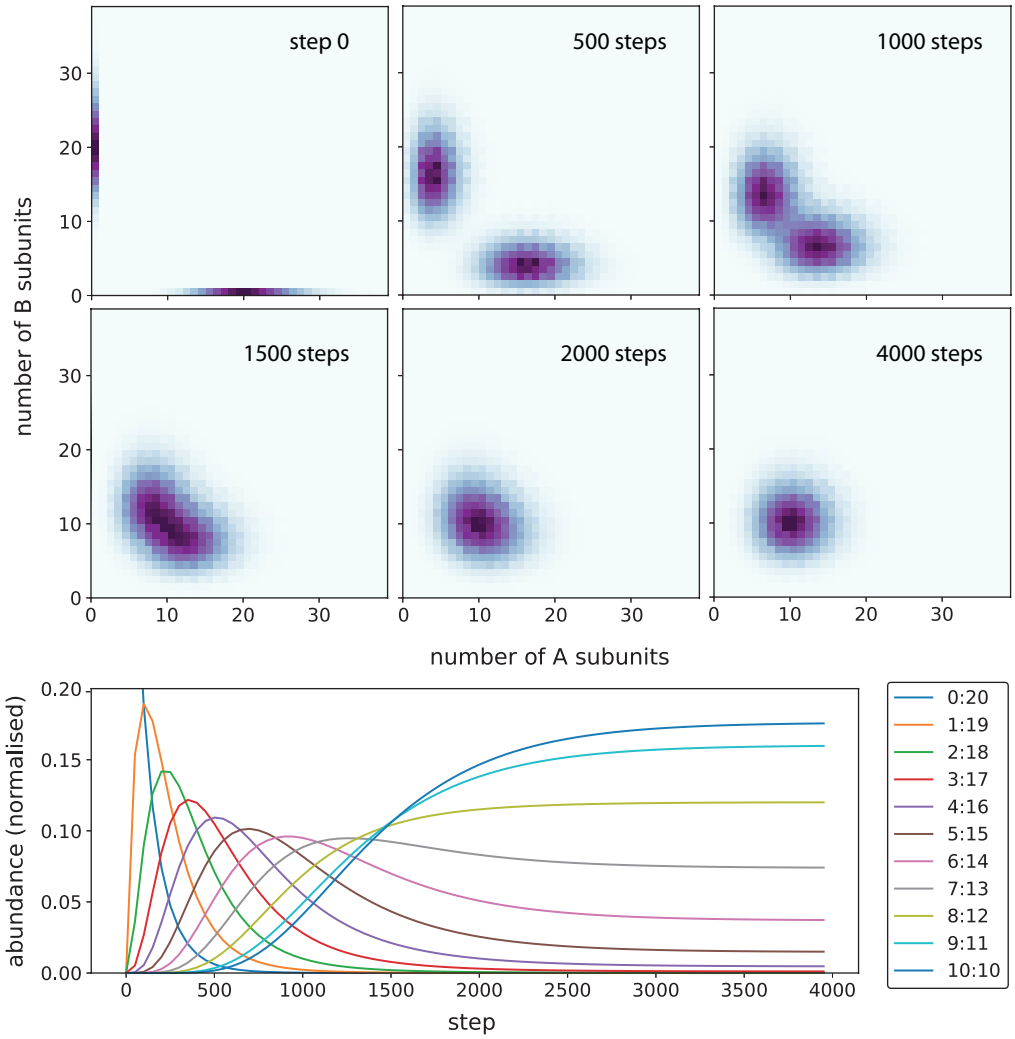


Figure 2.14: The kinetics of co-assembly. Upper: Oligomer co-assembly 'followed' using the Euler method from two purely self-assembled species up to the point that shows the resulting non-preferential co-assembly. Lower: The normalised abundances of different 20-mer sub-stoichiometries as a function of the number of steps show that as the simulation reaches equilibrium, the relative abundances stay relatively constant.

purely (AB)-hetero-dimers, and then swap (AB)-hetero-dimers for either of the two homo-dimers to reach the final desired composition.

First, we assume the two species, A and B, hetero-dimerise completely–that is, they only form (AB) dimers. This way, we can build a simple distribution as shown in Fig. 2.15.

It should be noted that since the order of monomer addition does not matter

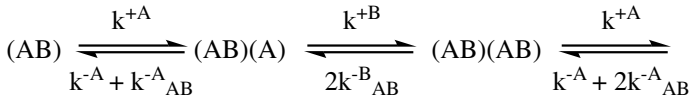


Figure 2.15: Obligate (AB)-hetero-oligomers can be built by subsequent addition of A and B subunits.

(Figure 2.16), k_{AB}^{-A} and k_{AB}^{-B} are dependent:

$$k_{AB}^{-A} = \frac{k^{-A}}{k^{-B}} k_{AB}^{-B} \tag{2.18}$$

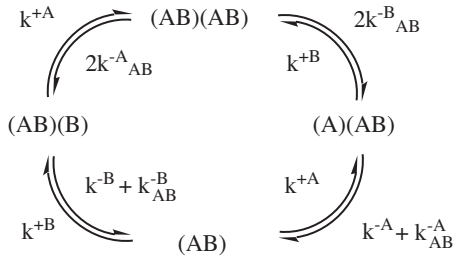


Figure 2.16: Hetero-oligomer formation is path-independent.

We can then obtain the relative equilibrium abundance (w.r.t. $[(AB)]$, the equilibrium abundance of a single (AB)-hetero-dimer) of any higher obligate (AB)-hetero-oligomer, $[(AB)_N]$, as:

$$[(AB)_N] = \frac{1}{N! \prod_{x=0}^{N-1} (x k_{AB}^{-A} + k^{-A})} \left(\frac{k^{+A} k^{+B}}{k_{AB}^{-B}} \right)^{N-1} [(AB)] \tag{2.19}$$

To extend this for cases where there is not a clear preference for hetero-dimerisation, we can make the following substitutions:

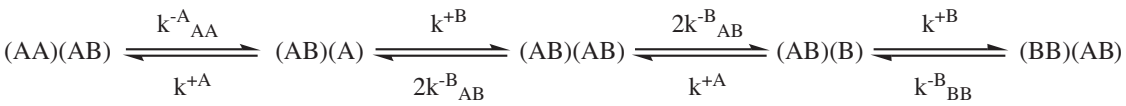


Figure 2.17: Example of substitutions needed to form homo-dimers within an oligomer.

Thus, given an equilibrium abundance, $[(AB)_N]$, of an 2N-mer composed solely of (AB)-hetero-dimers, we can calculate the relative abundance of any hetero-dimer, $[(AA)_{aa}(BB)_{bb}(AB)_{ab}]$, of the same size, 2N, as:

$$\frac{(aa+bb+ab)!}{aa!bb!ab!}\left(\frac{k_{AB}^{-B}k^{+A}}{k_{AA}^{-A}k^{+B}}\right)^{aa}\left(\frac{k_{AB}^{-A}k^{+B}}{k_{BB}^{-B}k^{+A}}\right)^{bb}[(AB)_N] \tag{2.20}$$

where $aa + bb + ab = N$

By combining Eq. 2.19 and Eq. 2.20, we can then calculate the equilibrium abundance of any $AA_{aa}BB_{bb}AB_{ab}$ hetero-oligomer as:

$$\frac{1}{aa!bb!ab!\prod_{x=0}^{aa+bb+ab-1}(xk_{AB}^{-A}+k^{-A})}\left(\frac{k^{+A}k^{+B}}{k_{AB}^{-B}}\right)^{aa+bb+ab-1}\left(\frac{k_{AB}^{-B}k^{+A}}{k_{AA}^{-A}k^{+B}}\right)^{aa}\left(\frac{k_{AB}^{-A}k^{+B}}{k_{BB}^{-B}k^{+A}}\right)^{bb}[(AB)] \tag{2.21}$$

Since we assume one off-rate from the homo-dimers, even though there are actually two (same) subunits in a homo-dimer, we need to make sure we multiply the homo-dimer off-rates by two before using the formula. To model assemblies where the mixing ratio is not exactly 1:1, we can multiply the two on-rates by r and $(1-r)$, where the product of these with their corresponding on-rates is proportional to their mixing ratio. We can calculate the mixing ratio, r_{mix} as:

$$r_{mix} = \frac{rk^{+A}}{(1-r)k^{+B}} \tag{2.22}$$

Note that, although we are using kinetic rates, the whole equilibrium solution is inherently time-independent, and thus corresponds to the 'kinetic' solution derived above in the infinite-time limit.

Abundance of odd oligomers

Equation 2.21 is only applicable to oligomers with an even number of subunits. In principle, there are two ways of forming an odd $[(AA)_{aa}(BB)_{bb}(AB)_{ab}(X)]$ oligomer, where X can be either A or B single monomer from an even one. They are either an addition of single X monomer to $[(AA)_{aa}(BB)_{bb}(AB)_{ab}]$, or a loss of X from a higher oligomer. For example, an $[(AA)_{aa}(BB)_{bb}(AB)_{ab}(A)]$ can be

formed by the loss of B from AB in $[(AA)_{aa}(BB)_{bb}(AB)_{ab+1}]$, or the loss of A from AB in $[(AA)_{aa+1}(BB)_{bb}(AB)_{ab}]$.

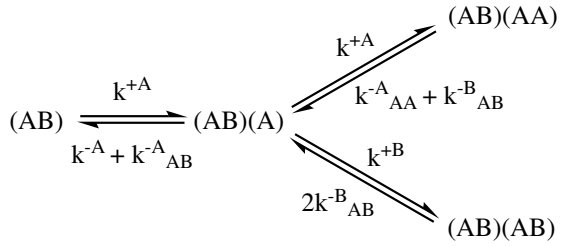


Figure 2.18: $(AB)(A)$ oligomer can be formed from either (AB), $(AB)(AB)$, or $(AB)(AA)$ oligomers.

We can thus obtain the relative equilibrium abundance of any odd oligomer (with a single A subunit) as:

$$[(AA)_{aa}(BB)_{bb}(AB)_{ab}(A)] = \begin{cases} \frac{k^{+A}}{k^{-A}+(aa)k^{-A}_{AA}+(ab)k^{-B}_{AB}}[(AA)_{aa}(BB)_{bb}(AB)_{ab}] \\ \frac{(aa+1)k^{-A}_{AA}+(ab)k^{-B}_{AB}}{k^{+A}}[(AA)_{aa+1}(BB)_{bb}(AB)_{ab}] \\ \frac{(aa)k^{-A}_{AA}+(ab+1)k^{-B}_{AB}}{k^{+B}}[(AA)_{aa}(BB)_{bb}(AB)_{ab+1}] \end{cases} \tag{2.23}$$

The same argument can be made for the formation of odd oligomers with a single B subunit.

$$[(AA)_{aa}(BB)_{bb}(AB)_{ab}(B)] = \begin{cases} \frac{k^{+B}}{k^{-B}+(bb)k^{-B}_{BB}+(ab)k^{-A}_{AB}}[(AA)_{aa}(BB)_{bb}(AB)_{ab}] \\ \frac{(bb+1)k^{-B}_{BB}+(ab)k^{-A}_{AB}}{k^{+B}}[(AA)_{aa}(BB)_{bb+1}(AB)_{ab}] \\ \frac{(bb)k^{-B}_{BB}+(ab+1)k^{-A}_{AB}}{k^{+A}}[(AA)_{aa}(BB)_{bb}(AB)_{ab+1}] \end{cases} \tag{2.24}$$

Approximation for systems biased towards even oligomers

From our experimental observations (shown in detail in the next chapter), the majority of known polydisperse sHsps, as well as their mixtures, exhibit a bias

towards even stoichiometries. This suggests that $k^{-A} >> k_{AB}^{-A}$ (and similarly $k^{-B} >> k_{AB}^{-B}$).

Taking this into account, if follows that:

$$\frac{1}{\prod_{x=0}^{aa+bb+ab-1}(xk_{AB}^{-A}+k^{-A})} = \frac{1}{\prod_{x=0}^{aa+bb+ab-1}(k^{-A})} = \frac{1}{(k^{-A})^{aa+bb+ab-1}} \tag{2.25}$$

And thus, Eq. 2.21 simplifies to:

$$\frac{1}{aa!bb!ab!}\left(\frac{k^{+A}k^{+B}}{k_{AB}^{-B}k^{-A}}\right)^{aa+bb+ab-1}\left(\frac{k_{AB}^{-B}k^{+A}}{k_{AA}^{-A}k^{+B}}\right)^{aa}\left(\frac{k_{AB}^{-A}k^{+B}}{k_{BB}^{-B}k^{+A}}\right)^{bb}[(AB)] \tag{2.26}$$

Calculating the equilibrium distribution using the simplified formula is less computationally expensive, and although for the majority of this work, we use the full formula in Eq. 2.21, the speed improvements offered by the simplification might be crucial in any future work with much bigger oligomeric sizes, or to calculate the equilibrium distributions for the mixing of more than 2 species–for which this model is fully extendable by introducing a third on-rate, and another set of off-rates.

Fig. 2.19 shows the difference in mass distributions obtained by the two formulas. For systems biased to even stoichiometries, the two distributions are well within any experimental error from MS.

We also use the simplified formula in the next section to find the formula for the most abundant oligomeric size.

Finding the most abundant oligomer

For self-assembling homo-oligomers, it is easy to estimate the most abundant stoichiometry. In the case when there is no even:odd preference, the distribution is simply the Poisson distribution with the maximum at $\frac{k^{+}[P_1]}{k_m^{-}}$.

Similarly, if $k_m^{-} >> k_d^{-}$ (as defined in Section 2.2.2), we can find the abundance of any even oligomer of the size $2N$ as roughly

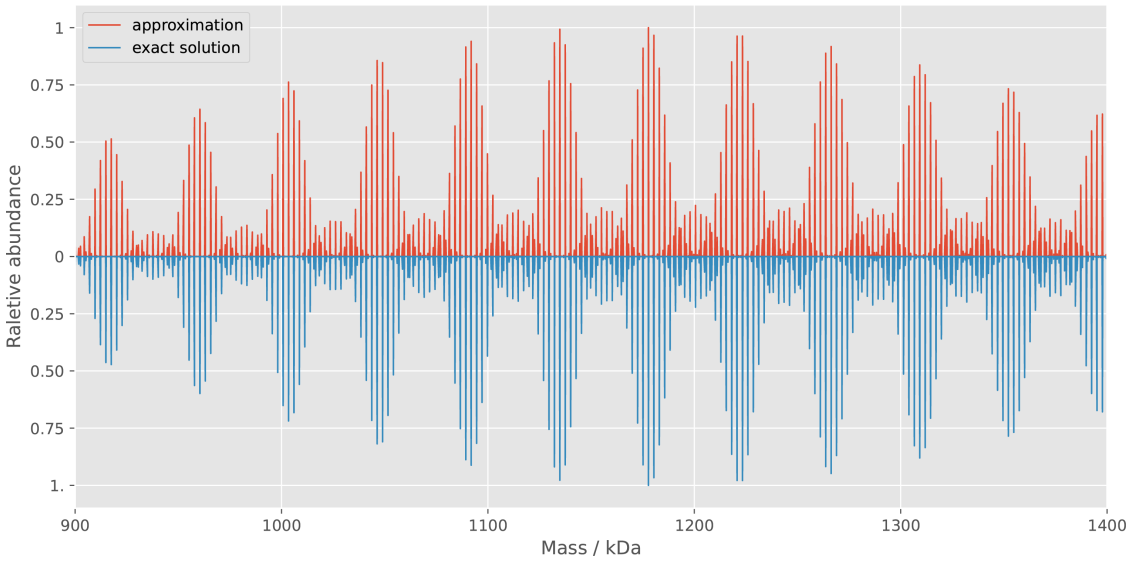


Figure 2.19: The exact and approximate solutions. Comparison of equilibrium distributions obtained using the approximation (red) and the exact solution (blue) using values for off-rates.

$$[P_{2N}] = \frac{1}{N!}\left(\frac{(k^+[M_1])^2}{k_d^- k_m^-}\right)^{N-1} \tag{2.27}$$

In this case, the function will have an inflection point (that corresponds to the maximum of that function) roughly at $\frac{(k^+[M_1])^2}{k_d^- * k_m^-}$ for reasons explained further below.

Since the contribution from different arrangements is very roughly symmetric around the most abundant arrangement, we can map all possible arrangements onto sub-stoichometries:

$$f : \sum[(AA)_{aa}(BB)_{bb}(AB)_{ab}]_{max} \longrightarrow [A_a B_b]_{max} \longrightarrow N_{max} \tag{2.28}$$

Thus, as an approximation, to find $[(A)_a(B)_b]_{max}$, it is enough to find $[(AA)_{aa}(BB)_{bb}(AB)_{ab}]_{max}$ such that $a = 2aa + ab$ and $b = 2bb + ab$.

Since we can rearrange the (simplified) formula for the relative abundance of a $[(AA)_{aa}(BB)_{bb}(AB)_{ab}]$-mer:

$$\frac{1}{aa!bb!ab!}\left(\frac{k^{+A}k^{+B}}{k_{AB}^{-B}k^{-A}}\right)^{aa+bb+ab-1}\left(\frac{k_{AB}^{-B}k^{+A}}{k_{AA}^{-A}k^{+B}}\right)^{aa}\left(\frac{k_{AB}^{-A}k^{+B}}{k_{BB}^{-B}k^{+A}}\right)^{bb}[(AB)] \tag{2.29}$$

into

$$\frac{\left(\frac{k^{+A}k^{+A}}{k_{AA}^{-A}k^{-A}}\right)^{aa}}{aa!}\frac{\left(\frac{k^{+B}k^{+B}}{k_{BB}^{-B}k^{-B}}\right)^{bb}}{bb!}\frac{\left(\frac{k^{+A}k^{+B}}{k_{AB}^{-A}k^{-B}}\right)^{ab}}{ab!} \tag{2.30}$$

This makes the problem somewhat separable as for

$$\frac{\partial[(AA)_{aa}(BB)_{bb}(AB)_{ab}]}{\partial aa}=0 \tag{2.31}$$

It is sufficient that

$$\frac{\partial\left(\frac{k^{+A}k^{+A}}{k_{AA}^{-A}k^{-A}}\right)^{aa}aa!^{-1}}{\partial aa}=0 \tag{2.32}$$

and so on for the (bb) and (ab) contributions as they are independent of each other.

Eq. 2.32 is not solvable analytically, as it is not smooth and continuous, but has a simple approximate analytical solution, as we know that any function of the form

$$f(x)=\frac{r^x}{x!} \tag{2.33}$$

will have a bell-shaped curve with a maximum roughly between $x = r - 1$ and $x = r$ as:

$$\frac{r^{r-1}}{r-1!}<\frac{r^r}{r!}=\frac{r^{r-1}}{r-1!}\frac{r}{r}>\frac{r^{r+1}}{r+1!}=\frac{r^r}{r!}\frac{r}{r+1} \tag{2.34}$$

Where r is a positive integer. By the same principle, it is possible to roughly estimate that the function

$$f(aa)=\frac{\left(\frac{k^{+A}k^{+A}}{k_{AA}^{-A}k^{-A}}\right)^{aa}}{aa!} \tag{2.35}$$

has an approximate maximum at

$$aa_{max}=\frac{k^{+A}k^{+A}}{k_{AA}^{-A}k^{-A}} \tag{2.36}$$

Similarly,

$$\begin{aligned} bb_{max} &= \frac{k^{+B}k^{+B}}{k_{BB}^{-B}k^{-B}} \\ ab_{max} &= \frac{k^{+A}k^{+B}}{k_{AB}^{-A}k^{-B}} \end{aligned} \tag{2.37}$$

We can then calculate the stoichometry of the most abundant oligomer, N_{max}, as:

$$N_{max} = r^2\left(\frac{k^{+A}k^{+A}}{k_{AA}^{-A}k^{-A}}\right) + (r-1)^2\left(\frac{k^{+B}k^{+B}}{k_{BB}^{-B}k^{-B}}\right) + 2r\,(1-r)\left(\frac{k^{+A}k^{+B}}{k_{AB}^{-A}k^{-B}}\right) \tag{2.38}$$

where r is the mixing ratio of protein a and protein b.

To explore both how well the oligomeric size of the most abundant arrangement and the predicted size correspond to the most abundant sub-stoichiometry, we simulated 1000 oligomeric distributions (with on- and off-rates chosen completely randomly). After filtering out distributions that did not have maxima within the range of oligomeric sizes we could simulate, we were left with 466 distributions. The agreement between both the oligomeric size of the most abundant arrangements and the predicted size, and the most abundant (simulated) sub-stoichiometry for 100 randomly chosen samples can be seen in Fig. 2.20. There seems to be an excellent agreement between all three. The Peasons' r coefficients were 0.997 and 0.996, and the 95 percent confidence interval for assignment error was (-2,2), and (-0.356, 2.582) for the oligomeric size of the most abundant arrangements and the predicted size and their correspondence to the most abundant sub-stoichiometry, respectively. These errors in the range of a couple of oligomeric sizes are expected as we're obtaining a real number for the predicted most abundant arrangement which, in reality, is of course an integer.

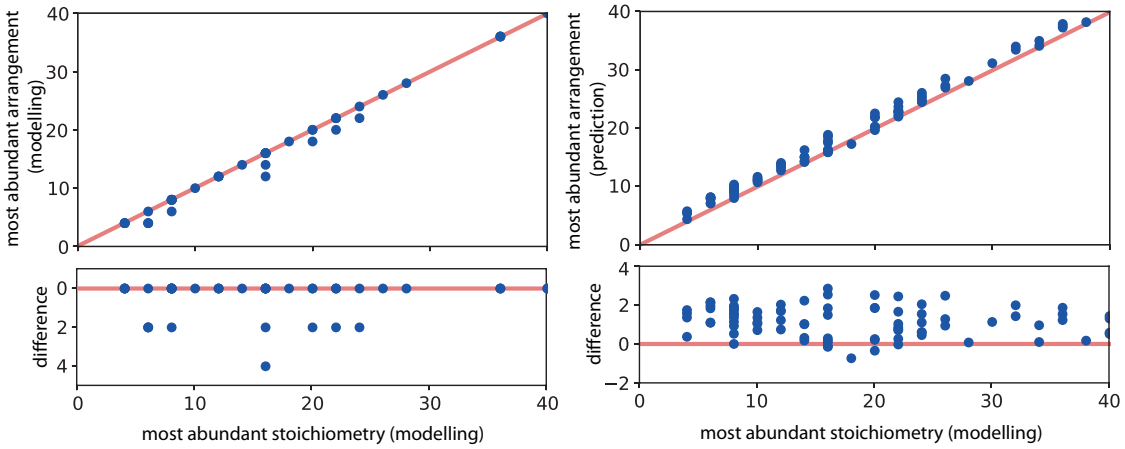


Figure 2.20: Comparison between the two methods for finding the most abundant sub-stoichiometry. a) There is a good agreement between the size of the most abundant dimer composition and the most abundant sub-stoichiometry (both were found by modelling). b) There also is a good agreement between the size of the most abundant dimer composition, as predicted by the derived formula, and the most abundant sub-stoichiometry. The red diagonal shows sizes for which the prediction matches the modelling exactly.

Rate correction for semi-assembling systems

The model does not make any assumption about how favourable it is for the hetero- or homo-dimers to co-assemble to form a bigger oligomer. While this assumption works well for favourable co-assembly, it gives a somewhat unintuitive result for cases of preferential self-assembly. One could expect two regions in the 2D histogram of preferential self-assembly, one that is A-rich and a B-rich one. This is not the case in this model. Without any correction for preferential homo-oligomerisation, the model predicts distributions such as the one in the bottom third of Fig. 2.13. This can be easily explained–the two species, A and B, don't form any (AB)-hetero-dimers, but their preferred (AA)- or (BB)-homo-dimers still co-assemble to form higher oligomers, which results in the 'checkered' pattern as shown in the figure, where there is a high abundance of A_aB_b species where both a and b are even. Hypothetically, this behaviour might be observed for example from the mixing of covalently-bound homo-dimers, which are not able to hetero-dimerise.

To correct for this, we can use the following modification to the off-rates:

$$\begin{bmatrix} k^{-A} \\ k_{AB}^{-A} \\ k_{AA}^{-A} \end{bmatrix} f(a,b) = \begin{bmatrix} k^{-Aw} \\ k_{AB}^{-Aw} \\ k_{AA}^{-Aw} \end{bmatrix}$$

and similarly

$$\begin{bmatrix} k^{-B} \\ k_{AB}^{-B} \\ k_{BB}^{-B} \end{bmatrix} g(a,b) = \begin{bmatrix} k^{-Bw} \\ k_{AB}^{-Bw} \\ k_{BB}^{-Bw} \end{bmatrix}$$

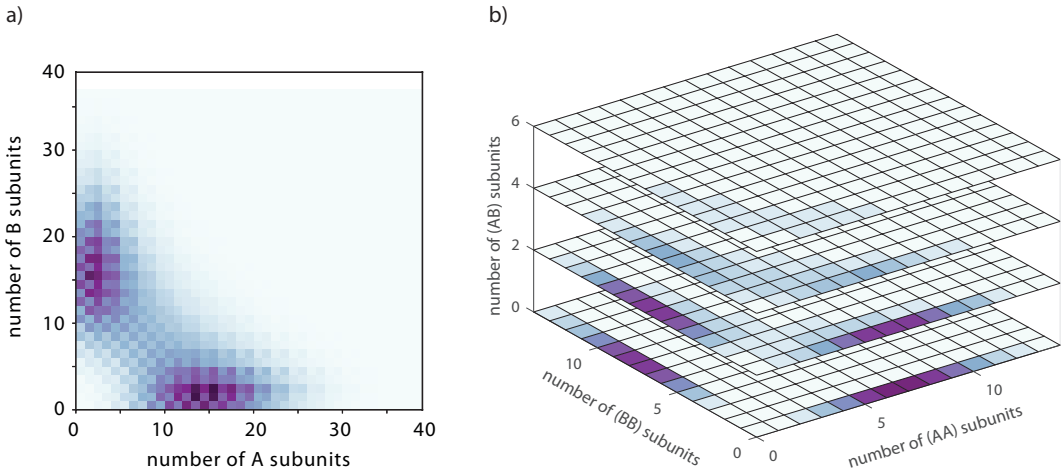


Figure 2.21: Correction for self-assembly. The assumption that off-rates are a function of internal composition allowed us to model co-assemblies close to pure self-assembly as can be seen in a). The dimer composition shows large contribution from pure (AA)- and (BB)-hetero-dimers, this time de-mixed with respect to each other.

Note that now the off-rates are a function of internal composition. If we set $f(a,b) = 1 + w_a \frac{b}{a+b}$ and $g(a,b) = 1 + w_b \frac{a}{a+b}$ as the corrective function to the A- and B-off-rates, respectively, we are able to obtain distributions such as in Fig. 2.21. In the formulas for $f(a,b)$ and $g(a,b)$, a and b is the number of A and B subunits in a given arrangement, respectively, and w_a and w_b are constants that describe how unfavourable co-assembly is. Note that if $w_a = w_b = 0$, $f(a,b) = 0$ and the off-rates are reduced to the rates used in the previous sections.

There are, of course, other functions of internal composition one could consider, but since we don't yet have good data for polydisperse systems that co-assemble only partially, we use this simple linear combination approach here as a proof that

the model is extendable to include partial co-assembly, should this be needed for future data analysis.

The most abundant oligomeric size is a function of the mixing ratio and relative interface stability

We can use the formula derived above (Eq. 2.37) to explore how the most abundant oligomeric size (as a rough proxy for the distribution size) varies with both the mixing ratio and the energy bias. Here, we are specifically interested in co-assembly without any energy bias, or a bias towards co-assembly, as we have systems to probe these modes of co-assembly (Chapters 3, 4 & 5).

The results can be seen in Fig. 2.22. On the left, we can see how preferential co-assembly varies with mixing ratio and energy bias for species with the same on- and off-rates (that is, $k^{+A} = k^{+B}$, $k^{-A} = k^{-B}$, and so on...). As expected, the most abundant oligomeric sizes at mixing ratios of 0 and 1 (that is, pure A and pure B self-assemblies, respectively) are the same due to the same rate. However, with mixing, there seems to be an increase in size that is maximal at the mixing fraction of 0.5. This effect also gets more pronounced with higher energy biases. This can be explained by the fact that as the ratio of hetero- to homo-dimer off-rates decreases, the stability of the oligomers increases, which manifests as an increase in the size of the most abundant species (which generally means an increase in the distribution size as well). In the case where the on- and off-rates for the two proteins are not the same (that is, $k^{+A} \neq k^{+B}$, $k^{-A} \neq k^{-B}$, and so on...), as shown in the same figure in the right plot, the dependence of the most abundant size on both the mixing ratio and stabilisation energy is less trivial and can depend on multiple factors. There is a general trend for a monotonic increase of size with mixing ratio

assuming the distribution at the mixing ratio of 1 is bigger, but the exact trend depends on the rate constants and range of mixing ratios.

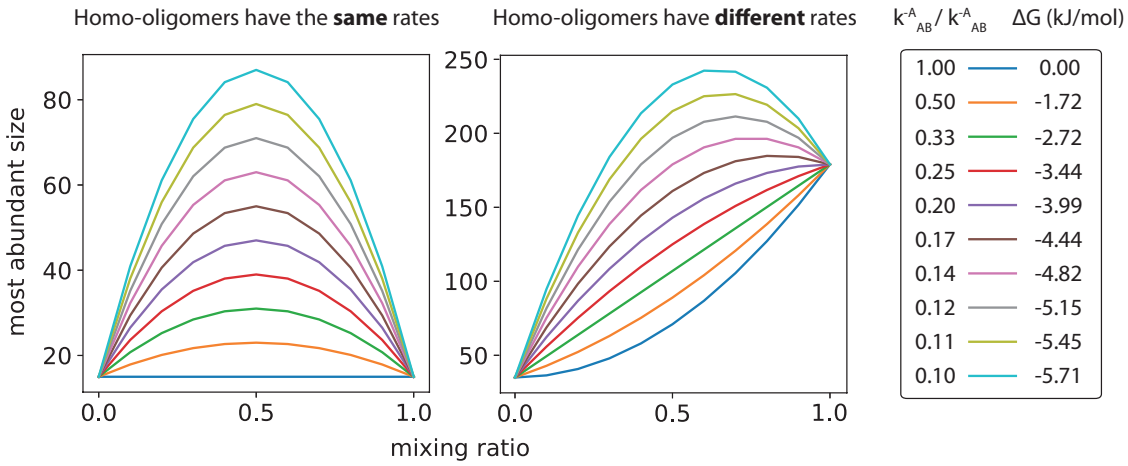


Figure 2.22: The most abundant oligomeric size as a function of either the ratio of off-rates from hetero- to homo-dimers or the Gibbs energy difference in the stability of hetero- and homo-dimers. Here, if we focus on the left figure as an example, when there is no energy difference between hetero- and homo-dimers, the average size of their co-assemblies is independent of their mixing ratio. As hetero-dimerisation becomes more favourable, the average size increases with the mixing ratio and has a maximum at the 0.5 mixing ratio (corresponding to maximised number of hetero-dimers within the oligomers), after which the size decreases again.

For the cases of preferential homo-dimerisation (not corrected as shown in the above section so that the off-rates vary with composition), there is a general trend for the most abundant oligomeric size to decrease upon mixing, with the exact shape of these 'mixing curves' dependent on the exact on- and off-rates. We will not explore the more preferential self-assembly further because of the lack of experimental data for these systems.

Co-assembly causes non-linear changes in the abundance of odd stochiometries

We next explored how the abundance of even and odd oligomeric sizes changes with mixing ratio. In the simplest case of the two mixing proteins having the same on- and off-rates (as above, $k^{+A} = k^{+B}$, $k^{-A} = k^{-B}$, and so on...), there seem to be

surprising changes in the abundance of odd oligomeric sizes with mixing.

To explore this, we let two proteins with a Poission-distributed oligomeric sizes (that is, there is no energy bias between odd and even oligomeric sizes) co-assemble, and follow the changes in odd and even abundances with mixing ratio (Fig. 2.23). Note that these are mass distributions, but we've used the same mass for both species to make any even:odd biases clearer. There seems to be a decrease in the abundance of odd oligomers with mixing. Also note that the dependence of this on the mixing ratio is non-trivial–we find the biggest even:odd bias happens at mixing ratios (of those we followed) of 0.2 and 0.8, with the odd oligomers being slightly more favourable at mixing ratios of 0.4 and 0.6.

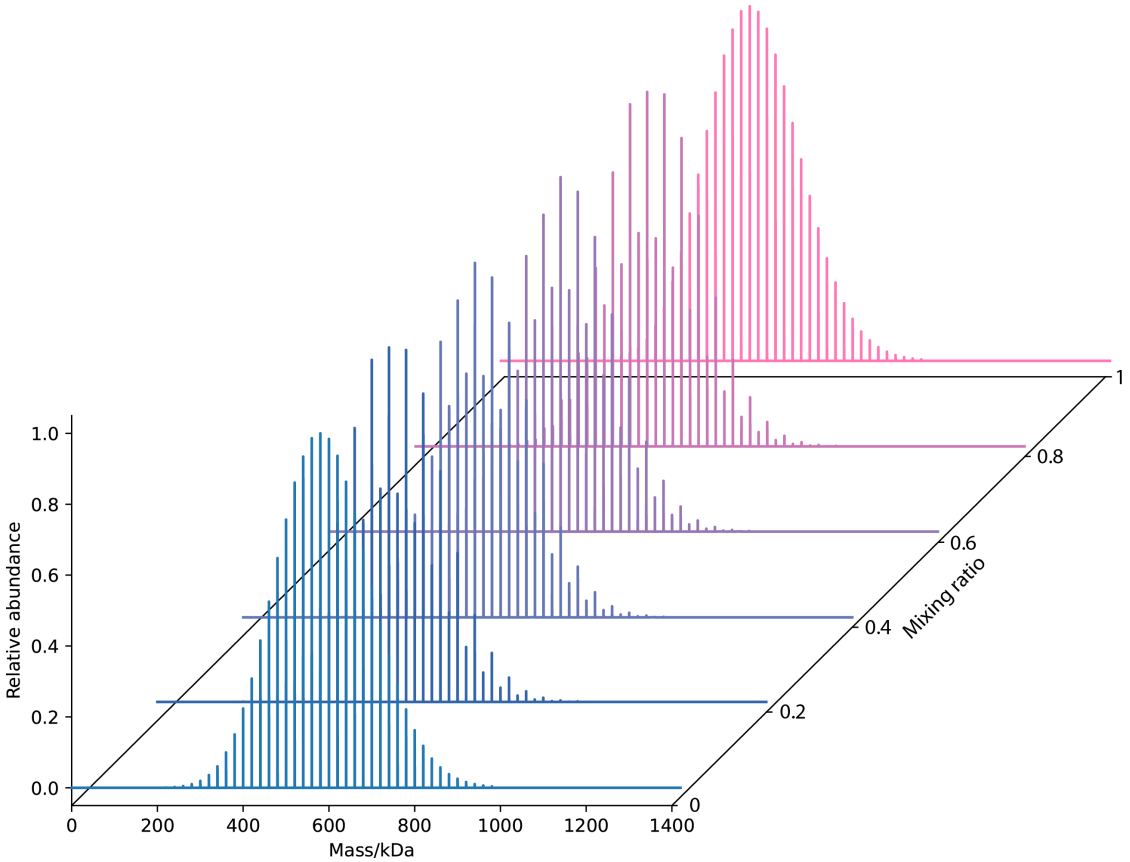


Figure 2.23: The dependence of even:odd bias can be non-trivial. In this example, we mix two species with the same on- and off-rates (as same masses for clarity), and see that mixing ratios of 0.2 and 0.8 have the biggest even:odd bias, 0.4 and 0.6 ratios have a slightly smaller even:odd bias. Pure proteins (mixing ratios of 0 and 1) don't show any even:odd bias, and thus their distributions are the Poisson distribution.

This behaviour is very difficult to explain, as the extent of the even:odd bias depends on all arrangements that belong to one particular sub-stoichiometry, which means it can't be easily described from first principles.

2.4 Acknowledgements

I would like to thank Professor Baldwin for his help in setting the 'kinetic' solution up and for his guidance with the chemical-kinetic model, as well as with developing the kinetic model of (homo-)oligomerisation ([34]). I would also like to thank Martin Galpin for developing the ingenious way of obtaining the 'multiplicities' for the statistical-mechanical model of monodisperse protein assembly. I would also like to thank Wiktoria Sadowska for providing me with the MS spectra for the 12-mer mixing which allowed me to apply the statistical-mechanical model on real-life data. Finally, thank you, Catie, Josh and Virginia, for reading through this chapter and your feedback.

3

Experimental quantification of co-assembly between HspB1 and HspB5

Contents

3.1 Introduction

3.1.1 HspB1 & HspB5

HspB1, also known as Hsp27, was initially characterised in response to heat shock as a protein that contributes to the process of protein refolding ([92]) in an ATP-independent manner. With further investigation, it was recognised that the protein is capable of responding to other stressors, such as chemical or oxidative stress ([93]), which makes it an attractive target as a biomarker of disease and a possible therapeutic. It also exhibits immunomodulatory and anti-inflammatory functions, and is generally widely expressed in various tissues ([94]).

HspB5, also known as alphaB-crystallin, is another highly expressed human sHsp. Is it also one of the major eye lens proteins, together with its closely related HspB4 ([95]), where they comprise roughly 35 percent of all lens proteins ([96]). Similarly to HspB1, it is also an ATP-independent chaperone, and has a plethora of other biological functions–from its role in the eye lens to a functional role in angiogenesis ([97]). It is also an attractive therapeutic target ([98]).

Both HspB1 and HspB5 have a highly polydisperse quaternary structure, with higher oligomers ranging in mass from roughly 200 up to 800 kDa. For HspB5, the subunits have a continuous distribution with the oligomerisation state ranging from 10- to 40-mers ([63]). Their quaternary structure, however, depends on other factors, such as pH ([34]), temperature and concentration ([99]), and phosphorylation state ([100], [101]).

3.1.2 sHsps co-assemble selectively with implications for their function

Many sHSPs are co-expressed in the same tissue. Further, there is a correlation between their expression ([102]) Many are known to form hetero-oligomeric species

([102], [103]), which might hint at the functional relevance of these co-assemblies. Indeed, hetero-oligomers formed by pair-wise mixing of HspB1, HspB5 and HspB6 seem to have different chaperone activity compared to their parent constituents have on their own ([103]). Some of these sHsp co-assemblies have been characterised in detail: HspB2 (also known as MKBP) and HspB3 are known to form hetero-oligomers in a strict 3:1 ratio ([5]). Others, however, have so far evaded a more detailed description of their modes of co-assembly as they form highly heterogeneous oligomers.

The co-assembly between the two of the most expressed human sHsps, HspB1 and HspB5, has been particularly challenging to quantitatively describe. Both HspB1 and HspB5 are highly polydisperse, forming large homo-oligomers in the range of 200-800 kDa ([34], [104]), as well as hetero-oligomers with apparent sizes between those of HspB1 and HspB5 ([105]).

The fact that both HspB1 and HspB5 are quite close in mass (21.7 and 20.1 kDa respectively) makes their possible co-assemblies difficult to study using traditional low-resolution methods as the mass difference between sub-stoichiometries within a particular stoichiometry can be expected to be approximately 1.6 kDa. In this chapter, we describe the co-assembly between these two proteins in unprecedented detail at both the level of stoichiometries and sub-stoichiometries using native mass spectrometry (MS) and mass photometry (MP). In the next chapter, we apply the chemical-kinetical model to both combine the MS and MP data, and extract energy biases for self- or co-assembly in this system.

3.1.3 Developments enabling native mass spectrometry of highly heterogeneous protein co-assemblies

As mentioned in the previous section, the expected mass difference between sub-stoichiometries is below 2 kDa. Since the whole mass envelope of the HspB1-HspB5 co-assemblies can be expected to be somewhere between 200 and 800 kDa, we need an instrument that is capable of transmitting these masses and has a high enough resolution. This has been made possible recently thanks to researchers at the University of Utrecht and Thermofisher Scientific who modified an Orbitrap platform (Fig.3.1) to overcome both of the aforementioned challenges ([69]).

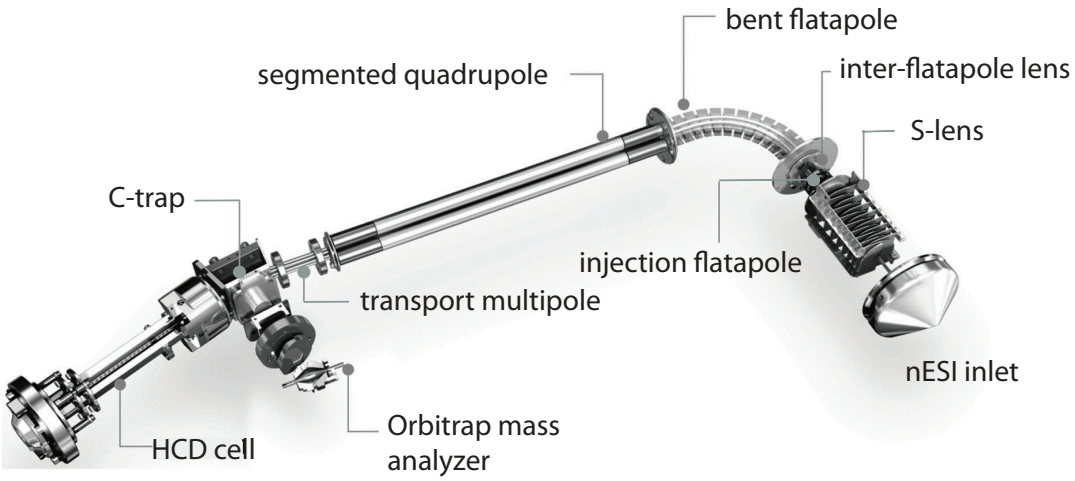


Figure 3.1: Schematic overview of the modifications to the QExactive UHMR instrument. In principle, the changes can be grouped into 3 major contributions: the introduction of in-source trapping (IST), reduction of radiofrequencies (RF) on ion guides, and change in C-trap pulses. The IST was introduced by lowering of the injection flatapole and raising the inter-flatapole lens potentials relative to the S-lens. This allows ion trapping during which the trapped ions can undergo collision with an inert background gas. This enhances resolution by desolvating the analyte and improves transmission by dampening the ion momentum. Secondly, lowered RF: the RF were lowered on all ion guides from the bent flatapole to the HCl cell. This helps with the transmission of large ions. Finally, by extending the interval for which ions are pulsed from the C-trap into the Orbitrap mass analyser, it is possible to give enough time to large ions to enter the Orbitrap for detection. Here then, it is the unprecedented resolution of the Orbitrap that allows us to assign stoichiometries with high confidence. The increased resolution comes from the fact that only the axial component of the motion around the inner spindle is measured without any contributions from the radial motion which decreases the peak width compared to traditional instruments.

Further, assuming HspB1 and HspB5 co-assemble in a statistical fashion, we can expect them to form sub-stoichiometries in the range of tens to hundreds. Assuming these sub-stoichiometries have a couple of charge states each, it is reasonable to expect hundreds of peaks in the native spectrum, and we can expect the number of peaks to increase even more upon collision-induced dissociation. Assignment of the resulting spectra manually would not be possible, and we thus need to use an automatic method for finding charge series and corresponding masses. This is possible using the UniDec software developed by a collaborator ([106]). Further, the collaborator also developed a method for Bayesian assignment of the CID product to their parent species.

Also, as the mass spectra produced by the Orbitrap tend to be biased towards lower masses, as well as the inherent bias coming from the method for recombining MS/MS regions (as will be discussed further in this chapter), we decided to use mass photometry to obtain relatively unbiased mass envelopes. This relatively new technique allowed us to obtain the mass envelopes with unprecedented detail and also served as an orthogonal, in-solution 'check' for the gas-phase native MS measurements.

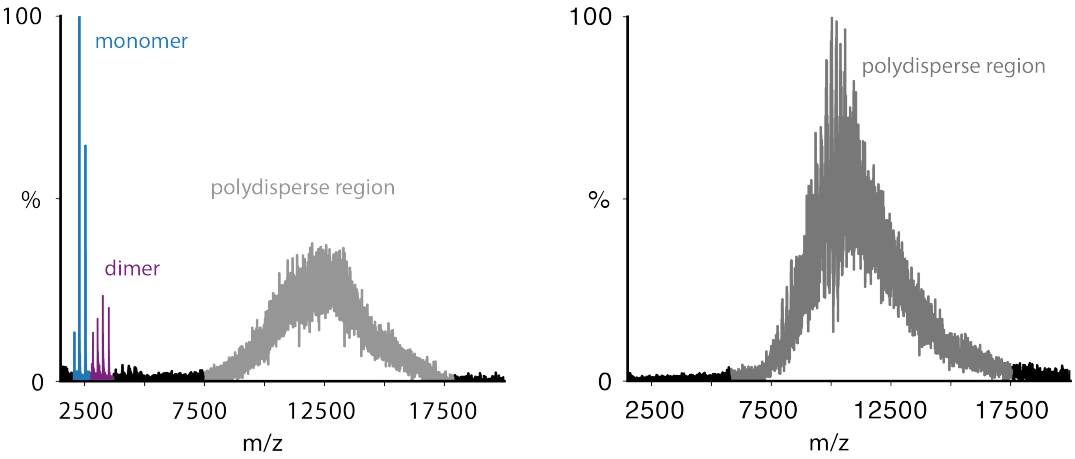
3.2 Results & Discussion

3.2.1 Quantification of polydispersity of HspB1 and HspB5 homo-oligomers

To explore how well the gas-phase native MS and in-solution MP mass measurements correspond to each other, we decided to design a workflow that results in optimal data collection for both methods using homo-oligomers initially before tackling the (more complex) HspB1-HspB5 hetero-oligomers.

Native mass spectrometry

We first recorded the native MS spectra for both homomeric HspB1 and HspB5 (Fig. 3.5). Both proteins presumably form higher oligomers, which cause the broad, baseline-unresolved signal in the higher m/z region. Additionally, there is an appreciable amount of HspB1 monomers and dimers, whereas HspB5 only seems to form higher oligomers. Both of these observations had been noted previously in [104] and [34], respectively.



(a) Native mass spectrum of HspB1. **(b)** Native mass spectrum of HspB5.

Figure 3.2: Native MS spectra of HspB1 and HspB5. Both HspB1 and HspB5 show a polydisperse region (grey). We can further see monomers (blue) and dimers (purple) in the HspB1 spectrum which are missing in the HspB5 spectrum.

To resolve the broad signal coming from higher oligomers, we performed a CID experiment by increasing the acceleration potential into the high-energy collision dissociation (HCD) cell. This allowed us to strip up to two subunits from the oligomers, which resulted in an appreciable loss of charge from the higher oligomers, moving their signal to a higher m/z region, and allowed us to assign the resulting baseline-resolved peaks to a range of oligomeric sizes as shown in the deconvolved mass spectrum in Fig. 3.3.

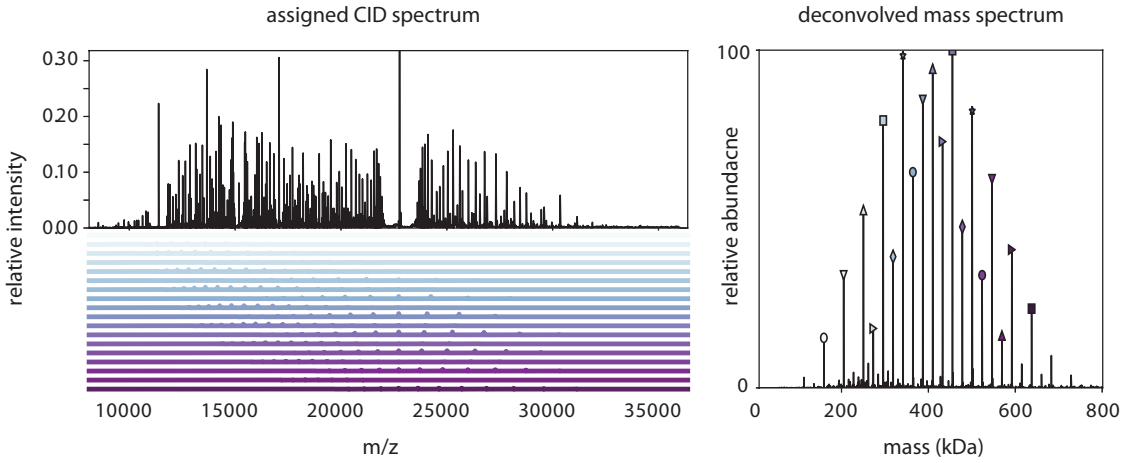


Figure 3.3: CID spectrum assignment. The orbitrap platform allows us to obtain baseline-resolved m/z spectra for the HspB1 homo-oligomers (left). We can than use UniDec to deconvolve the m/z spectrum into a mass distribution (right). Note that this mass spectrum is not CID-corrected yet.

Traditionally, the m/z mass spectrum assignment has been done by extracting the areas under the peaks closest to the 'resonance' peak (whose m/z is the same as the mass of the monomer subunit; all oligomers whose charge is the same as the number of their subunits contribute to this peak which often makes it quite prominent in CID spectra of polydisperse systems) as each of these peaks can be uniquely assigned to only one particular oligomeric size ([63], [107]). The unprecedented resolution of the Orbitrap mass analyser combined with recent software developments ([106]) allows to assign most of the CID spectrum which increases the confidence of this assignment.

The deconvolved spectrum above is a combination of singly- and doubly-stripped oligomers. To correct for the loss of subunits caused by CID, we assign the relative contributions from the n-1 (singly-stripped) and n-2 (doubly-stripped) regions to the oligomeric distribution, and correct for those accordingly.

The corrected mass distributions for both HspB1 and HspB5 are shown in Fig. 3.5.

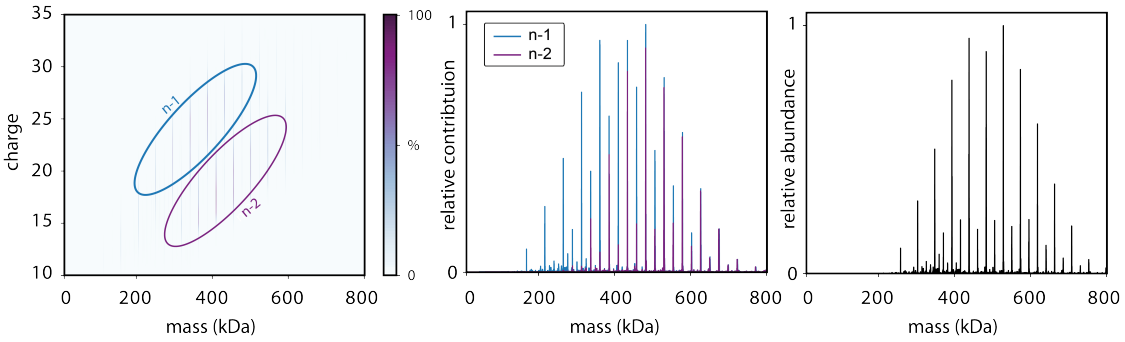
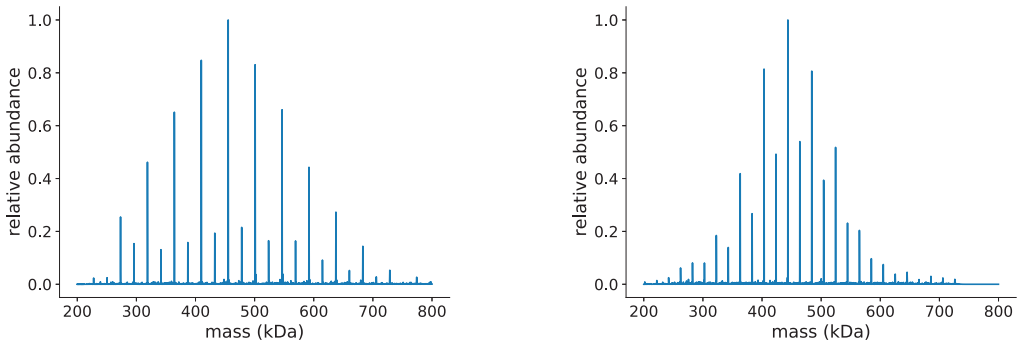


Figure 3.4: CID correction for homo-oligomers. To correct for CID, we need to identify contributions from the n-1 and n-2 regions (left and middle plots). We can than correct for CID by simply adding either one or two monomer masses to the mass distributions corresponding to the n-1 and n-2 regions, respectively, and then adding these two corrected distributions together.



(a) Mass distribution of HspB1. **(b)** Mass distribution of HspB5.

Figure 3.5: Mass distributions of HspB1 and HspB5. Most oligomers are between 200 and 800 kDa. There is a preference for forming even-numbered oligomers for both proteins.

The mass distribution for HspB5 is similar to distributions published before ([34]). Comparatively, the mass distribution for HspB1 shows a higher relative preference for even stoichiometries. This might also cause the relative broadening of the HspB1 distribution compared to HspB5–based on results from Baldwin et al. ([34])–as odd stoichiometries become increasingly less favourable, the distribution starts to broaden compared to a Poisson distribution, which is the equilibrium distribution for the case when there is no bias for even or odd stoichiometries–as shown in Fig. 3.6.

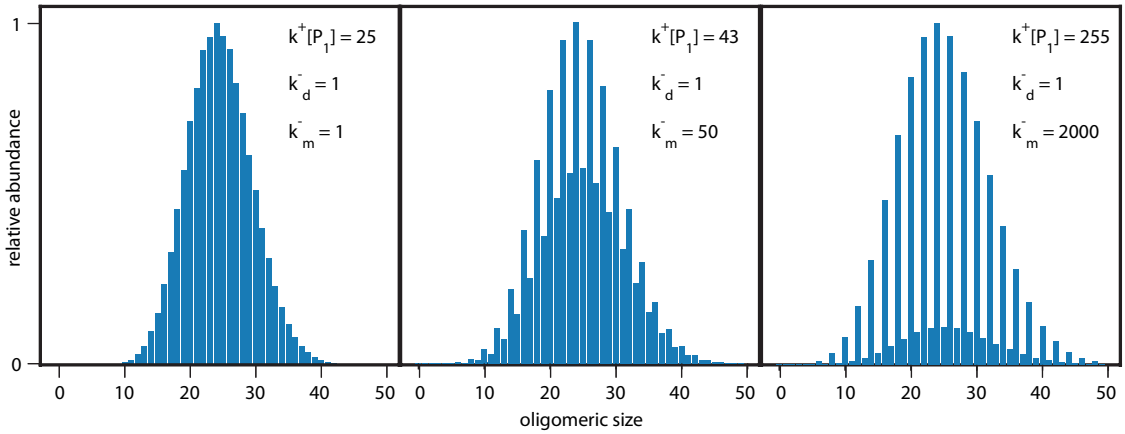


Figure 3.6: The width of the oligomeric distribution changes with the even:odd bias. In this simulation, we varied the on- and off-rates to keep the main oligomer roughly the same while varying the even:odd bias. As the even:odd bias increases, the width of the resulting distribution also increases.

Mass photometry

MP is a method that is orthogonal to the native MS experiments presented above. It is a single-molecule, in-solution method that uses light scattering to estimate the mass distribution of species in a given sample. A more detailed introduction can be found in Chapter 1, but to reiterate briefly: In an interferometric scattering experiment, the signal scales with the polarisability, which itself is a function of the refractive index and proportional to the physical volume causing this scattering. As the specific volumes and refractive indices of individual amino acids vary only slightly (less than 1 percent), the scattering signal is proportional to the mass of the protein, independent of its shape. After the scattering profile of a given sample solution is recorded, further processing is done by averaging N frames before the frame of interest and N frames after the frame of interest. With sufficiently high N, the background can be averaged out, and we are left with only changes in scattering caused by protein species interacting (binding or unbinding) with the glass measurement surface. Fig. 3.11 briefly shows an oligomer binding to (dark) and then a smaller species unbinding from (light) the glass surface obtained after averaging.

Since the monomer mass of HspB homologs is roughly 20 kDa, we had to optimise the data acquisition and analysis. This was done so that the peaks corresponding to even stoichiometries were as resolvable as possible.

First, we had to find the optimal concentration at which to perform the measurements as MP requires sample concentration in the range of tens of nM. Using too low of a concentration results in a low density of binding events, longer experiment time and the added possibility that the oligomeric distribution could shift to lower sizes after dilution depending on the magnitude of its dissociation constant(s). On the other hand, a sample concentration that is too high might bias the mass distribution. We found that, independent of the concentration we used to equilibrate the sample before the final dilution, the ideal MP sample concentration was around 400 nM monomer concentration for both HspB1 and HspB5. This corresponds roughly to 20 nM 'species', or oligomer, concentration.

Using a higher number of frames for averaging favours the visibility of smaller species with a loss of resolution in the higher mass range. A smaller number of averaged frames on the other hand comes with high background noise. The averaging window was optimised for each measurement, but was usually between 8-12 frames. The effect of averaging on the MP histograms can be seen in Fig. 3.7

We then had to further optimise the two filters used for filtering the binding and unbinding events in the native MP movies. DiscoverMP, the analysis software, uses two filters for finding events in a two-stage strategy. The first filter (T-test; called 'filter 1' in Fig. 3.8) detects step changes in reflectivity of the glass surface, indicative of binding and unbinding. Higher values of filter 1 reject more noise, potentially helping with resolution, but this also means smaller events get rejected, as can be seen in Fig. 3.8. The second filter (fast radial symmetry; called 'filter

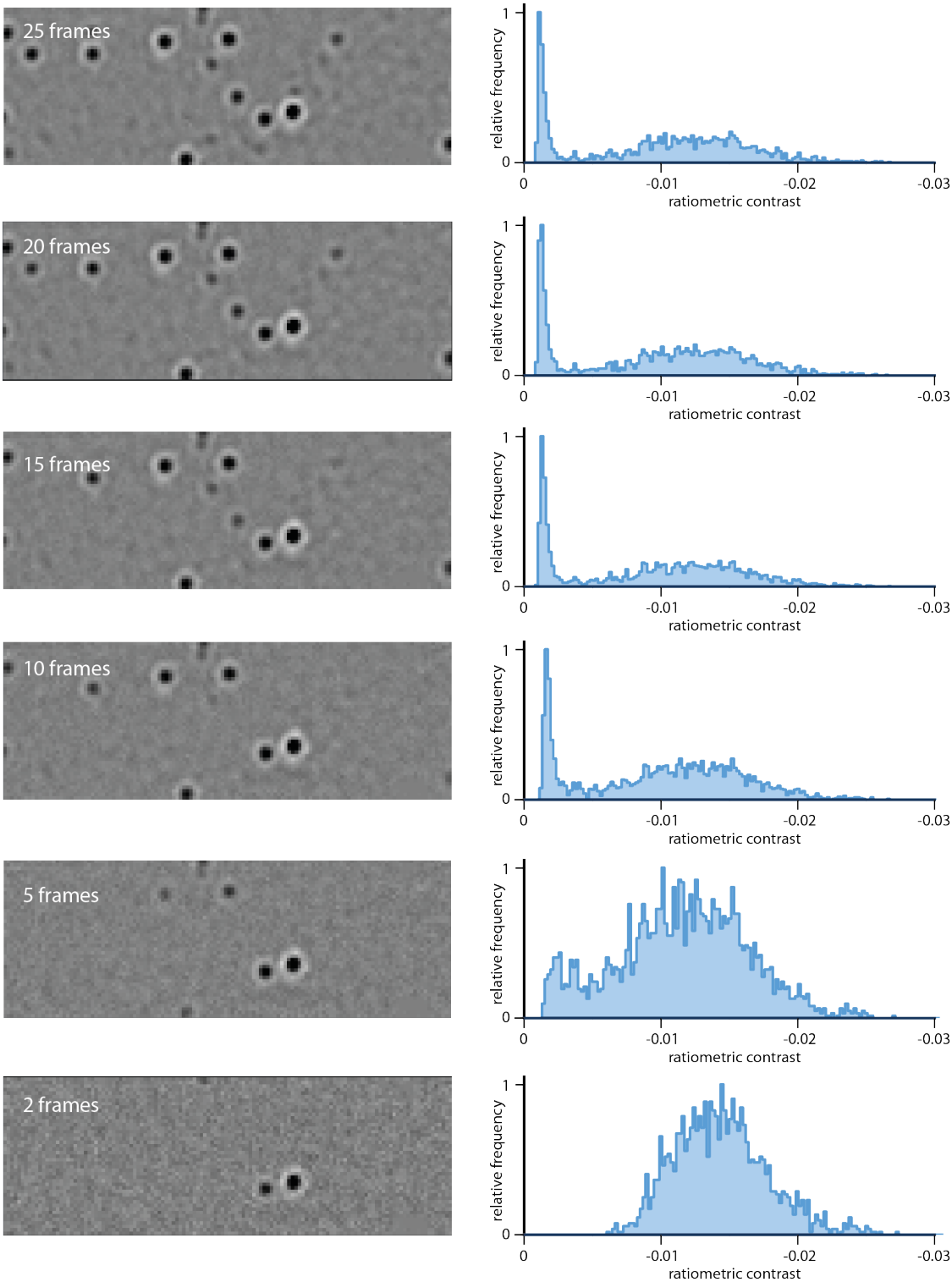


Figure 3.7: The effect of averaging on resulting ratiometric distributions. Since the monomer masses are close to each other (on the MP mass scale), we need to optimise the averaging to obtain the best possible resolution.

2' in Fig. 3.8) assesses whether the respective event in the ratiometric image matched the expected signature.

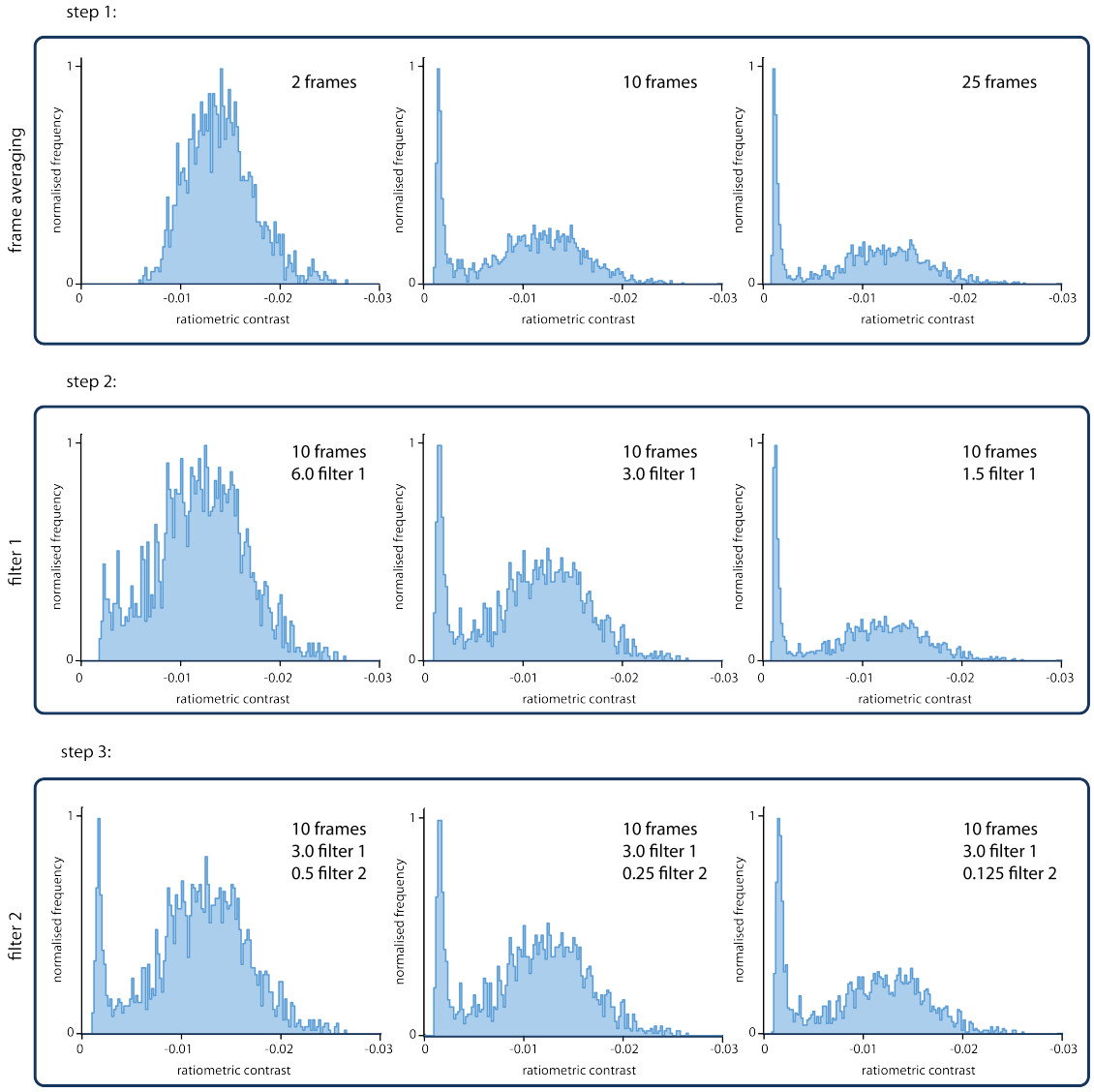
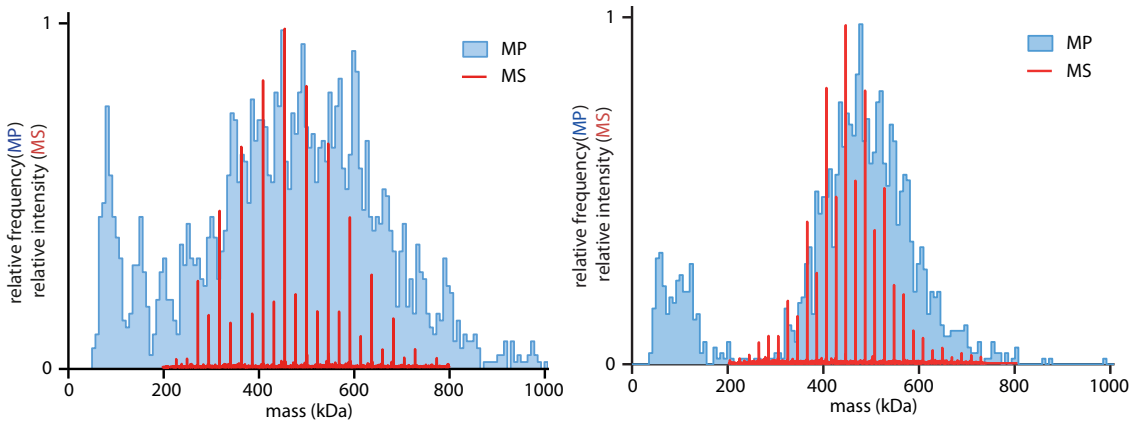


Figure 3.8: The effect of averaging on resulting ratiometric distributions. Since the oligomeric masses are close to each other, we need to optimise averaging to obtain the best possible peak separation.

In our measurements, filter 1 has a bigger impact on the final 'resolution' of the MP mass histogram, whereas filter 2 mostly affects the proportion of lower-mass events that are matched without significantly affecting the shape of the higher-mass distribution.

The resulting MP mass histograms, and their overlap with mass distributions obtained by MS can be seen in Fig. 3.12.



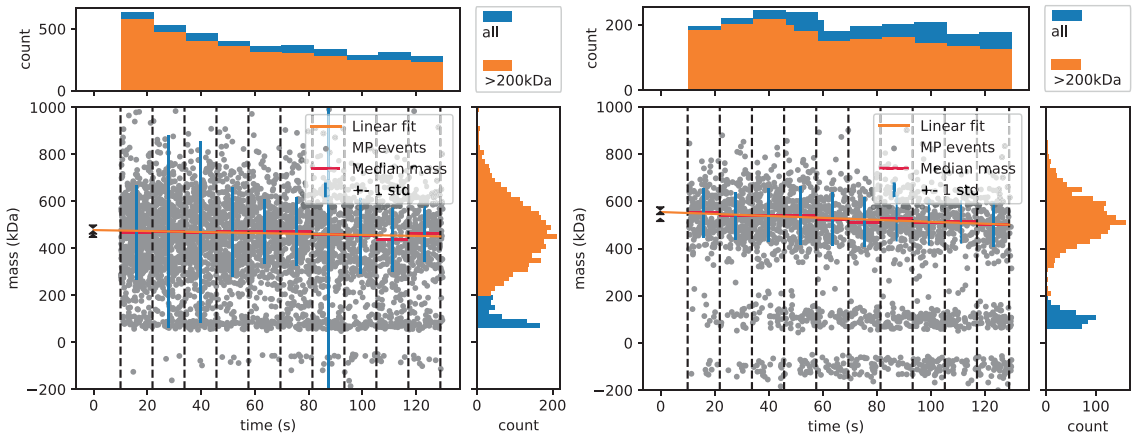
(a) Mass distributions of HspB1.

(b) Mass distributions of HspB5.

Figure 3.9: Comparison between MS- and MP-acquired mass distribution. Even though MP is lacking the roughly Da resolution of MS, the mass distribution is not biased towards lower masses.

There seems to be a good overlap for even stoichiometries between the MS and MP measurements. The MS mass spectra were obtained by CID which did not allow us to see smaller oligomers, but from the native MS spectra (Fig. 3.5), we expect to see smaller species in the HspB1 MP histogram. The smaller oligomers in the HspB5 MP histograms were surprising, and their potential origin is discussed in the next section. Both the MS and MP spectra seem comparatively somewhat broader for HspB1 than HspB5. This behaviour is, again, expected as explained previously–higher even:odd bias at similar average size causes a relative broadening of the distributions. Even though the MS and MP distributions are centred around the masses for both HspB1 and HspB5, the MP distributions are somewhat broader in the higher-mass region. This comes from the fact that the MS measurements are more biased towards lower masses compared to MP.

As the MP measurements are done at roughly 50-150 times dilution compared to MS, we wanted to ensure that the dilution does not result in a shift towards lower



(a) HspB1 **(b)** HspB5

Figure 3.10: MP time courses for HspB1 (a) and HspB5 (b). The time courses show little to no decrease in the mass of the landing species. HspB5, however, shows progressively more small (<200 kDa) binding and unbinding events with time. The fitted medians and averages correspond to the orange part of the MP mass distributions (y-axis histogram). The number of binding events in each time bin is shown in the x-axis histograms. HspB1 shows a roughly exponential decrease of binding events, whereas HspB5 binding events seem to be at a steady state.

masses since the dissociation constants for these proteins are generally not known. We employed the following method to examine any potential decay in size with time, as well as to back-calculate the original distributions should this be needed: we bin every 10 seconds of the measurement and calculate the average and median mass for higher (> 200 kDa) oligomers. After this, we fit both the medians and the averages, and back-calculate the 'original' distribution by extending the linear fit by the 'dead time'-the time from sample dilution to the start of measurements. With longer experimental times, we could expect the mass shift to follow a roughly exponential decay rather than a liner decay, but in the window period we use here, there was not a statistical difference between the two fits. The fitted values as well as the mean and median mass of the whole, 'full-time' MP envelope can be seen in Table 3.1.

The values differ on the scale of roughly one or two monomer masses, well within the margin of error for these measurements. This shows that the homo-oligomers do not disasssemble into smaller oligomers significantly on the time-scale of the

	HspB1	HspB5
fit	477	510
average	488	500
median	465	486

Table 3.1: The MP distribution's medians, averages, and fits to medians. For both HspB1 and HspB5, the differences between each are smaller than 2 monomer masses, which is well within the experimental error of the MP measurements. This suggests that on the time scale of the measurement, the oligomers do not dissociate to an appreciable extent.

measurements (2 minutes + roughly 10 seconds of a dead-time).

Lower mass region on mass photometry

From the native mass spectrometry measurements, we know that human HspB1 has a substantial population of monomers and dimers, whereas HspB5 does not (Fig. 3.2a and Fig. 3.2b, respectively), at least at micromolar monomer concentrations. Thus, it was interesting to see smaller species of HspB5 in the MP measurements. Multiple contributing factors may be at play here: First, the mass photometry measurements are conducted at much lower concentrations than native MS. Theoretically, there is a critical concentration ([34]) below which the higher oligomers disassemble into smaller oligomers. This behaviour has been also predicted for other co-assembling systems ([86]).

However, we tried to dilute the sample equilibrated at micromolar, MS concentration in a matter of seconds. In this time frame, it's very unlikely that an appreciable amount of the higher HspB5 oligomers would disassemble into smaller species, since the same behaviour was observed with a protein solution (and the sample buffer) at lower temperatures of around 4 °C–a temperature where any monomer exchange is expected to be very slow.

The other explanation is a dissociation of smaller species from surface-bound oligomers, something we can indeed see on MP measurements of HspB5, as shown

in Fig. 3.11. Here, the unbinding events (red part of the histogram) are recorded as having a negative mass.

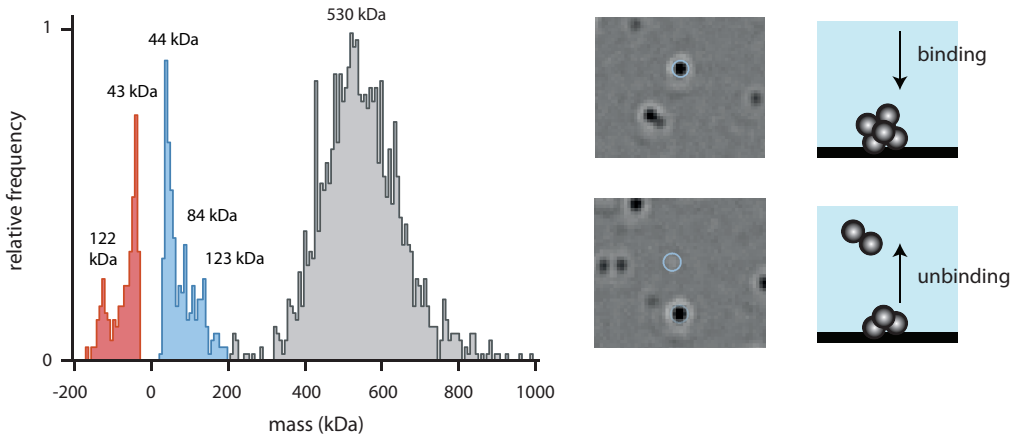
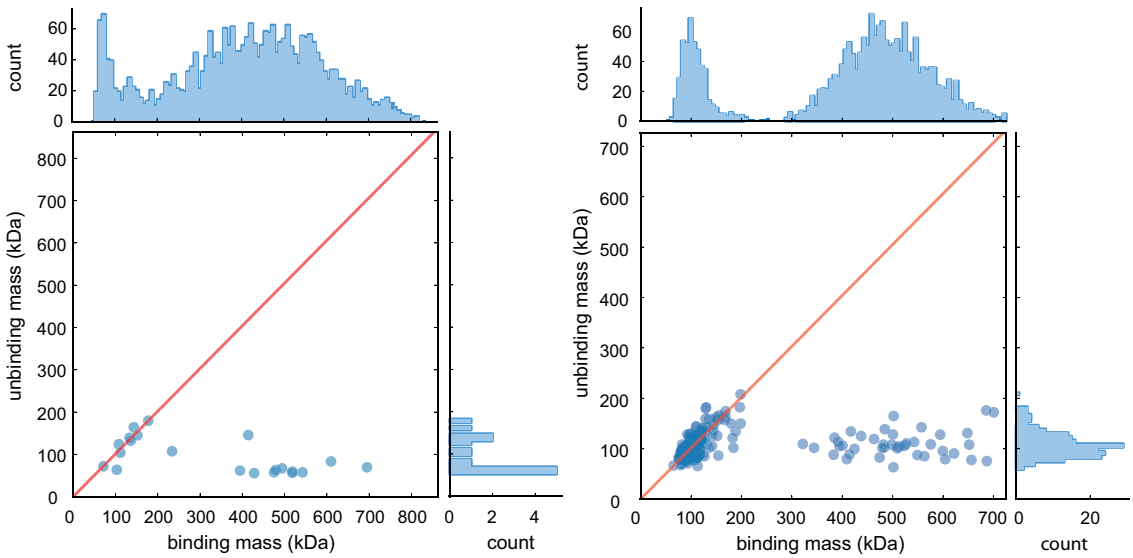


Figure 3.11: HspB5 is prone to dissociation from the glass surface. Left-The higher oligomers are shown in grey; this mass envelope agrees well with the measured MS mass envelope. There are, however, additional binding (blue) and unbinding (red) events in the MP histogram. The unbinding events (light, bottom right) often come from the same region where a higher oligomer (dark, top right) was previously bound.

To explore this behaviour, we employed two different methods - particle tracking and surface coating.

With particle tracking, the goal is to find the particles from which the smaller species dissociate. We have developed a particle tracking algorithm that assigns unbinding events to the closest binding event (in space) on the assumption that the particle only moves minimally between the binding and unbinding events. This seems to hold true for both HspB1 and HspB5 as we see relatively little particle movement in the MP movies.

The result can be seen in Fig. 3.12. For HspB1 (Fig. 3.12 a), there is a relative lack of unbinding events, and so it is very likely that the smaller oligomeric species we see are indeed present in the solution even before any interaction with the glass surface and/or the laser beam, and are not a product of smaller HspB1 species unbinding from the glass surface, and binding again.



(a) HspB1 **(b)** HspB5

Figure 3.12: The binding events and their corresponding unbinding events for both HspB1 and HspB5. Here, on the binding events are on the x-axis, while the corresponding unbinding events are on the y-axis. The circles represent found binding-unbinding pairs. HspB1 seems to be quite stable when it comes to unbinding, whereas smaller species seem to dissociate from higher HspB5 oligomers comparatively quite easily. It is then possible for the smaller species to undergo multiple binding-unbinding cycles.

Interestingly, that may not be the case for HspB5 (Fig. 3.12 b). Here, the majority of the unbinding events come from the lower-mass binding events. One possible mechanism to explain this behaviour is that initially, lower-mass species unbind from larger oligomers (bottom right region in the plot), upon which these smaller species might experience repetitive binding and unbinding cycles, which then causes the large amount of lower-mass species in the MP mass spectrum; something that we don't see with native MS.

Since, under normal conditions, the glass surface is negatively charged, it might repel negatively charged proteins more readily than neutral and positively charged ones. To explore how the surface charge affects the possible unbinding, we coated the glass surface with (3-Aminopropyl)triethoxysilane (APTES) to obtain a positively charged surface ([108]).

For the sake of comparison, we performed the APTES experiment with both

HspB5 and HspB1, but the effect on HspB1 binding and unbinding seemed negligible. This might be because the pH at which we performed the measurement, pH 8, is quite close to the pI of APTES, which is reported to be about 8.7. This means that it is possible that the APTES surface was insufficiently charged at pH 8 ([108]).

Alternatively, since from native MS we know that the smaller species exist in-solution in an appreciable extent, and don't result from the interaction with the glass surface, it is feasible that any surface manipulations would have a minimal effect on the lower mass region. This, however, does not explain fully why HspB1 binds strongly with the APTES-coated surface even when both the surface and the protein are positively charged.

Interestingly, the APTES coating had a considerably bigger impact on the interaction of HspB5 with the surface (Fig. 3.13).

At pH values higher than HspB5's pI value, the oligomers bound completely with very little unbinding (Fig. 3.13 a, upper figure). Presumably, this is because of the attraction between positively charged APTES and negatively charged HspB5 oligomers, which in this case slows down the unbinding of both smaller and bigger oligomers. Most importantly, there are almost no smaller species present in this case. This lends further credence to the hypothesis that the smaller HspB5 species (Fig. 3.12) are a result of initial loss of a smaller oligomer from a bigger one, which then binds again, and possibly undergoes more binding/unbinding cycles. The masses of binding and correspond unbinding pairs are shown in 3.13 c).

In the case where both the protein and the APTES surface had the same, positive, charge (Fig. 3.13 a, lower plot), we see binding followed by rapid unbinding, presumably due to the repulsive electromagnetic forces. The masses of binding and correspond unbinding pairs are similar, and so lie on the red diagonal in Fig. 3.13 d). Still, when the bigger oligomers were not allowed to stay attached to

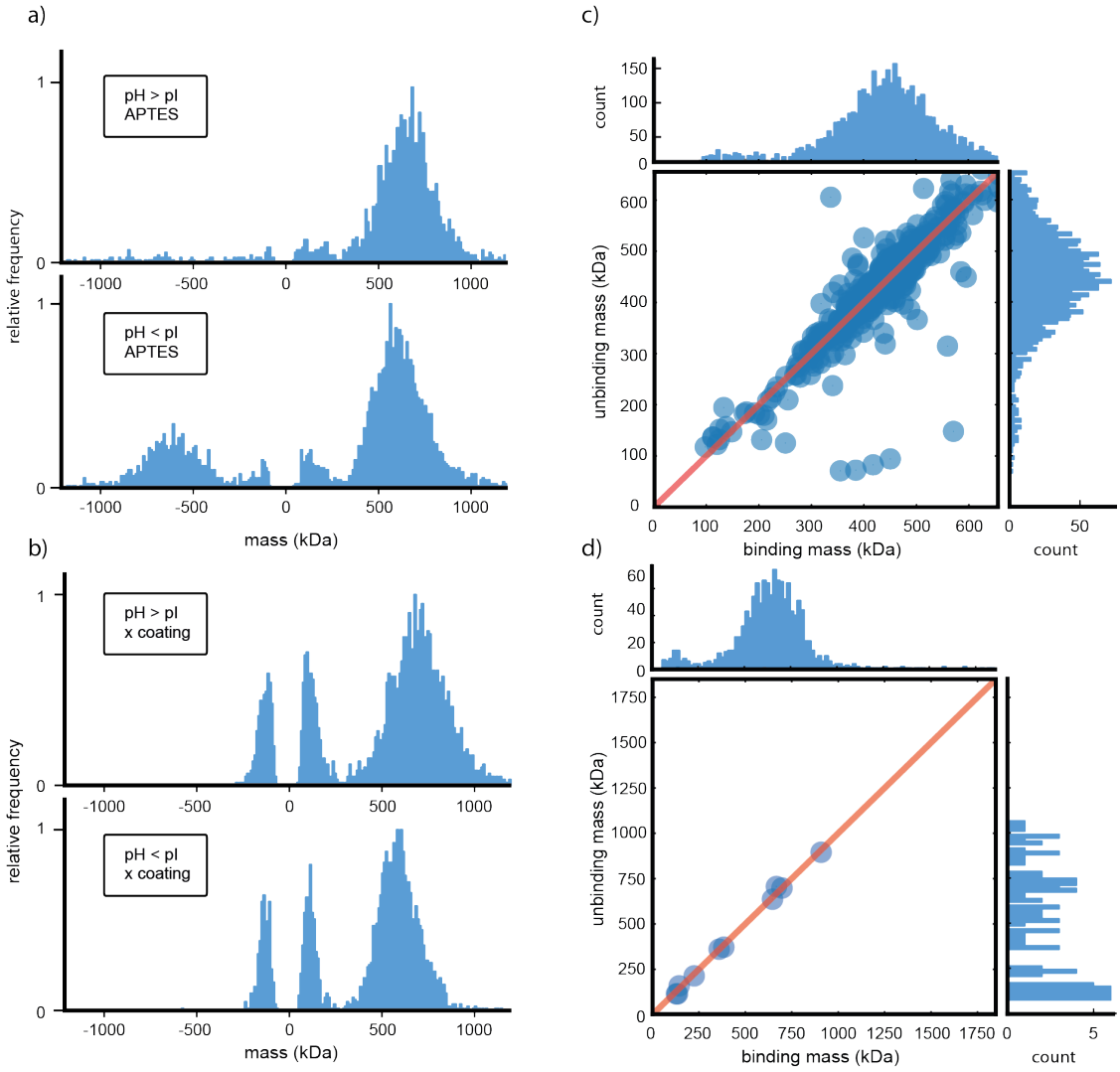


Figure 3.13: Binding and unbinding of HspB5 under different pH and glass coating. Whereas without coating (b), there isn't a drastic change in the unbinding behaviour, coating with APTES (a) seems to affect the rate of the unbinding of smaller species.

the glass surface, we see comparatively very small amounts of the smaller species binding. This might suggest that for HspB5, the interaction between the larger oligomers and the (uncoated) glass surface might speed up the dissociation of smaller species compared to the solution. This might be due to the extra energy supplied by the laser to the glass-solution interface.

For the sake of comparison, we also recorded the above measurements (that is, for pH $>$ pI and pH $<$ pI) without any coating. There does not seem to be a

large quantitative difference between them, or between them and measurements done at neutral pH.

For both APTES and no coating, the oligomeric distribution is somewhat broader and more resolved for pH > pI compared to pH < pI. This is a known phenomenon - at higher pH values, the oligomeric distribution of HspB5 becomes more biased towards even oligomers ([34]), which intrinsically leads to a larger range of occupied stoichiometries, as explained in the native MS subsection of this chapter. This can be seen quite well in Fig. 3.13 b.

3.2.2 Experimental quantification of co-assembly between HspB1 and HspB5

In this section, we show how we can obtain a detailed experimental description of co-assembly between HspB1 and HspB5 using both native MS and MP, how it varies with mixing ratio, and how these methods complement each other. Then, in the next chapter we explore how we can combine these experimental results with the chemical kinetics model described in Chapter 2 to obtain the full distribution of these complex mixtures. The measurements were done on mixtures at three different ratios, roughly corresponding to 1:3, 1:1, and 3:1 HspB1:HspB5 mixing ratio. For the rest of the HspB1-HspB5 mixing work, the mixing ratios are reported as HspB5:HspB1 - the amount of HspB5 is first and that of HspB1 comes second. Similarly, any molar fractions mentioned in the text are the molar fraction of HspB5 in the mixtures.

Native mass spectrometry

Unlike the HspB1 and HspB5 homo-oligomers, HspB1-HspB5 hetero-oligomeric co-assemblies are not amenable to a simple whole-spectrum CID measurement (Fig. 3.14). This is possibly due to the high number of sub-stoichiometries present in the HspB1-HspB5 mixture, as well as their relatively bigger size which could

make it harder for the ions to be accelerated enough to cause CID. In case the co-assembly is purely non-preferential, we can expect up to hundreds of different sub-stoichiometries, which would be difficult to assign even with the great charge reduction cause by CID. To circumvent this, we employed a tandem MS strategy, wherein we selectively use a comparatively small m/z region, which we then CID.

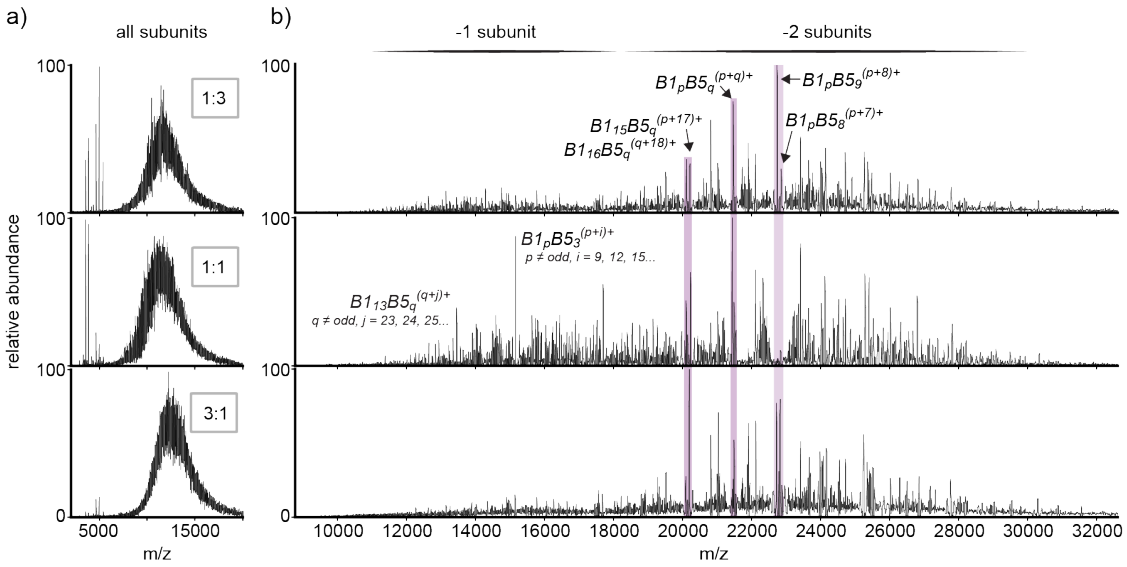
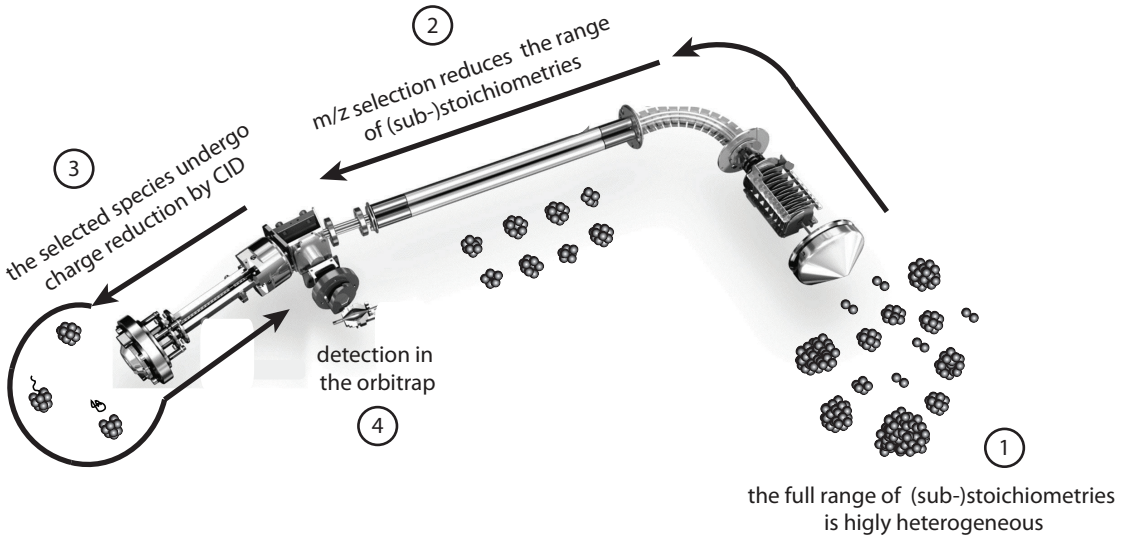


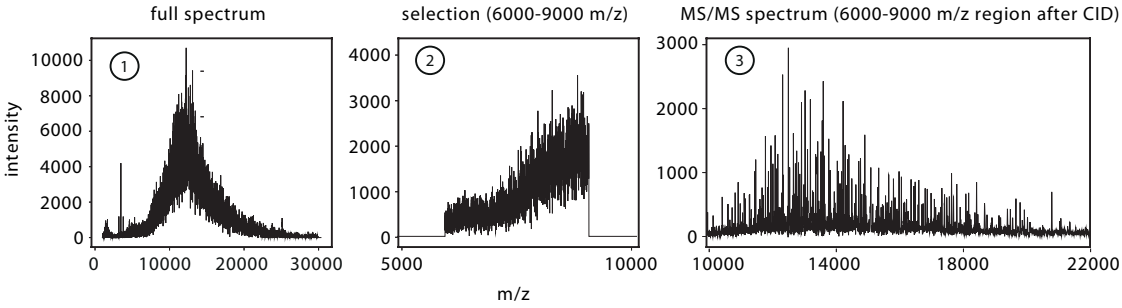
Figure 3.14: Similarly to HspB1 and HspB5 homo-oligomers, we attempted a whole-spectrum CID to reduce the oligomeric charges which leads to a spread to higher m/z which made the homo-oligomeric distribution resolvable (Section 3.2.1). However, the whole-CID spectra of the HspB1-HspB5 mixtures are not baseline resolved. This might be due to the relatively higher number of different oligomeric sizes in the mixtures compared to homo-oligomers of HspB1 and HspB5, which are baseline-resolvable. Nevertheless, there still are features (purple strips) that are assignable. These presumably come from particular combinations of p HspB1 and q HspB5 subunits with particular charge losses. This figure was adapted from [109] with permission.

Tandem mass spectrometry

As explained in Chapter 1, in a typical tandem MS (also called MS/MS) experiment, we select a region of m/z in the quadrupole before CID is done on the selected species. The m/z selection can greatly reduce the number of (sub-)stoichiometries that undergo CID before being detected, which might make the recorded spectra more resolvable (Fig.3.15b).



(a) After injection into the instrument (1), the oligomers are filtered (2) so that particles with only a particular range of m/z values have stable trajectories in the quadrupole. After this, the species are activated in the HCD cell (3), which results in a loss of one or more subunits.



(b) Without filtering (1), the recorded spectrum comprised of all oligomers that were present in solution and transmitted. MS selction (2) allows us to pick a particular m/z region, and narrows the number of species we analyse at the same time. CID (3) of the selected region results in a baseline-resolved spectrum.

Figure 3.15: Overview of an MS/MS experiment. The top figure shows movement of the oligomeric species through the instrument and the bottom one shows the resulting recorded spectra.

We were able to record baseline-resolved, deconvolve-able spectra using successive 520 m/z windows. 520 m/z was chosen so that it includes both a main 500 m/z window as well as 10 m/z 'overhangs' that overlap with the lower and higher windows to ensure continuous coverage. We were able to collect 5 MS/MS windows

between roughly 8740 and 11260 m/z for all three mixing ratios, with the windows for the individual mixing ratios varying slightly. We found that windows higher than 11260 m/z were not baseline-resolved and had poor signal intensities. The voltage we are able to apply in the HCD cell to cause collision-induced dissociation is limited. The presumably bigger species in the higher m/z ranges might not have been kinetically activated enough to cause CID as the velocity of the ions accelerated in an electromagnetic field is roughly inversely proportional to their mass. Often, collisions with heavier inert gasses help the CID process. We performed these measurements with both N_2 and Ar in the HCD cell without a significant difference. The spectra analysed below were all recorded with N_2 as the collision gas.

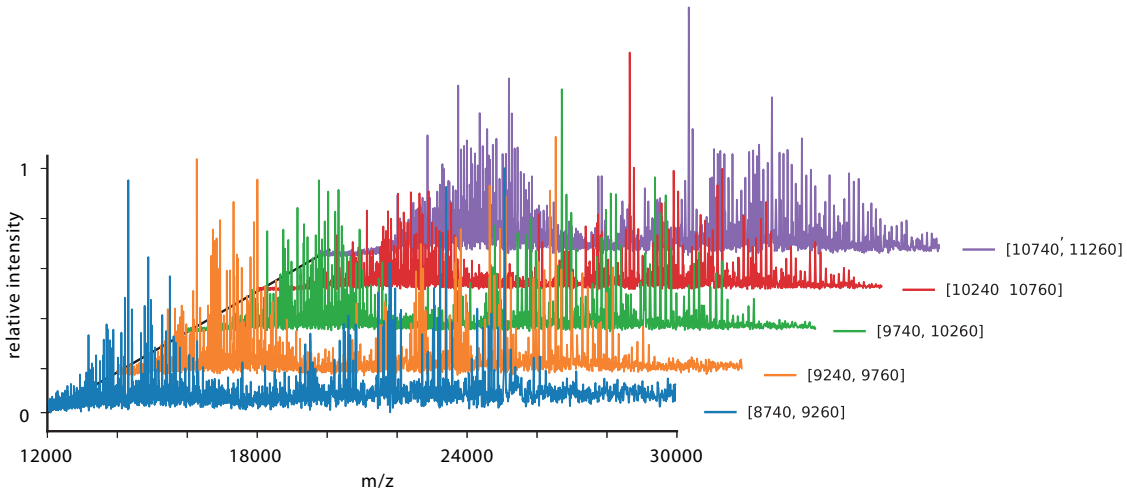


Figure 3.16: Recorded MS/MS spectra of the HspB1-HspB5 mixture at the 1:1 mixing ratio. The m/z ranges for selection shown to the right. There is a gradual shift to the higher m/z with higher selection window. This hints at potentially sequentially heavier species within the selected windows.

Fig. 3.16 shows the five tandem MS spectra recorded between 8740 and 11260 m/z. In all spectra, there are two regions: the left region, roughly below 20000 m/z is the "(n-1)" region originating from oligomers that have lost one subunit. The loss of two subunits results in a comparatively bigger charge loss, which means the "(n-2)" region is shifted towards higher m/z (in this case, roughly above 20000

m/z). There is a gradual shift of the m/z envelopes towards higher m/z with higher MS/MS windows. This is due to sequentially bigger species being present in higher tandem MS regions.

Correction for subunit loss

Correcting for the loss of two subunits is non-trivial. One possibility is recording both the low- and high-m/z regions so that on top of the higher oligomeric distribution, we also have access to the information about what subunits were lost. We could then use this ratio, and assuming the probability of losing either HspB1 or HspB5 are mutually independent, could try and back-calculate the original, 'parental', mass distribution.

There are multiple problems with this approach–the loss of subunits might not be completely independent, (n-1) and (n-2) might have different contributions to monomer-loss ratio, and most importantly, this approach applies the correction for either A or B to whole mass spectrum regions, not specific masses only. This can be highly problematic.

Imagine we have an A_5B_5-mer that we CID and want to analyse the (n-1)-region. In this region, we find two species–A_4B_5 and A_5B_4 in a 1:1 ratio. Let's say that by checking the lower-m/z region, we find that the parental oligomer(s) lost the A and B subunits also at a 1:1 ratio. Broadly, indiscriminately applying this information to the CID species would do the following:

First, if we correct for the loss of A:

$A_4B_5 + A \longrightarrow A_5B_5$

$A_5B_4 + A \longrightarrow A_6B_4$

And then the loss of B:

$A_4B_5 + B \longrightarrow A_4B_6$

$A_5B_4 + B \longrightarrow A_5B_5$

The resulting (wrongly) corrected parental distribution would have three oligomers– A_6B_4, A_5B_5, and A_4B_6, in a 1:2:1 ratio.

We devised a Bayesian approach for identifying the most likely masses coming from the loss of either A or B (Fig. 3.17).

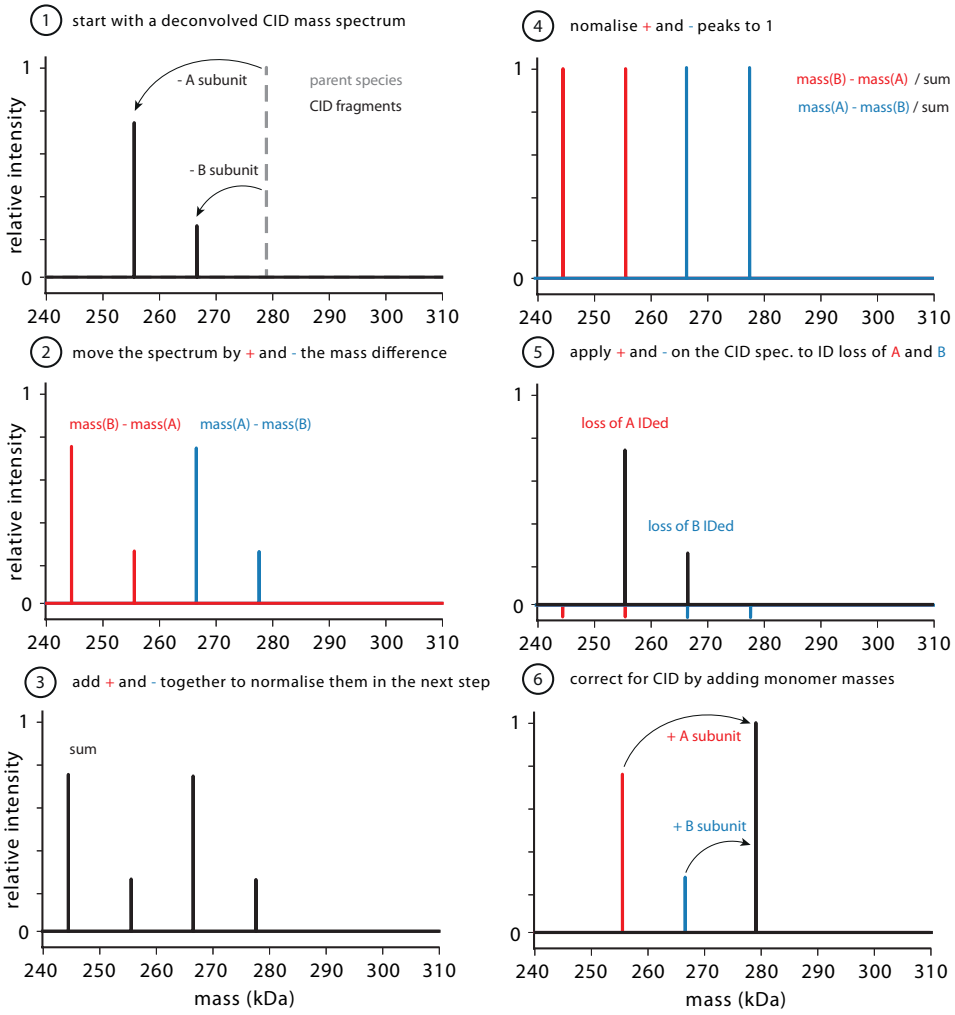
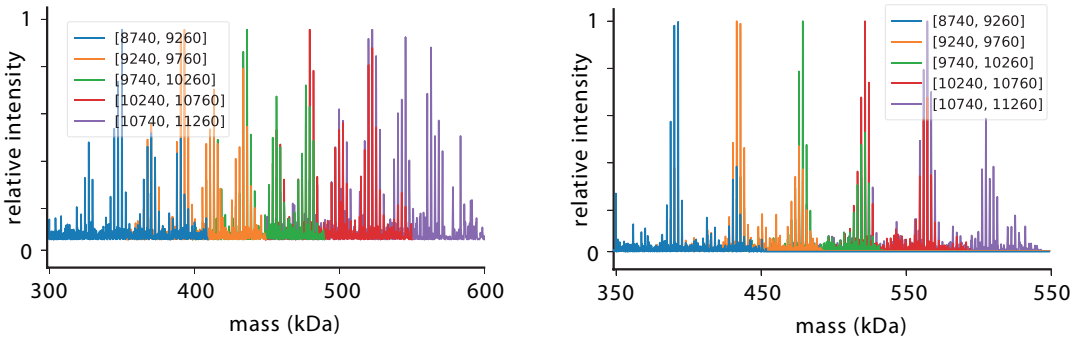


Figure 3.17: Algorithm for CID correction for n-1. (1) shows the proposed recorded CID spectrum (black) that results from unequal loss of the A and B subunits from a parent species (shown in grey). Moving the CID mass spectrum by either (mass(A)-mass(B)) or (mass(B)-mass(A)) on the mass axis (2) allows us to identify peaks coming from either a loss of subunit A or subunit B, respectively, from a parent species common to the both of them. Dividing (2) by their sum (3) results in a normalised 'spectra' that show the positions of potential losses. By multipliyng (1) by the (4), we can ID the potential contributions, and correct the losses independently. Note that the only priors for this process are the monomer masses for the two subunits.

Here, we again have an A_5B_5-mer (with arbitrarily chosen monomer masses:$M_w(A)$

= 12345 Da and $M_w(B)$ = 23456 Da), which loses either the A or B subunit in some set ratio. Assuming that mass peaks close in mass (relative to the monomer masses) results from of a loss of either monomer from one parental species, we can shift the CID mass spectrum by $M_w(A)$-$M_w(B)$ and $M_w(B)$-$M_w(A)$ to identify likely species that lost one B or A subunit, respectively.

The reason behind this is that if there are two species that originated from the same parental species by the loss of either A or B subunit, taking the mass of the (parent-A monomer) species, adding the mass of A and subtracting the mass of B will result in a hypothetical (parent-B monomer) species. Any overlap between the shifted and the original CID mass distributions will then help us identify a species that possibly lost a B subunit, and vice versa. We use the same reasoning to correct for (n-2) – loss of two consecutive monomers.



(a) Mass spectrum before CID correction. **(b)** Mass spectrum after CID correction.

Figure 3.18: CID correction applied to the deconvolved MS/MS spectra of the 1:1 HspB1-HspB5 mixture. After correction, the abundance of odd stoichiometries is comparatively lower compared to before correction. This is because we're correcting for both n-2 and n-1, where n-1 is giving rise to the abundance of odd oligomeric sizes before correction. Colours were chosen to show the m/z ranges for the MS/MS selection windows.

Combining information from native and tandem MS for intensity correction

Once we corrected the mass distributions for CID, we needed to combine the separate MS/MS measurements to obtain the total, complete mass distribution. To do this,

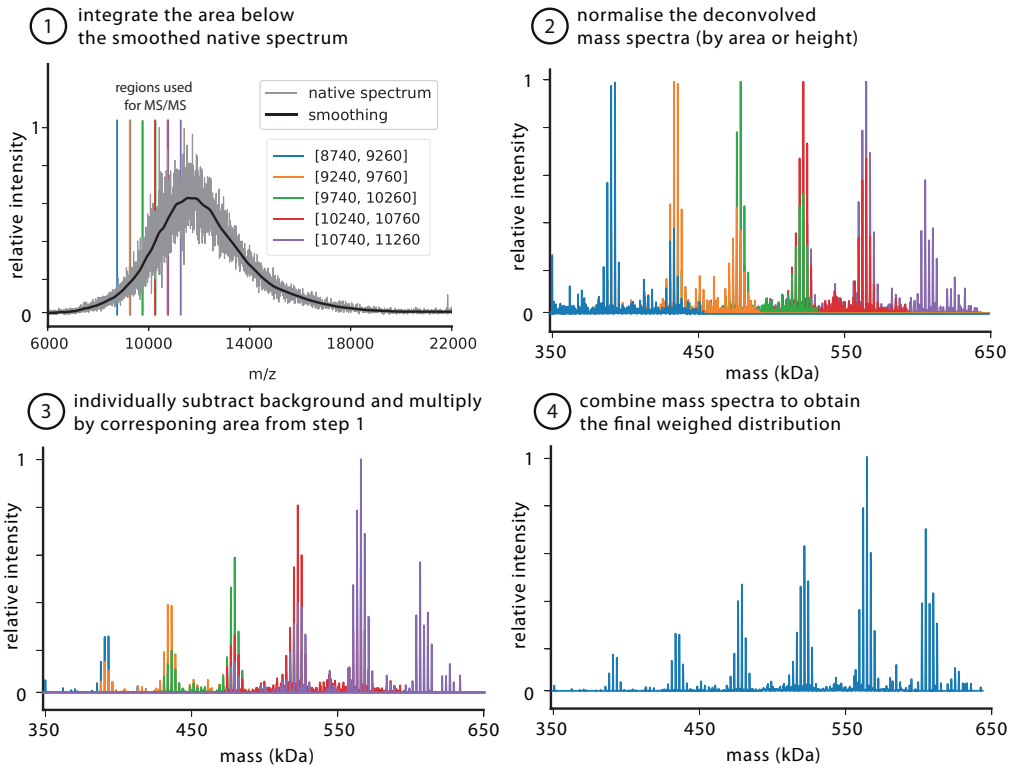


Figure 3.19: The rescaling algorithm shown for the 1:1 HspB1-HspB5 mixture. We first (1) smooth the spectrum using a gaussian filter and integrate the area corresponding to each individual m/z range under the smoother spectrum. In the next step, we normalise the MS/MS spectra by either height or area (in this figure, we show normalisation by height), after which we multiply (3) the normalised distributions by their corresponding areas in the native spectrum (1), and add them together (4) to obtain the final weighted mass spectrum.

we employed a strategy wherein we smoothed the corresponding native spectrum and integrated the area under the smoothed curve within the m/z range used for MS/MS. We then used the integrated values to multiply the CID-corrected MS/MS mass spectra before combining them together. The reasoning behind this is that the native spectrum contains information about the relative abundances of each MS/MS region–information we can't obtain from the MS/MS measurements themselves. This allowed us to obtain a skewed, reverse-U-shaped envelope. It should be noted that the envelope is asymmetric, and somewhat biased towards lower masses because the transmission, and thus signal intensity, in MS decreases with increasing m/z. The resulting MS spectra, together with MP mass spectra, can be found in Fig. 3.20.

Further, as there is inevitably an increase in the contribution from smaller species (due to their charge envelope distribution) to progressively higher windows in the native spectrum that is used for the correction, it is also possible the correction slightly overestimates the mass of the most abundant oligomer.

Mass photometry

The mass photometry spectra of the HspB1-HspB5 mixtures were recorded and processed the same way as the homo-oligomers with attention to getting well-resolved peaks if possible. Again, a comparison between the MS/MS and MP mass distribution can be seen in Fig. 3.20. We show the MP mass spectrum as a kernel density estimation (KDE) rather than a histogram. There are two reasons for this: first, the comparison with the MS mass spectra is somewhat clearer. Second, using the smooth KDE curve allowed us to run a code to find the best possible resolution for the peaks. For this, we used a simple algorithm to maximise the width of a 'median' peak multiplied by its height. Even for the highly complex mixtures, we could obtain reasonably resolved peaks for the 1:3 and 1:1 mixing ratios; this is possibly due to the fact that the relative abundance of odd stoichiometries is quite low (from MS), which might help with resolving peaks corresponding to even stoichiometries. Compared to the MS mass spectra, the MP mass distributions are centred on slightly lower mean masses for the 1:1 and 3:1 ratios. This might be because of imperfect MS/MS intensity correction as discussed in the tandem MS/MS section above. As expected from the MP homo-oligomeric mass distributions, there seem to be increasing amounts of unbinding events (not shown) with increasing HspB5 fraction in the oligomers, which results in the relative broadening and increased abundance of the lower-mass peak between below 200 kDa (Table 3.2).

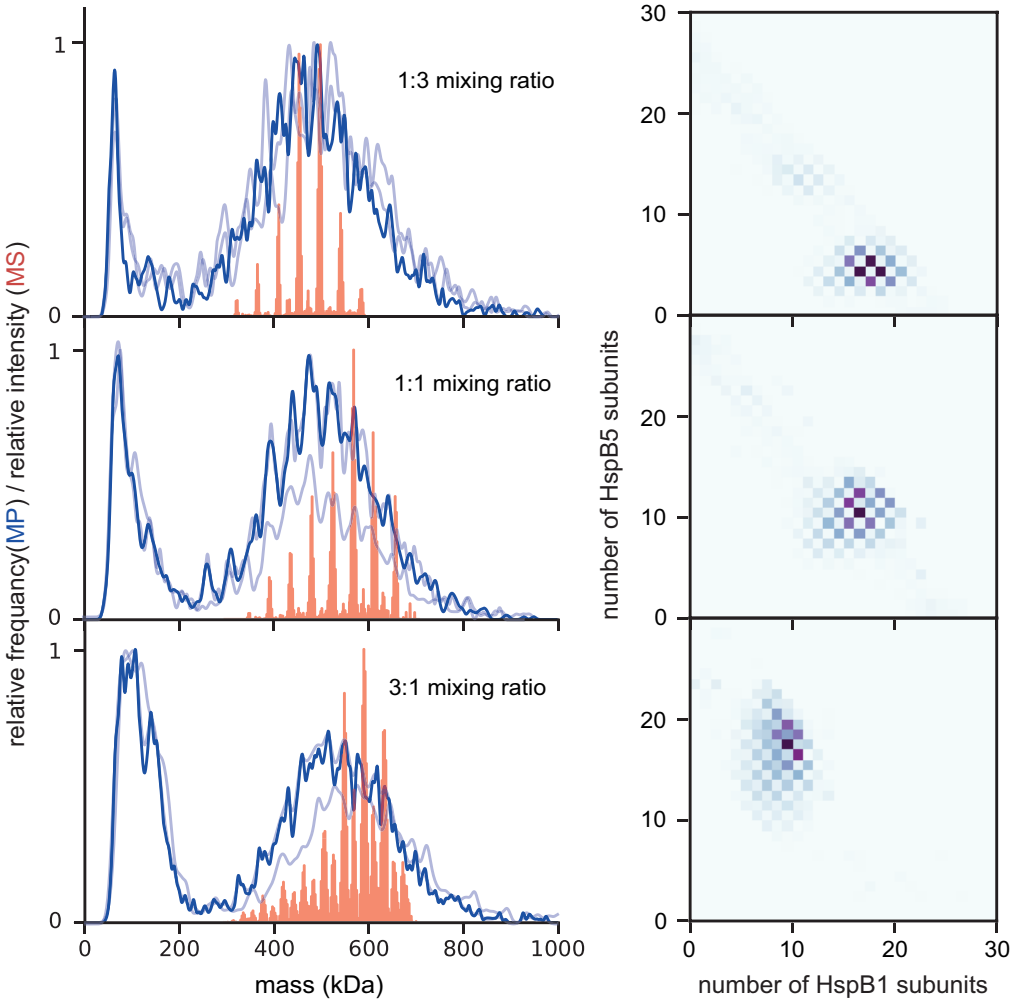


Figure 3.20: The comparison between tandem MS (red) and MP (blue) mass distributions on the left shows that tandem MS is biased towards lower masses. The imperfect overlap might be both due to the intensity correction in the lower mass region and decrease in transmission in the higher-mass region of the MS mass spectrum. The MP spectra are expected to be relatively unbiased. Repeats of the MP measurements are shown with reduced opacity. On the right are the mass distributions shown as 2D histograms. The distributions of sub-stoichiometries changes with the mixing ratio. The colour scheme for the 2D histogram is the same as in Fig. 3.4–darker colours correspond to a higher abundance.

ratio	binding (high)	binding (low)	unbinding	unbinding (fraction)
1:3	2588	374	98	0.033
1:1	1292	252	132	0.085
3:1	1320	572	583	0.308

Table 3.2: The number of binding and unbinding events for the three mixtures. The 3:1 mixture exhibits the highest number of unbinding events, perhaps because of the (relatively) higher contribution from a HspB5-like unbinding behaviour that we see with HspB5 homo-oligomers. High mass range is defined as masses above 200 kDa, whereas the low mass range are masses between 0 and 200 kDa

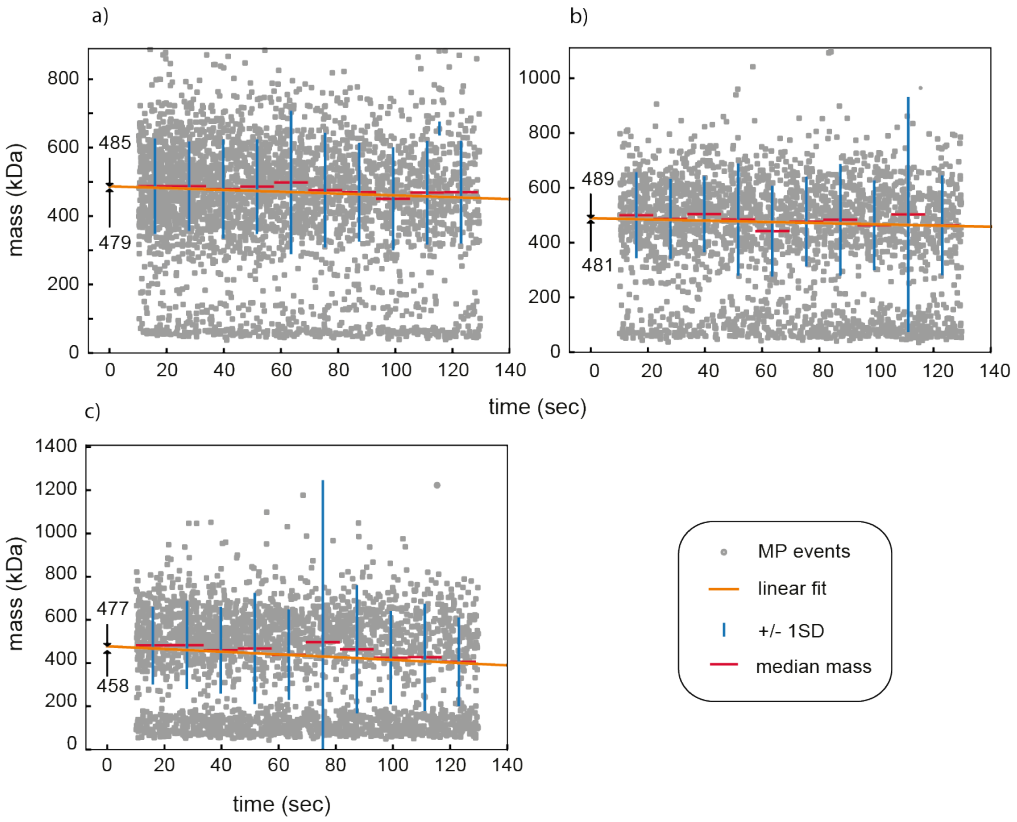


Figure 3.21: Time fit to the MP data for the HspB1-HspB5 mixtures with back-calculation to the time of dilution, on average about 10 seconds before we started the measurements. The back-calculated, fitted mass (upper arrow) and the median mass (lower arrow) do not differ, within a margin of error of the measurement (roughly one to two subunit masses), for the 1:3 (a)) and 1:1 (b)) and 3:1 (c)) mixing ratios. c) shows the biggest difference, although still only on the order of roughly one subunit mass.

Similarly to the homo-oligomers, we also checked the stability of the hetero-oligomers after dilution to 'MP concentration' (roughly 400 nM monomer concentration) by again fitting time bins (Fig. 3.21). For all three mixing ratios, the fitted and median masses were within roughly one monomer mass, suggestive that in the time frame of the measurements, the higher oligomers stayed stable. In the two most HspB5-rich mixing ratios (1:1 and 3:1), there is a gradual increase of smaller species with time–a phenomenon seen with HspB5 alone as discussed in the part of this chapter focused on homo-oligomers.

3.3 Conclusions

In this chapter, we have developed an experimental framework for studying the oligomerisation properties of highly polydisperse homo- as well as hetero-oligomers of the two most expressed human sHSPs, HspB1 and HspB5. We have shown how we can use tandem MS and intensity correction to obtain mass envelopes, and how these 'translate' into the 2D histograms. Further, as an unbiased, in-solution check, we have used mass photometry that, after optimisation, yielded MP mass spectra with resolvable envelopes even for the highly heterogeneous HspB1-HspB5 mixtures. Finally we compared both of these envelopes to find that the MS/MS distributions, even though they are higher resolution, are somewhat skewed in extent to the lower-mass region and slightly overestimates the most abundant oligomer compared to the MP envelopes which are roughly symmetric around the most abundant oligomeric size/mass.

3.4 Acknowledgements

I would like to thank Miranda Collier for recording the raw MS/MS spectra, and Mike Marty for his contribution to the CID correction algorithm and the development of UniDec ([106]), without which this work would not be possible. Finally, I would like to thank Josh Bishop for developing the particle-tracking algorithm we used to study the dissociation we see with MP and applying it to my MP data. I would also like to thanks Catie and Wiktoria for their feedback on this chapter.

3.5 Methods

3.5.1 Protein expression and purification

HspB5

BL21(DE3) competent cells (Novagen) were transformed with a pET23a(+) vector (Novagen) which had an insert coding for the full-length HspB5 protein. After DNA sequencing to ensure the transformation didn't produce any mutations, two one-litre flasks of Luria broth (LB) were inoculated with the transformed *E. Coli*. The cells were grown at 37 °C until the solution's optical density (OD) reached values between 0.6 and 0.8, after which expression was induced by adding IPTG to the final concentration of 1 μM. The cells were then grown overnight at 20 °C, harvested, flash-frozen with liquid nitrogen and kept at -80 °C before purification.

The purification process was described previously ([110]). Briefly, the cells were lysed by microfluidising in phosphate-buffered saline (PBS) with 5 mM β-mercaptoethanol (BME) at pH 7.4 with EDTA-free protease inhibitor cocktail (Roche) The lysate was then spun down at 20000xg (RCF) for 20 minutes. The clear portion of the lysate was filtered and the proteins were then purified by anion-exchange chromatography (IEX) followed by size-exclusion chromatography.

First, the clear lysate was loaded into two 5-mL HiTrap Q columns (GE Healthcare) connected in series and the protein was eluted using a salt gradient of 500-900 mM NaCl in the lysis buffer. After this, the appropriate collected fractions were concentrated and the concentrated protein solution was exchanged into the storage buffer (PBS at pH 7.4 with 1 mM DTT) by SEC using the Superdex 200/20 column. This also functioned as another, orthogonal purification step as only the purest fractions were then concentrated. The protein concentration was determined by absorbance at 280 nm using an extinction coefficient value of

9970 $M^{-1}cm^{-1}$ calculated from the protein sequence using the ProtParam website ([111]). The samples were then aliquoted and flash frozen with liquid nitrogen and stored at -80 °C.

HspB1

The full-length HspB1 protein was encoded in the pETHSUL vector (kindly donated by the Weeks group) containing an N-terminal His-tag with a SUMO sequence for digestion. The vector was used to transform BL21(DE3) competent cells (Novagen). Single colony was chosen and grown in a starter culture (10 mL of LB) overnight at 20 °C. Next, 500 μL of the starter culture was then used to inoculate 1 L of the ZYP auto-induction medium. Cells were expressed overnight at 24 °C with shaking at 180 rpm until OD reached values above 4, after which the temperature was lowered to 18 °C and the cells were left, again with shaking at 180 rpm, for 24 hours before they were harvested.

After harvesting, the cells were resuspended in roughly 50 mL 250 mM NaCl, 50 mM phosphate, 12.5 mM imidazole, 5 mM BME buffer with two protease inhibitor cocktail tablets (Roche) added, and lysed by microfluidising. The lysed supernatant was centrifuged for 20 minutes at 20000xg (rfc), and the clear fraction was filtered before being loaded onto two 5-mL nickel columns (GE Healthcare) connected in series, and then the protein was eluted using a gradient up to 500 mM imidazole. The resulting fractions were pooled and buffer exchanged overnight back into low-imidazole buffer with concurrent SUMO cleavage using the Ulp1 hydrolase (expressed and purified in-house). The hydrolysis step was done at 4 °C out of prudence to ensure minimal truncation. Next day, the sample was again loaded onto the column, but this time, the flow-through was collected. The flow-through was then concentrated and loaded onto the Superdex 200/20 column. This served

as a further purification step as well as a buffer exchange method. The protein was exchanged into a PBS buffer (pH 7.4) with 1 mM DTT.

3.5.2 Mass spectrometry

HspB1 and HspB5 were stored at -80 °C. On the day of measurement, the proteins were thawed and buffer exchanged into 200 mM ammonium acetate (pH 6.9) using BioRad biospin columns. For homo-oligomeric measurements, the proteins were then equilibrated at 37 °C for 60 minutes at 15 μM monomer concentration to ensure the recorded distribution was the equilibrium distribution at that temperature. Similarly, for the hetero-oligomeric 1:3, 1:1 and 3:1 mixtures, the mixtures were equilibrated again at 37 °C for 60 minutes, but at 60 μM monomer concentration before they were diluted by a factor of three before the MS measurements.

The data was collected using in-house gold-plated capillaries on a Q Exactive^{TM} mass spectrometer in positive ion mode with source temperature of 170 °C, capillary voltage of 1.4 kV, and S-lens RF 200 %. In-source trapping was set to -200 V for hetero-oligomers and HspB1, and -180 V for HspB5. Ion transfer optics and voltage gradient throughout the instruments were optimized for ideal transmission as follows: injection flatapole 10V, inter-flatapole lens 8V, bent flatapole 6V, transfer multipole 4V, C-trap entrance lens 3V. The RF MP2 and MP3 voltages were overridden to remain constant at maximum value (800+V). For the measurements shown, we used nitrogen as the collision gas as a heavier gas, argon, did not seem to result in a significant improvement to the resulting spectra.

Spectra were acquired with 10 microscans to increase the signal-to-noise ratio with transient times of 64 ms, corresponding to resolution of 17,500 at m/z = 200 for a transient time of 64 ms, and AGC target of $1.0x10^6$. The noise threshold parameter was set to 3 and the scan ranges used were and the spectra were recorded

in the m/z region between 12000 and 30000 m/z. The tandem MS spectra were recorded in sequential selection windows of 520 m/z.

Mass deconvolution was achieved using the UniDec software ([106]). Before deconvolution, the data were not smoothed, subtracted, binned, or otherwise pre-processed. We used a charge range of $10^+ - 50^+$. Further, the fitted peak FWHM was optimized for each spectrum (generally within 0.5 and 1 Th). All other UniDec parameters were left at their default values. The deconvolution was performed in the mass range of 100 to 800 kDa. After deconvolution, n-1 and n-2 charge series were assigned (as shown in Fig. 3.4), and the spectrum was corrected for CID.

After CID correction, the HspB1-HspB5 mixing spectra were rescaled, noise-subtracted, and added together as described in the text. 2D binning was done using mass windows of 200 Da centred at the expected mass of the sub-stoichiometries using sequence monomer masses.

3.5.3 Mass photometry

Microscope coverslips (No. 1.5H thickness, 24 x 50 mm, VWR (630-2603)) were cleaned by two 5-minute sonication cycles, first in pure milli-Q water, then in a 1:1 mixture of Milli-Q water and iso-propanol. The coverslips were dried with nitrogen before being stored in a dust-free environment before use.

Immediately before use, silicone gaskets (CultureWell™ reusable gasket, 3mm diameter x 1 mm depth, Grace Bio-Labs) were attached to the microscope coverslips and the quality of the cleaning process was checked by recording a blank data acquisition. After making sure that the gaskets were clean, the laser focus was found using 30 μL of the appropriate buffer, 200 mM ammonium acetate in this case, after which 25 μL of the buffer was removed from the well and replaced with

25 μL of the sample solution. This resulted in a lag time of approximately 10 seconds between sample dilution and the start of MP measurement.

Samples were equilibrated at the same ratios and concentration as in the mass spectrometry experiments (1 hour at 37°C). Samples were diluted to 400 nM monomer concentration immediately before measurement. The experimental MP setup was the same as described previously ([112]).

4

Quantitative description of the HspB1-HspB5 co-assembly

Contents

4.1 Motivation

In the previous two chapters, we developed a model of co-assembly of polydisperse proteins (Chapter 2) and an experimental methodology (Chapter 3) to study these co-assemblies. In this chapter, we apply the model to find statistically non-significant energy biases in the HspB1-HspB5 co-assemblies with a possible preference for hetero-dimerisation. We then prove this bias experimentally using peptide binding experiments and MP measurements. Finally, we use molecular dynamics to complement the experimental results.

4.2 Introduction

4.2.1 Overview of the experimental results

In the previous chapter, we showed that we can use tandem MS to obtain a detailed description of the HspB1-HspB5 co-assemblies, and that simply by changing their mixing ratio, we can change what kinds of co-assemblies they form. This might suggest the co-assembly is not strongly preferential for neither self- nor an obligate n:n co-assembly. We further showed that due to practical hurdles, we are not able to use MS/MS to probe the whole native MS mass envelope. We then used mass photometry to obtain complete, unbiased mass envelopes for the mixtures which clearly showed that the MS/MS-derived mass envelopes were relatively narrow compared to the MP-derived envelopes. This could be due to imperfect abundance correction in the lower-mass region and poor oligomer transmission on MS in the higher-mass region. To overcome this, we now use the kinetic model of polydisperse co-assembly to combine both the MS/MS and MP measurements and extract any co-assembly energy biases and explore the 'directionality' of these interactions using further MS and MP measurements.

4.3 Results & discussion

4.3.1 Strategy employed to obtain the complete distributions

To obtain the complete mass distribution, we combine the MS/MS and MP mass distribution using the chemical kinetic model.

The strategy we employed is as follows–We only fit the high-resolution MS/MS data in the region where both MS/MS and MP agree; this is essentially the whole region below the most abundant mass on MP. The resulting, 'cut' MS/MS mass distributions used for further fitting, and how they compare to the MP envelopes, can be found in Fig. 4.2. As explained in more detail further, we intensity-corrected using by both normalising the individual MS/MS mass distributions by area (left) and height (right) to see how well the mass distributions compare to the MP measurements. We then fitted distributions corrected by both methods and found that final fitted results did not vary significantly between the two normalisation methods.

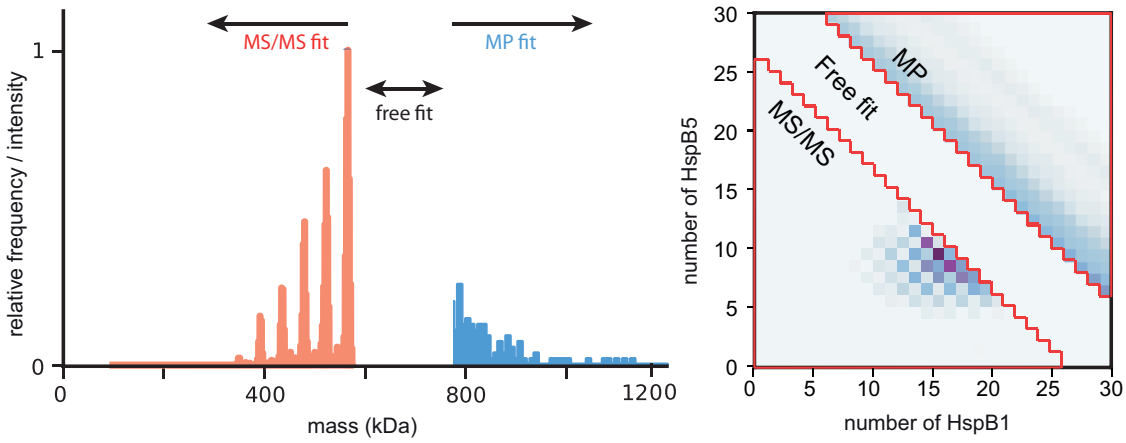


Figure 4.1: Contributions from the MS/MS data and MP data for fitting. In the lower-mass region (red), we fit the higher-resolution MS/MS data. To restrict the fit in the higher-mass range, we use the fraction of oligomers from MP (blue). The fit is free to vary between these two regions. Shown both on a mass axis on the left as well as on the 2D histogram on the right.

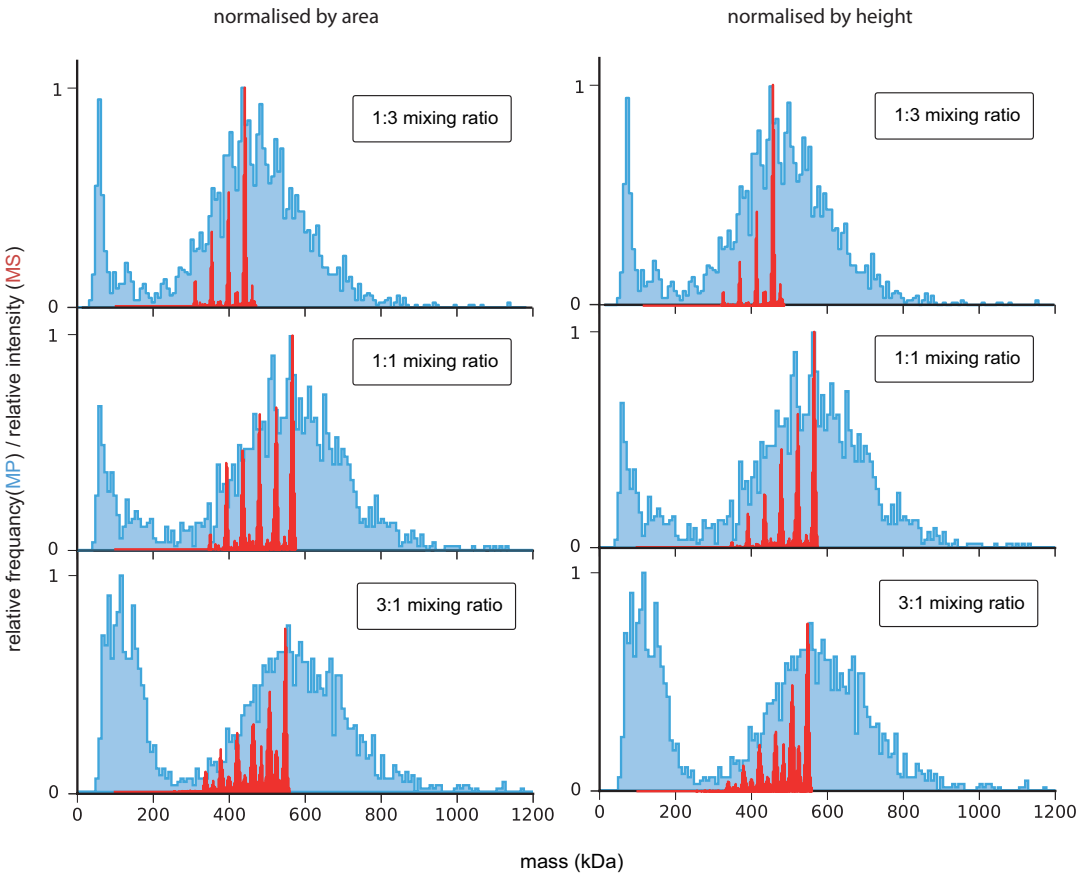


Figure 4.2: Comparison between MS/MS normalisation by area and height. The normalised MS/MS spectra are shown in red and the MP histogram for a comparison are in blue. For the MS/MS spectra, we only show the regions used for fitting. We used both methods for separate fitting as shown later.

As shown in Chapter 2 (pages 67 and 68), the co-assembly between two polydisperse proteins causes changes in the abundance of odd stoichiometries that are not easily explainable as, for example, the linear combination of the biases coming from each of the two homo-oligomers. It also seems that upon mixing, there is a trend for the odd stoichiometries to become less populated (Fig. 2.23).

These trends can be seen in our experimental data for the HspB1-HspB5 mixtures, too (Fig. 4.3). Relative to the homo-oligomers, the hetero-oligomers seem to be more biased towards even stoichiometries. There also seems to be an increase in the relative abundance of odd stoichiometries with an increasing HspB5 mixing ratio.

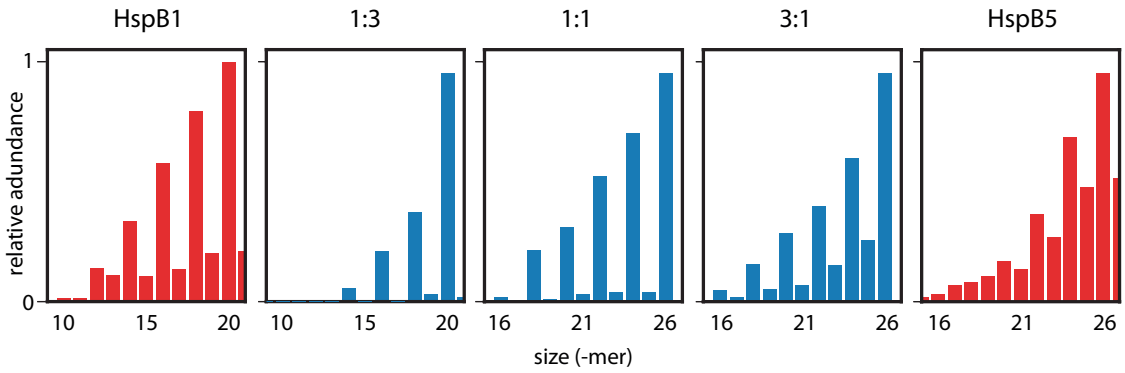


Figure 4.3: There is a higher even:odd bias in the HspB1-HspB5 mixtures compared to homo-oligomers. As shown in Chapter 2, the even:odd bias is not a simple linear combination of the two homo-oligomer biases.

4.3.2 Data preparation and fitting

The MS/MS data were rescaled and combined as explained in the previous chapter, with the rescaling done both by area and height. In the case of rescaling by height, we take the deconvolved mass distributions and normalise them so that the most abundant oligomer has the abundance of 1 before multiplying the individual mass spectra by the area under the curve in the corresponding region of the smoothed native spectrum and adding them together. Similarly, when we rescale 'by area', we normalise each MS/MS mass spectrum by making sure that the sum of the mass distribution is equal to 1 before we, again, multiply the normalised mass spectra by the corresponding area in the native spectrum and add them together. We then cut the high-resolution mass envelopes so that only the region below the most abundant mass on MP was left for fitting. This translates to the masses of 475, 575, and 559 kDa for the 1:3, 1:1, and 3:1 mixing ratios, respectively. The fit was then allowed to vary penalty-free between these masses and a hard cut-off corresponding to 35-mers in size. From the MP experiments, we know that the fractional abundance of oligomeric sizes above (and including) 35-mers should be at most 0.095 for all mixing ratios. The fraction was found by summing the number of binding events with a

mass above 730 kDa and then dividing this number by the number of all binding events with a mass higher than 200 kDa. We chose the lower cut-off of 200 kDa for two reasons: firstly, we are only able to model one mass envelope, which in general is the higher-oligomer mass envelope; secondly, by only counting the events above 200 kDa, we can be ambivalent about the smaller species that might be the product of glass-induced dissociation from higher oligomers as shown in the previous chapter.

Fits whose fractional population above this limit was higher were essentially rejected by having a high penalty associated with them. We then used an error-squared penalty for the MS/MS region of the fit. All eight on- and off-rates were fitted for all three ratios simultaneously by least-squares fitting.

4.3.3 Fitted distributions correspond accurately to the experimental data

The final best fits to all three ratios can be seen in Fig. 4.4. The left half shows again the input distributions used for fitting (MS/MS in red and MP in light blue), and the resulting fits as mass distributions (orange), or 2D histograms (right half of the figure). Even though we didn't use the MP distribution explicitly for our fits–we only used the fraction of oligomers higher than the 730 kDa threshold to reject distributions that were too large, it seems the fitted mass distributions match the MP data well.

The fitted 2D distributions also seem to match the experimental 2D distributions (Fig. 3.20) well. Compared to the experimental MS distributions, the fitted distributions are slightly broader at the sub-stoichiometry level. This could be due to multiple reasons: first, the spread of oligomeric sizes and the spread of sub-stoichiometries are inherently linked, although non-trivially, through the simultaneous fitting to three different ratios. Since the MP forces the fit to be

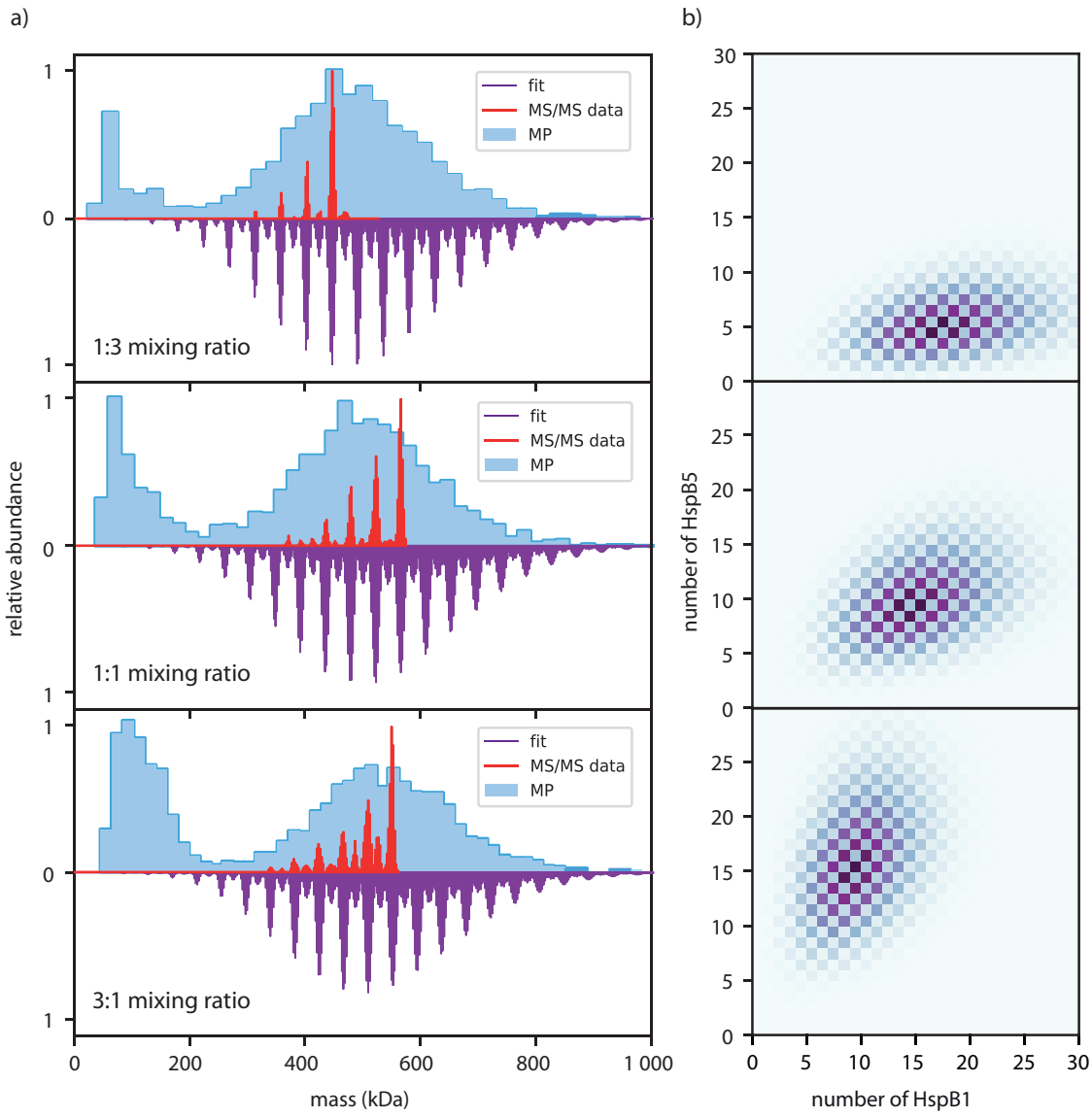


Figure 4.4: Fitting inputs and outputs for the HspB1-HspB5 co-assemblies. a) shows the inputs: MS/MS data (red) and MP data (blue) and the resulting fitted mass distributions (purple) for the three mixing ratios. b) shows the fits as 2D histograms.

roughly between 200 and 800 kDa, this results in the relative broadening of the sub-stoichiometries. When we only fit the MS distributions without any MP input, the result is a distribution that matches the MS/MS data better, but are much broader along the size axis, having maxima at roughly 1.5 MDa–a result that is incompatible with the MP measurements, as well as results described elsewhere ([113]). Secondly, the MS/MS data itself could have been made artificially narrower by background

subtraction–we subtracted the background from each MS/MS region individually, before combining the mass distributions for rescaling, but each subtraction was done by subtracting a constant minimum until most of the noise disappeared, which inherently makes the sub-stoichiometry distribution narrower (Fig. 4.5).

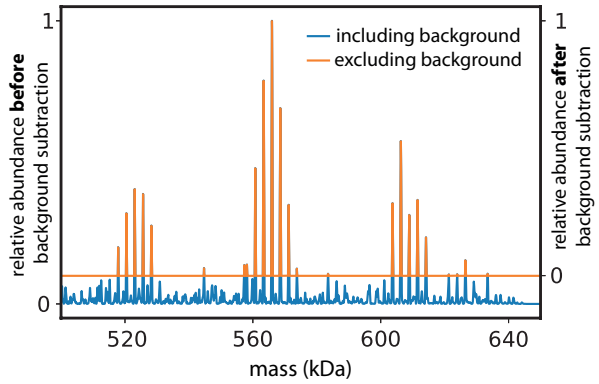


Figure 4.5: Background subtraction causes narrowing on the sub-stoichiometry level. yt subtracting a constant background level, we somewhat artificially narrow the the mass distributions used for fitting.

To further explore the fit to the MP distributions, we devised a method (Fig. 4.6) wherein we simulated (a) what the resulting MP distributions (b) might look like. To do this, we first sample a given number of events (that is the same as the number of recorded MP landing events in the higher-oligomeric region above 200 kDa) using the fitted distributions as probability distribution from which we sample (Fig. 4.6 a; upper and middle plot), and then broaden the sampled events to obtain a mass distribution with resolution similar to MP. For these simulations, we used a Gaussian broadening with $\sigma = 18$ kDa. We found this value from comparing the narrow MS and broader, but still resolved MP HspB1 mass distributions (Chapter 3, Fig. 3.12).

We repeated this process thirty times to obtain a confidence interval band (orange band in Fig. 4.6 b) for each mixing ratio. The simulated MP seems to be slightly broader than the measured MP histograms, but there was a good agreement on the mass of the most abundant species. The relative broadness could be due to the processing of the MP movies since using different filter values (Chapter 3, Fig. 3.8) slightly changes the shape of the recorded mass distributions.

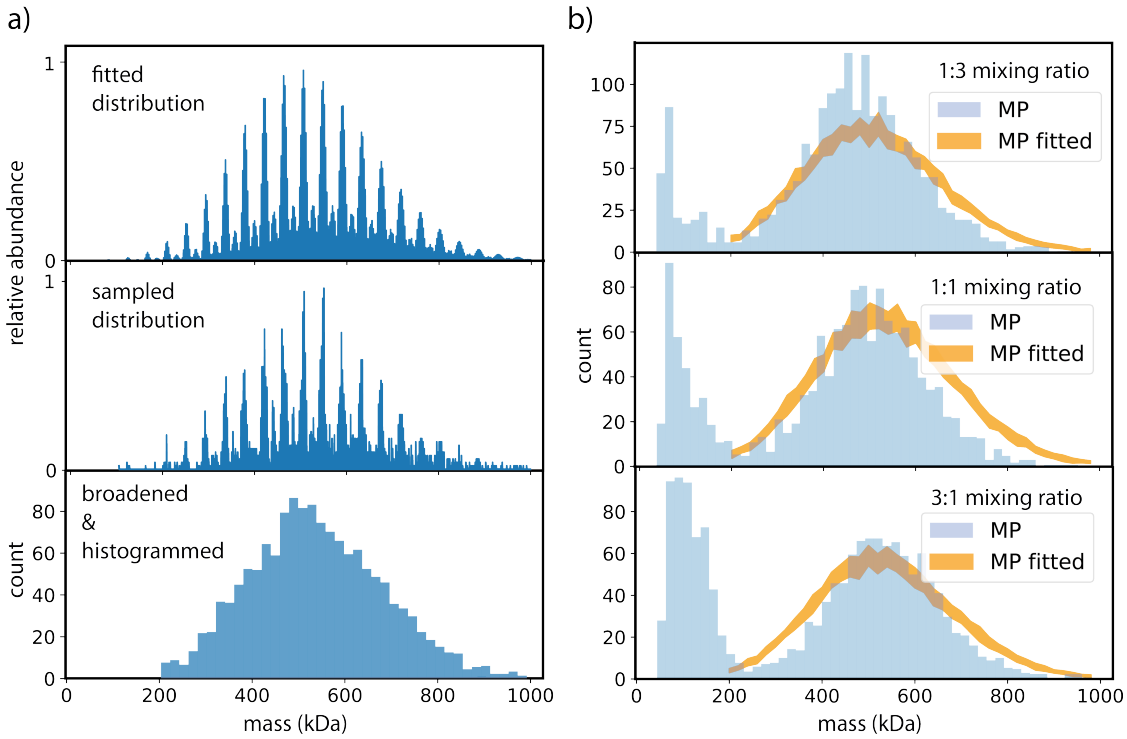


Figure 4.6: Strategy for simulating MP from the fits (a) and comparison between the fitted and measured MP envelopes (b). In a), we first sample a limited number of events as realistically with MP we only record between a thousand to two thousand events, and it is not expected we sample the whole distribution completely. We then broaden and histogram the sampled mass distribution to 'simulate' an MP histogram. We do this multiple times to obtain a confidence range for the simulations to account for the limited sampling. The overlap between the simulated and measured MP distributions can be seen in b). In general, the distributions seem to agree well. The simulated MP distributions are slightly broader for all three mixing ratios compared to the measured distributions, but this might be due to the averaging and filters used to obtain the measured Mp mass enveloped as explained in Chapter 3.

Another possible explanation is that the distribution does not have as strong even:odd preference as the MS/MS data suggest. As discussed in Chapter 2, the dependence of the even:odd bias is not straightforwardly linear. Even so, we might be underestimating the amount of odd stoichiometries with imperfect CID correction or background subtraction as explained above, making the even:odd skew larger, and since the oligomeric distributions tend to be broader with a bigger even:odd skew Chapter 3, Fig. 3.6, this might be another contribution to the relative broadness of the fitted distributions.

4.3.4 The mixing between HspB1 and HspB5 shows a possible, statistically non-significant bias towards co-assembly

As the next step, we extracted energy biases from the kinetic modelling. To obtain confidence intervals for the goodness of fit of the distributions to the model, we performed Markov-chain Monte Carlo (MCMC) ensemble sampling. The only priors used set the maxima for the different on- and off- rates as described in the Methods section. We also modified the steps to account for the fact that ultimately, the energy bias is a function of the ratio between the different off-rates, and so simple $position - step$ and $position + step$ don't result in the same absolute energy difference from the initial position. To account for this, we modified the steps as a negative exponential function of the given step which resulted in more optimal steps in the MCMC walk.

We can then define the hetero-dimer stabilisation energy as:

$$\Delta\Delta G_{stab} = \Delta G_{BA} + \Delta G_{AB} - \Delta G_{AA} - \Delta G_{BB} \tag{4.1}$$

In terms of the kinetic rates, $\Delta\Delta G_{stab}$ can be defined as:

$$\Delta\Delta G_{stab} = -RTln\left(\frac{k_{AA}^{-A}k_{BB}^{-B}}{k_{AB}^{-A}k_{AB}^{-B}}\right) \tag{4.2}$$

since we effectively have to combine the Gibbs energy change associated with these two processes:

$$(AA) \underset{k^{+A}}{\overset{k^{-A}{}_{AA}}{\rightleftharpoons}} (A) \underset{k^{-B}{}_{AB}}{\overset{k^{+B}}{\rightleftharpoons}} (AB)$$

Figure 4.7: (AA)-homo-dimer to (AB)-hetero-dimer swap.

and

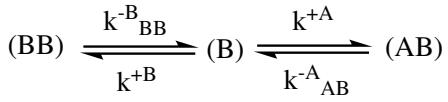


Figure 4.8: (BB)-homo-dimer to (AB)-hetero-dimer swap.

The result of applying this formula to the MCMC walk can be seen in Fig. 4.9. Left and right halves show fits to MS/MS distributions normalised by area and height, respectively.

The hetero-dimer stabilisation energy is slightly negative, meaning that the hetero-dimers are possibly marginally more stable within the mixed oligomers, but still, the possibility of no energy difference is well within the 95 percent confidence interval. It is quite remarkable, that even approximately 450 million years after their gene-duplication event, these two proteins still co-assemble in virtually the same fashion as would be expected right after the duplication event. We explore this by using phylogenetics and MD further in this chapter. It also seems that any hetero-dimer stabilisation energy is coming preferentially from the stability of HspB1 in the hetero-dimers (Fig. 4.9 red violin plot–marked (B1-B5)), as the stability of both homo-dimers, (B5-B5) and (B1-B1), and well as the (B5-B1)-directional hetero-interaction seem to be comparable and the (B5-B1) interaction contributes only slightly to the hetero-dimer stabilisation energy.

Still the confidence intervals on these fits are quite large. This might be coming from the way the MCMC chains 'walks' around the energy space–even though we tried to correct for this by modifying the steps as explained above. Further, the error function we used for the walk might be biasing the walk towards exploring a much broader energy space as we used the absolute difference as the basis of the error function rather than e.g. the square of the difference. Using a square of the difference would presumably make the explored space narrower around the best fit.

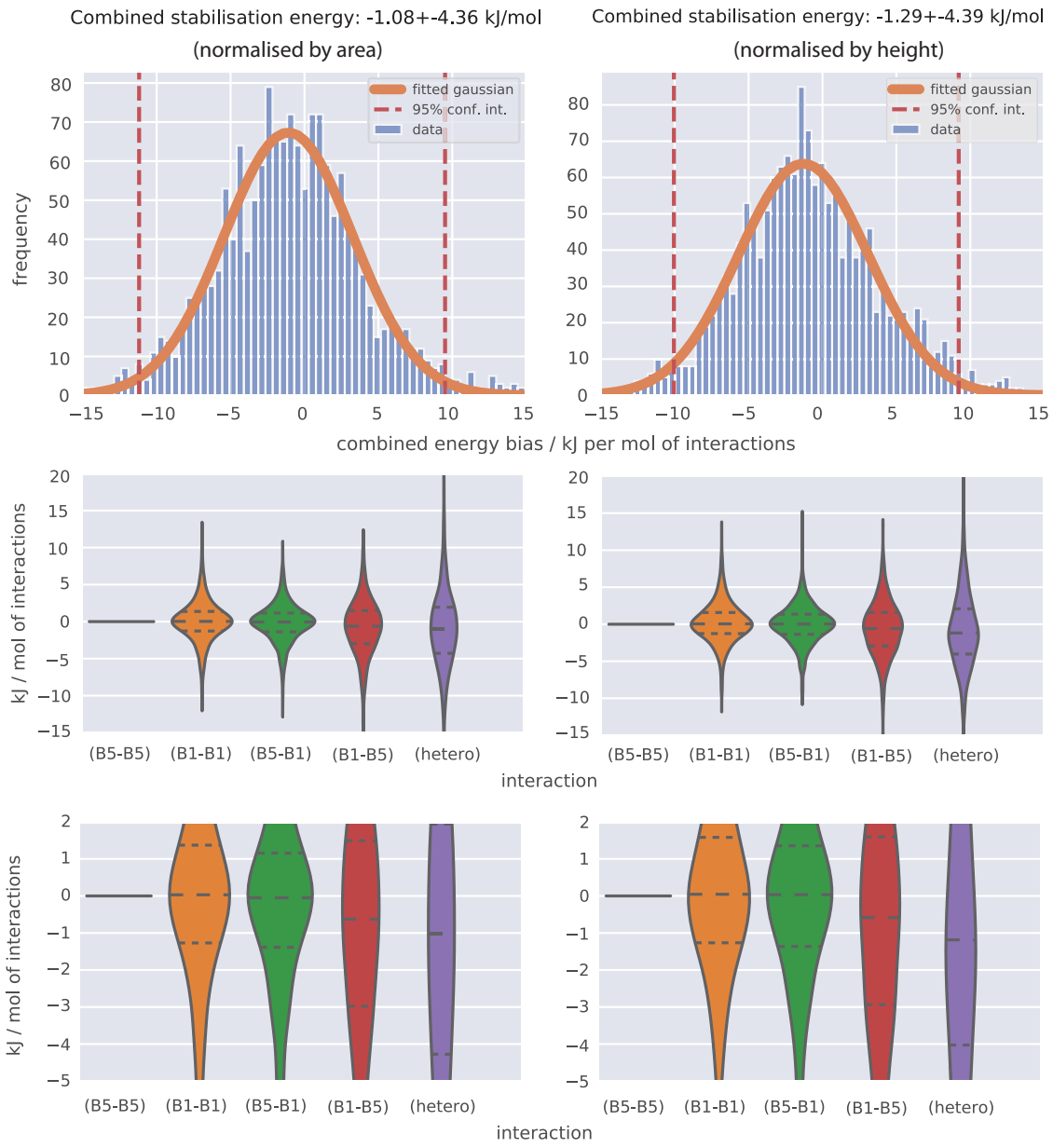


Figure 4.9: The combined stabilisation energies and contributions from the differential stability of the hetero- and homo-dimers for all three mixing ratios combines. Compared to the (B5B5)-homo-dimer, only the directional (B1-B5)-hetero-dimer interaction is slightly different. This suggests that any bias in the HspB1-HspB5 co-assembly might be directional. We show fits do MS/MS distributions normalised by both area and height, but don't find statistically significant differences.

Peptide binding experiments show that the HspB1 CTD-mimicking peptide slightly prefers the HspB5 aCD than its own

As the confidence interval on these fits was quite large, we next wanted to explore whether we can observe any slight bias experimentally. For this, we first mixed

HspB1 and HspB5 aCDs to find they dimerise without any preference as at three different mixing ratios (2:3, 1:1, and 3:2), the fraction of a hetero-dimer was, within a margin of error, close to a binomial fraction–suggestive of energetically unbiased aCD-aCD hetero-dimerisation (Fig. 4.10 a). If there is any bias, it presumably comes from the interaction of either the NTDs or CTDs with the core domain. As the NTD is either intrinsically disordered ([114]), or highly flexible ([115]), we decided to only study the contribution from the CTDs. For this, we used 4 peptides. Two peptides (which we further call Ctp–the C-terminal peptide) had the exact sequence of a part of the CTD from HspB1 and HspB5; the remaining two were the Ctp sequences scrambled. This served as a control for the specificity of these interactions. The peptide-binding titration spectra, and the resulting titration curves can be seen in Fig. 4.10 b) and c), respectively.

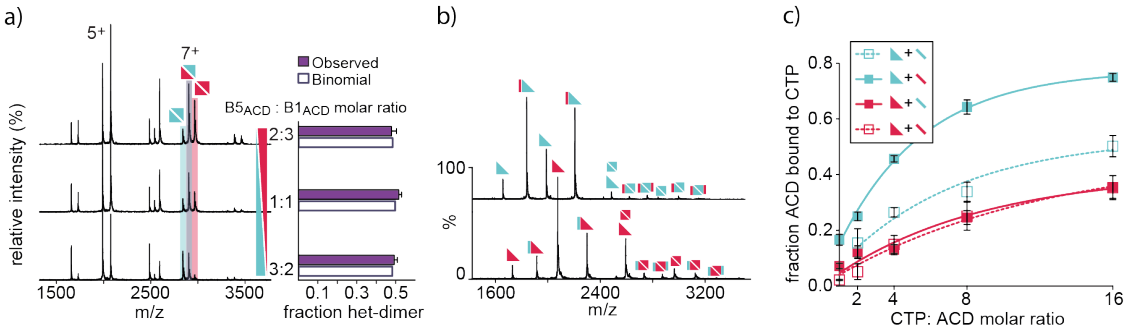


Figure 4.10: The aCD-Ctp titration experiments show slight preference in aCD-Ctp binding. First (a), we mixed the two aCDs to find that they co-assemble in a binomial fashion. We then performed peptide binding experiments (b) and (c). The mass spectra for the cross-protein (HspB1 aCD and HspB5 Ctp, and vice versa) binding experiments can be seen in (b). The abundances for the dimeric species (roughly around 3000 m/z) were quite low, so we used the monomer abundances to fit titration curves (c) at five different aCD:Ctp ratios. Figure adapted from [109] with permission.

From the measured dissociation constants, K_d, we can then calculate the as $= -RTln(K_d)$. The resulting Gibbs energies as well as their difference, $\Delta\Delta G$ can be found the in the table below.

	HspB1(aCD)	HspB5(aCD)
HspB1(CTp)	-21.8±0.3	-26.4±0.1
HspB5(CTp)	-22.0±0.1	-23.8±0.1
difference	-0.2±0.4	-2.5±0.3

Table 4.1: Difference in Gibbs energies extracted from the peptide-binding experiments.

The binding experiments show that, indeed, the Ctp-aCD might be biased with the HspB1 Ctp showing a slight (-2.5 ± 0.3 kJ/mol) preference for the HspB5 aCD compared to the HspB5 Ctp.

MP distributions as a function of mixing ratio and equilibration concentration give further credence to the binding bias

Since the MP mass distribution should be less biased compared to the tandem-MS-derived mass distribution, we decided to test the experimental prediction from the Chapter 2-which is that the mixing ratio has an effect on the oligomeric size that may not be simply linear, and also that the main oligomeric size is dependent on $k^+[M]$, where k^+ in the on-rate for monomers of concentration $[M]$; this means that decrease in the concentration we equilibrate the mixtures at should result in progressively lower average masses of the oligomeric distributions. We performed a set of four such MP 'scanning' experiments for monomer equilibration concentrations from 1 to 20 μM and 11 different mixing ratios. The results can be found in Fig. 4.11.

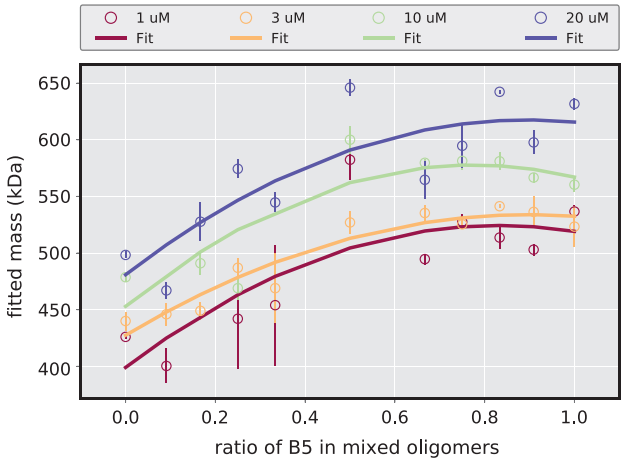


Figure 4.11: The HspB1-HspB5 co-assemblies show dependence on both the equilibration concentration and the fracion of HspB5 in the oligomers. There is an initial steep increase in the average size followed up until roughly 0.5, followed by a relatively flat region between 0.5 and 1.0 HspB5 fraction. We fitted quadratic trendlines to each concentration to show the flattening in mass differences above the 0.5 mixing fraction.

We can indeed see a dependence of the average measured mass on equilibration concentration for almost all mixing ratios, where the average measured mass shifts towards higher values with increased equilibrium concentration as expected since the on-rates are pseudo zeroth order ($k^{+}[M]$, where the effective rates depend on the monomer concentration), and the average size of the co-assembly is inherently linked to the two on-rates for proteins A and B.

The dependence of the average mass on the mixing ratio is slightly more complicated-there is an increase in average mass up until the roughly 1:1 mixing ratio, after which the sizes stay almost constant. This behaviour could be explained by the fact that the hetero-dimer interactions are 'asymmetric', as we've shown with both the peptide-binding experiments and chemical kinetics modelling. This property is also well-reflected in molecular dynamics (MD) simulations we show in the next section.

Next, we wanted to see whether we can extract the hetero-dimer stabilisation energies using on- and off-rates for homo-oligomers and MP average sizes for mixtures only. Since the hetero-dimer stabilisation energy is only a function of off-rates, there should not be any dependence on the equilibration concentration; the fitted energy bias should be the same for all four equilibration concentrations.

To test this assumption, as well as to see whether extracting energy biases only from MP data for the co-assemblies, we extracted k^{+A}, k^{+B}, k^{-A}, k^{-B}, k_{AA}^{-A}, k_{BB}^{-B} purely from MS-derived homo-oligomer distributions (Fig. 3.5), and fitted k_{AB}^{-A} through the calculated average masses to the MP data set.

The result can be seen in Fig. 4.12. The fitted average masses seem to recapture the data set quite closely. Surprisingly, the fitted hetero-dimer energies (-1.09 ± 0.19 kJ/mol of interactions) correspond well to the energies extracted by fitting the detailed MS/MS mass distributions (-1.08 ± 4.36 kJ/mol and -1.29 ± 4.39

kJ/mol for normalisation by area and height, respectively). Note that the confidence interval on the fit to MP average masses is smaller than the corresponding error to the MS/MS fit. It might thus be possible in the future to combine both fitting methods to narrow the confidence interval on the fits.

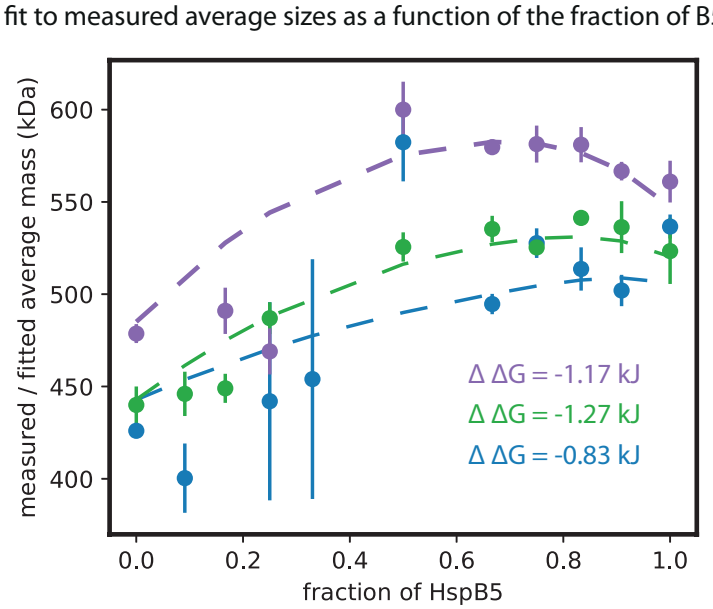


Figure 4.12: Hetero-dimerisation energy biases extracted by fitting the average sizes from MP at three different equilibration concentrations–1 μM (blue), 3 μM (green), and 10 μM (purple). Experimental data shown as data points, and fits are shown as dashed curves. As expected, the three extracted biases are quite close to each other, and also remarkably close to the bias extracted from the MS/MS data.

MD simulations

As the next step, we wanted to explore the possible reasoning behind the fact that the stabilisation energies were so close to the case of no bias as well as to rationalise the peptide-binding experiments further. To do this, we ran MD simulations of various combinations of aCDs and Ctps from HspB1 and HspB5. The simulations were based on crystal structure published before and were run for 300 ns for all combinations. Only the last 100 ns was then used for analysis. A more detailed description of the MD simulations can be found in the Methods section.

We chose the averaged frequencies of contact between the various Ctp and aCD residues with a cut-off of 4 Angstrom as a measure of affinity the various Ctps and aCDs have to each other. Using this metric, it seemed that while the

averaged frequencies are quite close to each other for all cases, there might be slightly stronger interactions between the HspB1 Ctp and HspB5 aCD–in a perfect agreement with the fitting, MP and MS data.

To visualise and explore this behaviour in more detail, we constructed a 2D histogram with the contact frequencies for all amino acid pairs for the Ctps and aCDs as shown in Fig. 4.13. The top two plots show the interaction frequency between the HspB1 peptide, and the HspB1 (top) and HspB5 (middle) ACDs. In these two plots, the arrows point to the region that differs to an appreciable extent between the two simulations. The bottom plot shows the projection of these frequencies on the two aCDs. The only quantitative difference between the two maps seems to be in the position corresponding to the residues 112 and 82 for HspB1 and HspB5 aCDs, respectively. The two residues on the HspB1 Ctp that show the biggest change between the two aCDs are 185 and 186. The extra stabilisation afforded to the HspB5 aCD and HspB1 Ctp could thus be coming from a possible hydrophobic interaction between the said residues in the HspB1 Ctp and the HspB5 aCD.

We further attempted to simulate the interactions of both HspB1 and HspB5 with their last common ancestor (LCA). We found the sequence for the LCA using phylogenetic reconstruction and then modelled the starting structure based on the published structures of HspB1 and HspB5. The structure was, however, highly flexible–pointing to a possibility that either the sequence was not reconstructed properly or the starting structure was not suitable for the simulation. There were no significant differences in the behaviour of the LCA aCD otherwise, both with Ctps and other aCDs other than the fact that the HspB5 Ctp binds quite strongly to the LCA aCD. This points to a slow evolution of the interface between these two proteins.

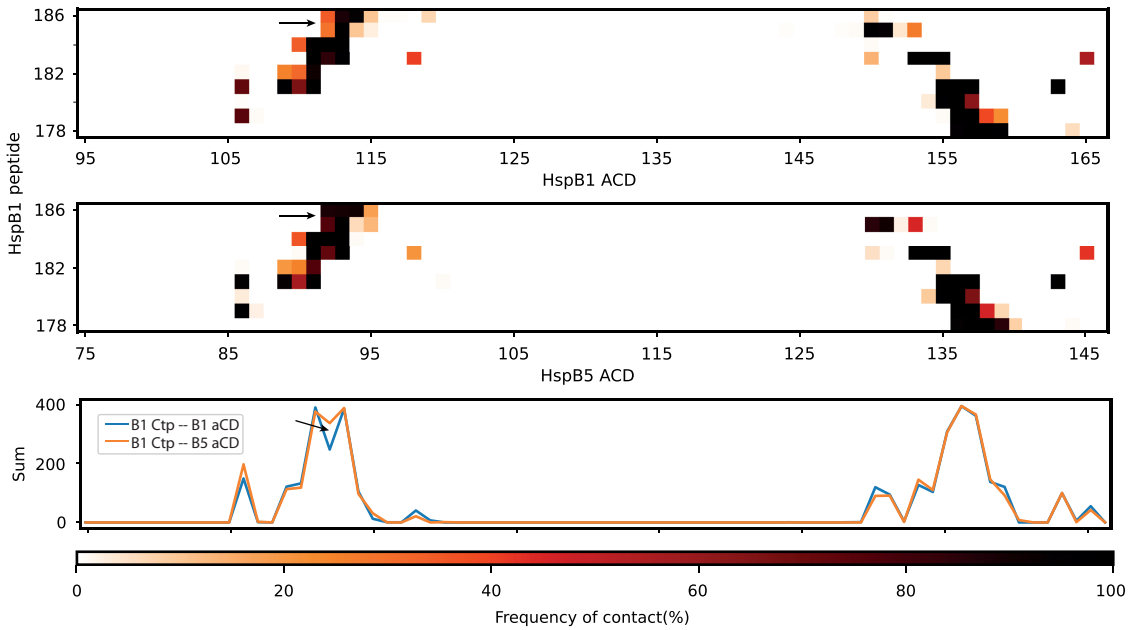


Figure 4.13: Frequency-of-contact maps for the HspB1 Ctp and HspB1/HspB5 aCDs. The frequency of contact between the HspB1 Ctp and the two aCDs, HspB1 aCD (upper) and HspB5 aCD (middle) shows a region (arrow) that might be responsible for the differential average frequency of contact between the two aCD-Ctp combinations.

4.4 Conclusions

In this chapter, we focused on the physico-chemical description of the HspB1-HspB5 co-assemblies. We first applied to the chemico-kinetic model of polydisperse co-assembly to extract any energy biases, and found that although not statistically significant, there might be some directional preference for co-assembly (or hetero-dimerisation). To study this possible 'directionality' of the interactions in more detail, we performed peptide/aCD-binding MS experiments and MP measurements.

In the binding experiments, we first show that the two aCDs mix non-preferentially with each other, suggesting that any 'directionality' in the hetero-dimer interactions might be caused by either the NTD or CTD. Since the NTD is disordered, and any peptide-binding would be difficult, we decided to probe the binding between CTD-mimicking peptides and the aCDs to find that HspB1 CTD-mimicking peptide indeed slightly prefers the HspB5 aCD than its own.

We also measured the average masses of different mixing ratios equilibrated at different concentrations to find that between the HspB5 mixing ratios of 0 to 0.5, there is an increase in the average mass after which the average mass changes are comparatively minimal. This suggests that the off-rate for HspB1 from the hetero-dimer is saller than from its own homo-dimers, this results in the increase in stability with the increase in the amount of hetero-dimers. Further, the off rate for HspB5 from hetero-dimers might be close to that from its own homo-dimers, which eplain the relative flatness of the curve once the oligomers start saturating with HspB5. This again points at some 'directionality' in the hetero-dimer interactions. There is a positive correlation between the average mass and equilibration concentration as perhaps expected. We further fit each equilibration concentration and saw that we could extract energy biases by only using the MS distributions of homo-oligomers, and MP distributions of the mixed co-assemblies. This might help narrow the confidence interval on the fitted energies extracted from MS/MS.

Finally, we asked the question of whether the two proteins co-evolve together to ensure co-assembly seemingly without any energy bias towards either self- or co-assembly. To explore this further, we obtained the sequence of the HspB1/HspB5 ancestor before the gene duplication event that gave rise to the two extant proteins, and performed MD experiments to study the binding between the extant and the ancestral aCDs and peptides. We found that the preference changes are minimal pointing to a very slow evolution of these two proteins.

4.5 Methods

4.5.1 Protein expression and purification

HspB1 & HspB5

For the peptide binding experiments, the HspB1 and HspB5 proteins were expressed and purified as described in Chapter 3.

Peptides

Peptides (HspB1$_{Ctp}$, HspB5$_{Ctp}$, HspB1$_{Ctp-ctrl}$, and HspB5$_{Ctp-ctrl}$) were purchased from Biomatik at 99+% purity. Lyophilized peptides were resuspended in deionised water to a stock concentration of 1 mM, flash-frozen, and stored at -20 °C until use. Peptides were diluted in ammonium acetate prior to mixing for MS experiments.

4.5.2 Mass spectrometry

For aCD-Ctp mixtures, aCD concentrations were held constant at 5 µM and peptide concentrations were varied from 5 to 80 µM (corresponding to mixing ratios of 1:1, 1:2, 1:4, 1:6, and 1:16 of aCD to peptide). Populations of free and peptide-bound monomers for construction of comparable binding curves were determined by deconvolution of raw spectra with the m/z range restricted to include only 5+ and 6+ monomer charge states. The upper bound of the mass was set to 11.3 kDa. We did not analyse the dimer region as the signal in this region was not intense enough for a confident analysis. Deconvolution and dissociation constant fitting were carried out using UniDec software ([106]). The affinity between HspB5 aCD and its Ctp that was reported before, 70 μM, was in good agreement with our measurement of 67 $\pm 4 \mu M$, providing evidence that excluding dimers has only a small effect on the results.

4.5.3 Mass photometry

Microscope coverslips (No. 1.5H thickness, 24 x 50 mm, VWR (630-2603)) were cleaned by two 5-minute sonication cycles, first in pure milli-Q water, then in a 1:1 mixture of Milli-Q water and iso-propanol. The coverslips were dried with nitrogen before being stored in a dust-free environment before use.

Immediately before use, silicone gaskets (CultureWell™ reusable gasket, 3mm diameter x 1 mm depth, Grace Bio-Labs) were attached to the microscope coverslips and the quality of the cleaning process was checked by recording a blank data acquisition. After making sure the gaskets were clean, the laser focus was found using 30 μL of the appropriate buffer, 200 mM ammonium acetate in this case, after which 25 μL of the buffer was removed from the well and replaced with 25 μL of the sample solution. This resulted in a lag time of approximately 10 seconds between sample dilution and the start of MP measurement.

Samples were equilibrated at the same ratios and concentration as in the mass spectrometry experiments (1 hour at 37°C). Samples were diluted to 400 nM monomer concentration immediately before measurement. The experimental MP setup was the same as described previously ([112]).

4.5.4 Markov chain Monte Carlo sampling

To obtain the confidence intervals for the fit, we ran a Markov chain Monte Carlo simulation. The simulation was set up with 18 walkers and 10 000 steps each. The first 4 auto-correlation times (τ) were discarded (roughly 1200 steps each), and the samples were further thinned $\tau/2$-times (roughly 150 times) to reduce any remaining correlation. The run was set up using the emcee Python module ([116]) with default settings. The likelihood function was defined as:

$$log(p) = -\frac{1}{2}abs(fit - data) \tag{4.3}$$

Further, the mixing ratios for all three distributions were independently optimised for each step with least-squares fitting.

The priors we set on the rates were as follows:

1. on-rates (k^{+A} and k^{+B}) were set to be between 0 and 500
2. single off-rates (k^{-A} and k^{-B}) were set to be between 0 and 1000
3. dimer off-rates (k^{-A}_{AA}, k^{-B}_{BB}, k^{-A}_{AB} and k^{-B}_{AB}) were set to be between 0 and 10

Since both of the on-rates were quite close to each other for most of the simulations, it might be reasonable to reduce the number of on-rates to just one common one instead of using two. This is supported by the fact that the on-rates might be diffusion limited (NMR data and [34]). The the (single) off-rates have to be different in order to model the diferent even-odd preferences for the homo-oligomers on their own. It would then be possible to keep keep the the homo-dimer-off-rates equal to 1. Since then the two off-rates are inherently linked as show in Chapter 2, it is possible that, at the barest level, we would only need to fit 4 rates: 1 on-rate, 2 single off-rates, and one hetero-dimer off rate.

4.5.5 Bias extraction from MP mixing data

To obtain the MP average measured masses, we recorded three separate measurements for all data points (meaning all equilibration concentrations as well as all mixing ratios), fitted a Gaussian to each, and calculated the average measured mass by averaging the three measurements for each data point. We then performed separate least-squares fitting to the three equilibration concentrations mentioned

in the main text to find k_{AB}^{-A} from which energy biases can be calculated as explained in the text.

4.5.6 Molecular dynamics

Initial structure preparation

To investigate the interactions between the aCD and the Ctp using molecular dynamics, we ran multiple simulations of both homo- and hetero-dimers with the aCDs and Ctp of all three proteins–HspB1, HspB5 and their last common ancestor. The hetero-dimers consisted of two complementary aCD-Ctp-exchanged complexes, ensuring the presence of every heterogeneous interaction.

The crystal structures of HspB1 and HspB5 homo-dimers were have been published before. The other dimer combinations, including ancestor homo-and hetero-dimers were prepared with one or more template(s) designed on the basis of the previous dimers. According to the ancestor sequence, the template(s) would have two symmetric monomers, each containing a 72-residue ACD chain and a 9-residue peptide. Sequence alignment by BLAST helped to remove the non-homologous region in HspB1 and HspB5 aCDs, which both had longer determined sequences than ancestor aCD did. A glutamine residue was added to both sides of the 7-residue peptide in HspB1 homo-dimer crystal structure (PDB:4MJH), and for HspB5, several crystal structures were integrated: The aCD structure was based on PDB:2WJ7, while the two peptides from 4M5T and 4M5S. The reorganised HspB1 and B5 homo-dimer structures were both minimised for 10ps (5000 conjugate gradient steps, 2 fs per step), and then merged to a template by MODELLER. For each dimer, including HspB1 and B5 homo-dimers, 5 structures were generated based on the template, of which the one with highest MODELLER score was selected for MD simulations.

MD simulations

Each .pdb file was first prepared with AmberTools21. The protein was modelled using the FF14SB force field and immersed in a 10-angstrom box of 0.15 M NaCl water solution (using the TIP3P water model). The MD simulations were run by NAMD2.14b2 engine using 2-fs time step, whilst the distance between any hydrogen and its mother atom was fixed, in order to constrain heavy atom distances (SHAKE algorithm). The systems were first minimised for 10 ps, then simulated in the nPT ensemble (300 K, 1 atm) for 0.5 ns with alpha carbons fixed by a 10 kcal/mol constrain. The constant temperature and pressure was maintained by Langevin dynamics, of which the piston period was 200 ns and the piston decay was 50 ns. The systems with modified volume were equilibrated in the nVT ensemble for 1 ns and the equilibrated system were suitable for the production simulations in the nPT ensemble at 300 K and the constant pressure value from the nVT simulation for 300 ns. The trajectory of each system was contracted, retaining one set of coordinates per ns.

4.6 Acknowledgements

I would like to thank Miranda Collier for conducting the aCD-aCD and peptide-binding experiments and Guodong Cao for setting up and analysing the MD simulation, as well as describing the setup of these simulations in the Methods section. Georg Hochberg used phylogenetic analysis to obtain the sequence for the ancestral aCD and ancestral peptide we used for the MD simulations.

5

Evolution of polydispersity and co-assembly in HspB vertebrate homologs

Contents

5.1 Motivation

In the previous chapters, we have shown that human HspB1 and HspB5 co-assemble together without almost any energy preference for homo- or hetero-dimerisation even after hundreds of millions of years after their gene duplication event. This result was very surprising in the context that protein paralogs most often prefer to self-assemble which allows them to evolve new functions ([82]). To see whether these two proteins co-evolve together to ensure this unbiased co-assembly, or whether they co-assemble in this way by chance, we decided to study HspB1-HspB5 co-assembly between different species. The reasoning behind this strategy is that if HspB1-HspB5 co-evolve together, mixing HspB1 and HspB5 from different species would result in a different mode co-assembly–more specifically, we wouldn't expect the two proteins from different species to co-assemble.

We also wanted to see how broadly applicable the model is. For this, we decided to explore the co-assembly between HspB1 and HspB6 as evidence suggests that these two proteins preferentially form hetero-dimers and so possibly co-assemble in a strict n:n fashion, such as shown in the bottom right corner of Fig. 2.1 on page 35.

While we were setting up these experiments, we mentioned that some HspB1 and HspB6 orthologs from the vertebrate clade have different oligomeric properties than their human counterparts–some markedly so. We wanted to explore this behaviour in more detail first, before we moved further with the mixing experiments. This is why the first half of this chapter focuses on the changes of polydispersity in different vertebrate HspB homologs before we then finally move to exploring the co-assembly between them.

5.2 Introduction

As mentioned in the previous chapters, the formation of higher sHsp oligomers might have a functional basis. The question of what role (if any) polydispersity plays in the function is not well-understood. From the previous chapters, we know that human HspB1 and HspB5, as well as their co-assemblies, are highly polydisperse. Contrary to this, a great number of bacterial and plant small heat shock proteins form higher, but monodisperse, oligomers. For example, *Methanococcus jannaschii* Hsp16.5 and *Sulfolobus tokodaii* Hsp14.0 form 24-mers only ([117], [118]), wheat Hsp16.9 forms only 12-mers ([119]), and *Schizosaccharomyces pombe* Hsp16 forms only 16-mers ([120]). We also know that upon temperature-activation and sHsp-substrate formation, HSP18.1 from *Pisum sativum* exhibits a degree of polydispersity that allows it to take up multiple substrate subunits ([32]). Further, some sHsps only form smaller oligomers–for example, HSP18.5 from *Arabidopsis thaliana* only forms monomers and dimers, human HspB6 only forms dimers ([121]), and conflicting evidence suggests that HspB7 is unable to oligomerise at all ([122]), although is has recently have been observed to possibly form dimer and higher oligomers ([123]). All these sHsps are still chaperone-active. In this chapter, we extend our understanding of the role of polydispersity on function by studying various HspB1, HspB5 and HspB6 orthologs, and find that both HspB1 and HspB6 orthologs have lost polydispersity in multiple vertebrate branches–seemingly without a loss of function. Finally, we hypothesise about what role neutral evolution could play in shaping polydispersity in a context when polydispersity is no longer strictly required.

While a more throughout introduction to the evolution and function of vertebrate sHsps can be found in Chapter 1, let us just reiterate the most important points here, too.

5.2.1 Phylogenetics of vertebrate HspB homologs

Franck et al. ([43]) surveyed the evolutionary diversity of vertebrate sHsps and found two additional HspB homologs in birds, amphibians and that they named HspB11 and HspB12, and further four (HspB12-HspB15) in teleost fish. Most of these paralogs originated during a period of extensive gene duplications early in vertebrate evolution.

Focusing in more detail on the different vertebrate lineages, they found that two further surveyed mammalian genomes (mouse and rat) only have the same 10 HspB homologs as humans. Meanwhile, *G.gallus* (red junglefowl) as a representative of the bird-reptile clade shares seven HspB mammalian orthologs (HspB1-5, HspB7 and HspB), but has three additional HspB homologs not found in mammals–HspB11a, HspB11b and HspB12. No other paralogs were found in other sequenced avian and reptilian genomes. At the time of the publication, amphibian genomes were not fully sequenced, but the authors still found five orthologs of human HspBs (HspB1-2, HspB4, HspB7-8) and multiple genes for HspB11-like homologs in *X. laevis* (African clawed frog). They further found a HspB5-like homolog in *R. catesbeiana* (American bullfrog). As for the teleost fish, a combined 12 HspB homologs were found in *T. rubripes* (the Japanese puffer) and *D. rerio* (zebrafish). Seven of those were paralogous to the mammalian HspBs.

5.2.2 The effect of the quaternary structure on the function of sHSPs

Both HspB4 and HspB5 are highly concentrated in the eye lens, where the total protein concentration can vary anywhere from about 400 mg/mL in humans ([24]) to well over a gram per milliliter in blue eyed trevally fish ([124]), where the high protein concentration is responsible for the high refraction index necessary for the

refractive power of the lens. Here, both the unusual structural properties and their chaperoning function play a functional role. The highly polydisperse, variable quaternary structure of these proteins presumably prevents their aggregation even at the high protein concentrations mentioned above. Indeed, this is one of the reasons that so far, it has not been possible to obtain a full-length crystal structure of a polydisperse sHsp at atomic resolution. However, even truncated HspB4 and HspB5 crystal structures allude to the multiple structural mechanisms that are important for the quaternary structure polydispersity. Namely, the palindromic sequence in the C-terminus; the flexible hinge loop preceding it; and the multiple possible registers ($AP_I - AP_{III}$) for the dimer interface ([26]) Additionally, HspB4 and HspB5 function as molecular chaperones, further prevent any protein aggregation, often seen with ageing as the eye lens proteins accumulate damage caused by UV radiation, oxidation, post-translational modification, and truncations ([28]), that may lead to cataract formation.

However, we know of at least one human sHsp, HspB1, that is polydisperse even though its prevalence in the eye lens is negligible ([125]), which prompts the question: How does the quaternary structure of these proteins affect their chaperone-like function?

Even though some sHsps, like human HspB6 and HspB7, do not form any higher stoichiometries in an appreciable extent, a wide range of sHsps, such as human HspB1 and HspB5, have the ability to form higher oligomers ([121]).

However, up to date, no systematic study on the effect of these higher oligomers, and their potential polydispersity, on the chaperone-like function has been conducted, and there seems to be evidence for both sides of the argument.

Jacobs et al. ([31]) suggest that the optimal chaperone-substrate binding results in chaperones themselves self-assembling. Using lattice modelling, they found that

the chaperone-chaperone interactions do not affect the stability of a protein solution with aggregation-prone protein substrates aggregating through strong, directional interactions. They also show that under fine tuning to make sure the chaperone-substrate is both reversible, but at the same time still protects the proteosome, the larger oligomers are a direct consequence of this tuning.

All these studies, and many more, seem to suggest that the larger oligomers might be reservoirs for the smaller, active species; whether they originated as a consequence for fine tuning for chaperone-substrate reversibility, or whether they form oligomers to sequester themselves from the cellular environment to not unnecessarily promote protein aggregation. The ability to form a wide range of oligomeric sizes may also allow for a broader range of how many and what size of these substates they are able to sequester. However, no direct evolutionary proof of this behaviour is available at the moment. It thus might be the case that higher oligomer formation might have more than one direct functional relevance–ranging from not unnecessarily promoting aggregation while still staying chaperone active to being able to recruit a more diverse range of proteins and protein stoichiometries for more polydisperse systems.

5.2.3 Human HspB6

Human HspB6 is a small heat-shock protein originally found as a co-purification product of human HspB1. It is highly expressed in skeletal, cardiac and smooth muscle ([126]), and plays an important role in smooth muscle relaxation ([126]), regulation of platelet formation, and has a protective role in the heart. Its expression has also been inversely correlated with the grade of malignancy ([127]), suggesting a potential role as a cancer biomarker.

Unlike HspB1 and HspB5, HspB6 does not form higher oligomers, and is mostly present as a dimer in solution. This could be because human HspB6 lacks the

'IXI' motif that is possibly needed for the formation of higher oligomers. However, it is known to form higher complexes in the presence of trimethylamine N-oxide (TMAO) as a crowding agent. It was suggested that this might be due to 'IXI'-like tripeptide sequence in the NTD which might effectively substitute for the CTD's IXI motif. Despite this, HspB6 still appears to be chaperone active, with the activity rivaling that of HspB5 ([128], [129]).

5.2.4 The experimental mass distributions show that HspB1-HspB6 seems to be highly biased towards co-assembly

After describing the co-assembly between human HspB1 and HspB5, the next attractive target seemed to be the co-assembly between HspB1 and HspB6. Both HspB1 and HspB6 are highly expressed in muscle tissue ([130]), and *in vitro* studies have shown that the two proteins form highly polydisperse complexes with an equimolar ratio of each protein ([129]). Heirbaut et al. also observed that even though both HspB1 and HspB6 alone can exchange subunits in a stochastic fashion, the HspB1-HspB6 complex (at 1:1 mixing ratio) seems to be composed solely of hetero-dimers ([131]). They also proposed that the highly-disordered NTDs in the two proteins are responsible for this bias. The same researchers used a library of deletion constructs and found a highly conserved RLFDQXFG motif that is necessary for subunit exchange. They also found that a site about 20 residues downstream from this motif is responsible for the resulting oligomeric size, and finally that a region unique to HspB6 is the reason for the preferential hetero-dimerisation ([132]).

5.3 Results & discussion

As mentioned in the Motivation section of this chapter, we first explore the evolution of polydispersity in the different HspB homologs before we move onto studying

their co-assembly in more detail.

5.3.1 The vertebrate HspB1 and HspB6, but not HspB5, orthologs slowly lose quaternary structure polydispersity

Fish and cartilaginous fish HspB1 orthologs are monodisperse

While setting out to explore the co-assembly of HspB1, HspB5, and HspB6 between different vertebrate species, we made a surprising discovery–some HspB1 and HspB6 orthologs have lost polydispersity during their evolution, while a host of others might be in the process of reducing their polydispersity.

One striking example of a loss of polydispersity in the HspB1 lineage is the elephant shark (*Callorhinchus milii*) HspB1 ortholog. As shown in Fig. 5.1 a, the protein only forms octamers to an appreciable extent. To prove that the octamers indeed were 'native' and not formed by, for example, misfolding during the purification process, we imaged the octamers using electron microscopy (Fig. 5.1 c). The sample seemed homogenous. We next tried reconstructing the 3D structure of this protein, but the imaging was too low in resolution. Even despite this, there seems to be a four-fold symmetry on the 'top view'. CD melting experiments (Fig. 5.1 d) also suggest a folded structure with a perfect match to secondary structure prediction from Phyre2 ([133]) with a standard melting behaviour. This also allowed us to fit a melting curve (using models II and III from [79], explained in more detail in the Methods section)–the measured melting temperature under mild denaturing conditions using 300 mM guanidine hydrochloride (GHCl) was 63.90 (95 % c.i.: 69.25, 64.44) °C. The octamers also seem to be chaperone-active (Fig. 5.1 b), with a (comparatively to the human HspB1 ortholog) higher chaperone activity in a temperature-induced citrate synthase aggregation assay, and similar activity in a DDT-induced aggregation assay.

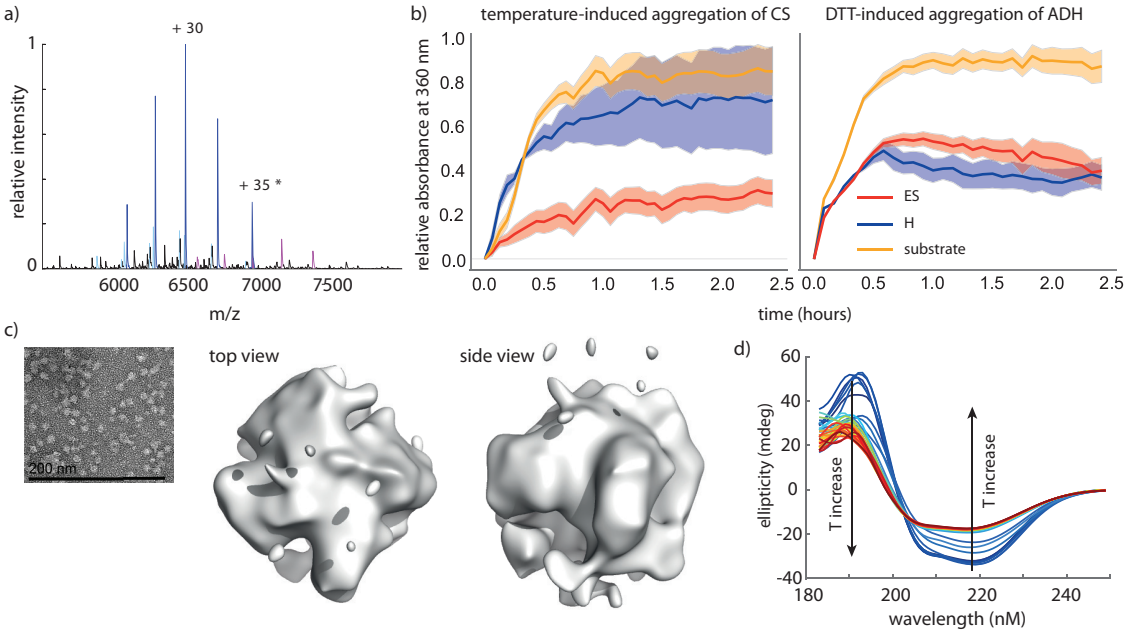


Figure 5.1: Biophysical characterisation of the elephant shark HspB1 ortholog. a) The native mass spectrum shows that the elephant shark HspB1 is predominantly an octamer (blue) with a very minor contribution from a decamer (purple). b) The protein also seems to be chaperone active in both temperature-induced and DTT-induced aggregation assays. Human ortholog is used as a positive control. c) Negative stain EM imaging shows homogenous oligomers. Albeit low resolution, the reconstructed 3D structure shows a degree of symmetry. The low-resolution reconstruction was provided d) The CD spectrum of the elephant shark protein shows folding in agreement with theoretical prediction. Melting behaviour is also as expected with a gradual decrease of the CD signal with increasing temperature from 20 °C (blue) to 85 °C (red).

In the HspB1 lineage, it seems other cartilaginous fish orthologs–little skate (*Leucoraja erinacea*) HspB1–as well as bony fish orthologs–spotted gar (*Lepisosteus oculatus*)–lost polydispersity. The little skate ortholog again forms mostly octamers, with some decamers (10 subunits) and dodecamers (12 subunits) also present (Fig. 5.2 middle). The spotted gar ortholog only forms dodecamer (12 subunits) to an appreciable extent. (Fig. 5.2 right).

Other vertebrate groups HspB1 orthologs have reduced their polydispersity compared to their last common ancestor

The rest of the HspB1 orthologs we studied here seem to have retained a polydisperse structure with a variable composition (Table 5.1 on page 141).

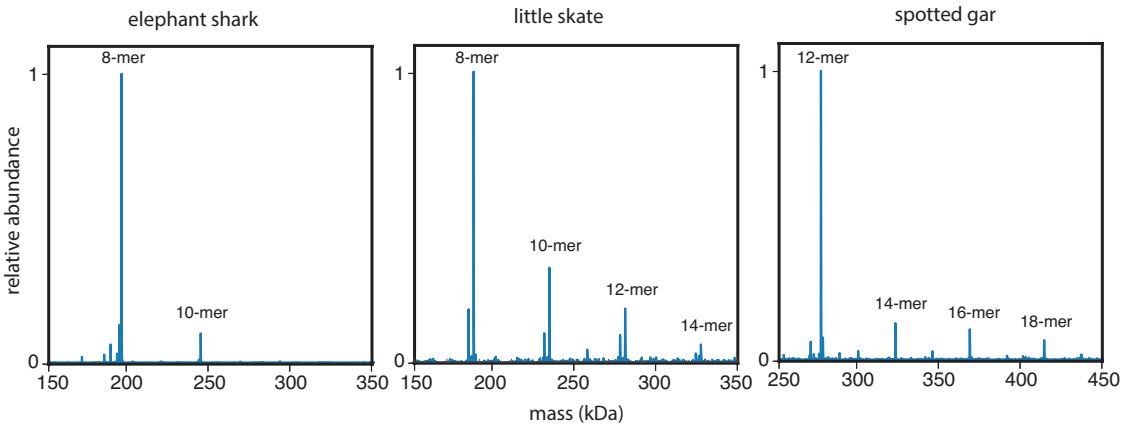


Figure 5.2: Bony and cartilaginous fish HspB1 orthologs show drastically reduced dispersity. The elephant shark HspB1 only forms octa- and deca-mers (left). The little skate HspB1 (middle) is comparatively slightly more disperse, but still shows reduced dispersity compared to the human HspB1. The spotted gar HspB1 only forms dodecamer to an appreciable extent (right).

To again make sure the proteins were folded correctly, we measured their ellipticities on CD (selection shown in Fig. 5.3). We also attempted to estimate the relative stability of selected monodisperse and polydisperse HspB1s by performing a melting experiment followed by CD. There were no significant changes in ellipticity at 220 nm (corresponding to the α-helix fingerprint, although there is an expected contribution at this wavelength from the antiparallel β-sheets in the aCD) with increasing temperature for both polydisperse proteins (African clawed frog and human), as well as the spotted gar HspB1. We then destabilised the protein with a range of GHCl concentrations and found the optimum concentration of GHCl to be 300 mM. At this concentration, we were able to observe secondary structure transitions for all five proteins. The two completely monodisperse proteins, elephant shark and spotted gar HspB1s, exhibited a two-state transition between a folded and an unfolded state. We were able to extract the melting temperature (and corresponding confidence intervals) of both using methods described in [79]: 63.90 (95 % c.i.: 69.25, 64.44) °C for the elephant shark, and 77.23 (95 % c.i.: 76.21, 78.26) °C for the spotted gar. Both polydisperse HspB1s as well as the little

skate protein, which exist in 3 oligomeric states, exhibited transitions that were not simply two-state. The unfolding seemed irreversible for all five proteins–the CD spectra recorded after the samples were cooled back down to 20 °C, and allowed to equilibrate at this temperature for an hour, did not show a significant difference to those recorded at 85 °C.

The next question was whether the reduction of polydispersity affects the proteins' chaperoning function. All the mono- as well as poly-disperse proteins also seem to be chaperone-active under two assays we chose: temperature-induced aggregation of citrate synthase (CS) and DDT-induced aggregation of alcohol dehydrogenase (ADH) (Fig. 5.4). These were chosen to be somewhat orthogonal to each other so that we could study the chaperone activity under different stressors. There might be a possible correlation of chaperone activity with the habitat temperature of the organism the protein comes from, but this should be confirmed with more proteins with a bigger temperature range. We observed no such trend in the DTT-induced assays.

species	average stoichiometry	variance
LCA1	24	6.223 (5.620, 6.827)
LCA2	19	10.14 (8.615, 11.660)
Human	18	8.449 (7.322, 9.576)
Frog	18	5.074 (4.375, 5.770)
Bird	15	4.562 (3.673, 5.454)
Gar	12	mono
Skate	8	8- to 12-mers only
Shark	8	mono

Table 5.1: Fitted average stoichiometries and quaternary structure variances of selected HspB1 orthologs. Both obtained by fitting a Gaussian to even stoichiometries as shown in Fig. 5.5. Starting from the last common ancestor, LCA1, all extant HspB1s reduced their average size, and most also reduced the variances of their higher oligomeric distribution.

To prove that the original state was indeed polydispersity, as opposed to the original state being a monodisperse protein and the terrestrial vertebrate then

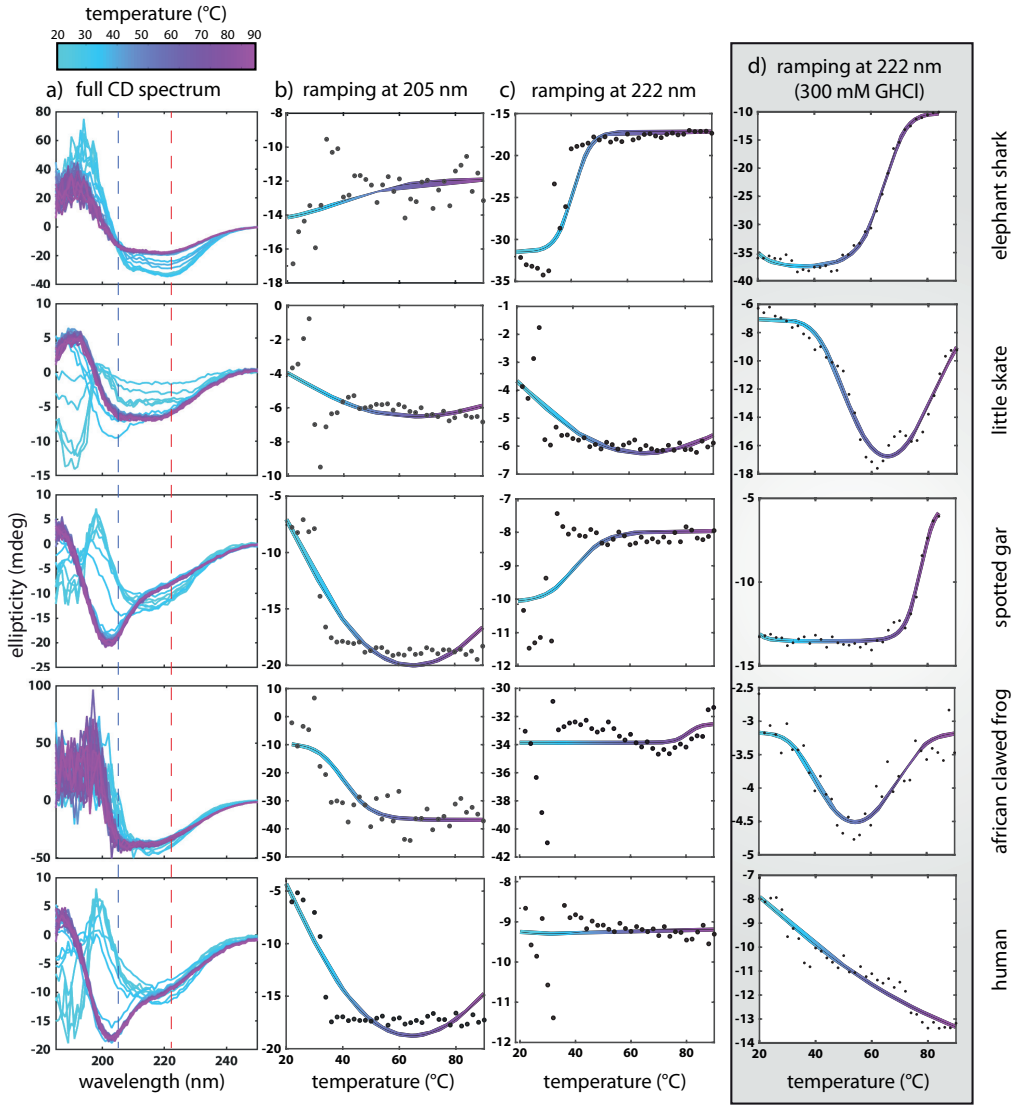


Figure 5.3: Far UV CD spectra and corresponding melting curves. a) Far-UV CD spectra for elephant shark, little skate, spotted gar, African clawed frog, and human HspB1s collected at monomer concentrations between 10 and 15 μM at varying temperatures from 20 °C (cyan) to 85 °C (magenta). Dashed lines correspond to the approximate maxima of contributions from α-helices (red, 222 nm) and random coils (blue, 205 nm). b) Ellipticity change at 205 nm as a function of temperature for each vertebrate ortholog. c) Ellipticity change at 222 nm as a function of temperature for each vertebrate ortholog. d) Ellipticity change as a function of temperature under 300 mM GHCl. Figure adapted from [27]

gaining polydispersity, we resurrected two last common ancestors (LCA-1 and LCA-2) of these proteins. LCA-1 is the last common ancestor of bony fish and tetrapods, and LCA-2 is the last common ancestor of coelacanth and tetrapods.

We then quantified all the HspB1 orthologs by extracting the main oligomeric

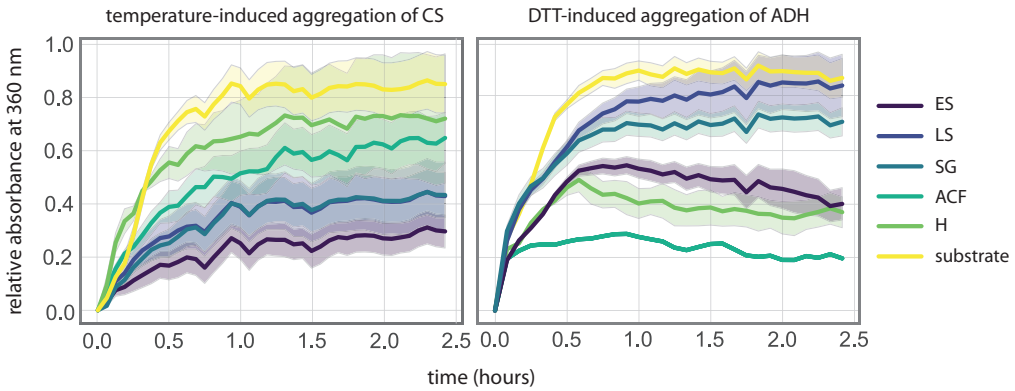


Figure 5.4: Activity assays show monodisperse HspB1s are chaperone-active. Both monodisperse and polydisperse HspB1 orthologs show chaperone activity in temperature-induced aggregation of CS (left) and DTT-induced aggregation of ADH. ES = elephant shark; LS = little skate; SG = spotted gar; ACF = african clawed frog; H = human.

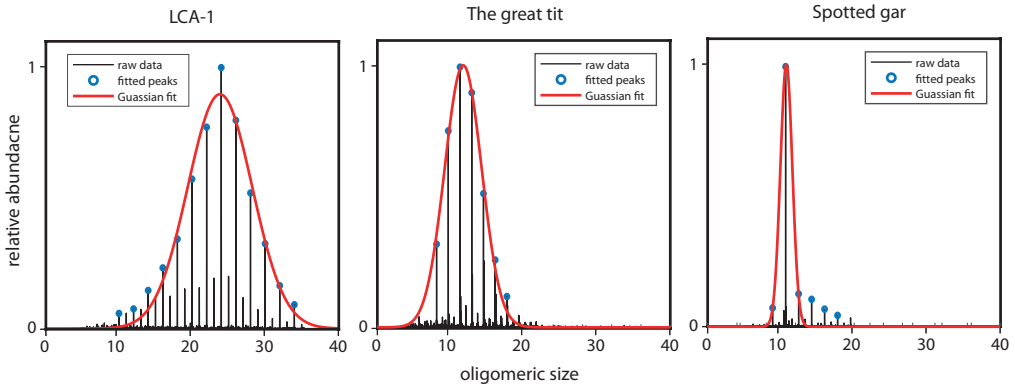


Figure 5.5: Gaussian curves were fitted to the even stoichiometries of all HspB1s, and their average sizes as well as quaternary structure variances were extracted. Here shown on the last common ancestor 1, LCA-1, as well as the extant great tit and spotted gar HspB1 orthologs.

mass and higher-oligomer variance based on fitting the even stoichiometries. Fig. 5.5 shows the fitting process on three examples; the result for all proteins studies are in Fig. 5.6). Both of the LCA proteins have a higher average stoichiometry than all the other vertebrate HspB1 orthologs, and LCA-2 also has the highest higher-oligomer variance, which we used as a measure of polydispersity.

In Fig. 5.6, the x-axis shows the average number of substitutions per site from the last common ancestor (LCA) of all the vertebrate HspB1 orthologs. In this context, this serves as a proxy for how much the proteins have evolved since the

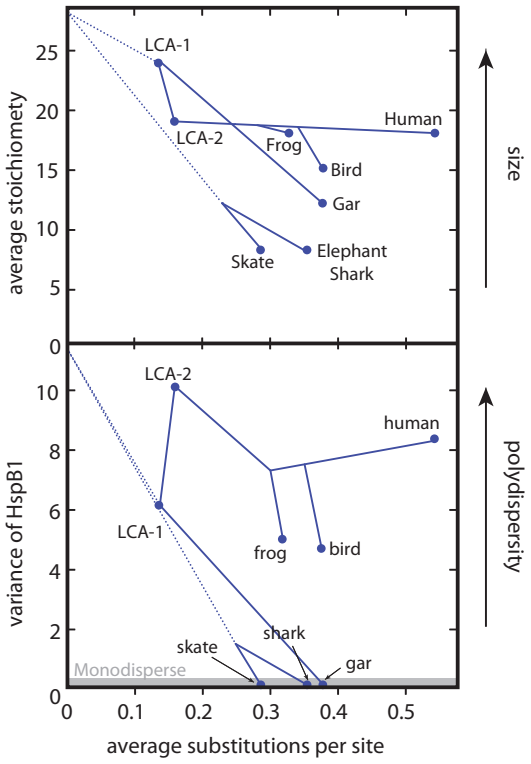


Figure 5.6: The average stoichiometry and variance as a function of average substitutions per site. Here, the average substitutions per site is a measure of how much the proteins evolved from the presumed last common ancestor of all of them. There seems to be a negative correlation between both the average stoichiometry and the variance, and the average substitutions. This might suggest that throughout evolution, the HspB1 orthologs slowly lose their dispersity. For simplification, we treat the little skate HspB1 as an octamer. Figure was kindly made by Prof. Justin Benesch.

LCA. We can then see that there seems to be a correlation between how much the various vertebrate HspB1 orthologs have evolved and both their average size and variance of their quaternary structure. As a general trend, there seems to be a slow and erratic decline of polydispersity (both in terms of the average size as well as the variance of the oligomeric distribution) with the extent of how far the proteins have evolved since their LCA.

5.3.2 Mammalian HspB6 have lost higher-oligomeric structure

Next, we wanted to see whether we could identify other losses of dispersity in different HspB1 paralogs, which might help us understand what criteria, if any, are important for sHsps to keep a high level of polydispersity. After reconstructing a phylogenetic tree (Fig. 5.10) for HspB1 homologs (including HspB1, HspB1-like proteins, HspB4, HspB5, and HspB6), we mentioned that HspB6 orthologs only lost the 'IXI' motif in mammals, but the motif was still present in amphibians, reptiles and even marsupials

(Fig. 5.7). We, indeed, observe a quaternary structure polydispersity on MS with CID (Fig. 5.8). After deconvolution and correction for CID, we find an oligomeric distribution with the usual preference for even stoichiometries, with the oligomeric mass envelope centred on 18-mers with higher-oligomeric distribution that varied from roughly 12- to 30-mers. As far as we are aware, this is the first known HspB6 ortholog that is polydisperse. As mentioned in the introduction to this chapter, the human HspB6 ortholog only forms dimers under non-crowded conditions.

```
Human      MEIPVPVQPSWLRRAS-APLPGLSAPGRLFDQRFGEGLLEAELAAL--CPTTLAPYYLRA     57
Koala      MEIPVPIQPSWLRRPAPLAFPSLTTPTRLFDQRFGEGLLEAELAAL--CPQALAPYYLRA     58
Frog       --MDVTIHHPWMRRPPLS---PSFFPSRILGQRFGEGVLESDLFPAMPMPMTLSPYYYSS     55
Gecko      --MDITIHHPWIHRPFGF---SPFFPPRLFDQRFGEGLLVSDMFSGMN---TLSPYYSRV     52
             : : ::  *::*           * *::.******:* :::        :*:***

Human      PSVALPVA----QVPTDPGHFSVLLDVKHFSPEEIAVKVVGEHVEVHARHEERPDEHGFV     113
Koala      PSVALRADPGPGQVTLAKGEFSVLLDVKHFSPEEINVKVVGEHVEVHARHEERPDEHGFI     118
Frog       PSIPQPSEVGLSEVKLDKDQFSVLLDVKHFSPEELNVKVVGDSVEVHAKHEERPDEHGFI     115
Gecko      PHIP---DTGLSEVKLDKDKFSVHLDVKHFSPEELSVKVVGDHIKVHAKHEERPDEHGYI     109
           * :         :*    ..*** **********: *****: ::***:*********::

Human      AREFHRRYRLPPGVDPAAVTSALSPEGVLSIQAAPASAQAPPPAAAK-------------     160
Koala      SREFHRRYRLPEGVDPAAVTSALSPEGILSIQAPATTPGTQAAERKVPIAQGSVPEAK--     176
Frog       SREFHRRYKIPPTVNPGAISSALSAEGLLSIQAPVTASGKQ-EERSIPIARKDK------     168
Gecko      SREFHRRYMIPKGVDLGSVTSALSPDGVLSITAPVAQGALP-QEHTIPISRQDNPAVAGK     168
           :******* :*  *: .:::**** :*:*** *  :
```

Figure 5.7: Marsupial, amphibian, and reptile HSPB6 all have the IXI motif. As shown with the frog HspB6, we could expect reptile and marsupial HspB6 orthologs to have some level of polydispersity. The motif is completely lost in extant mammalian HspB6 orthologs.

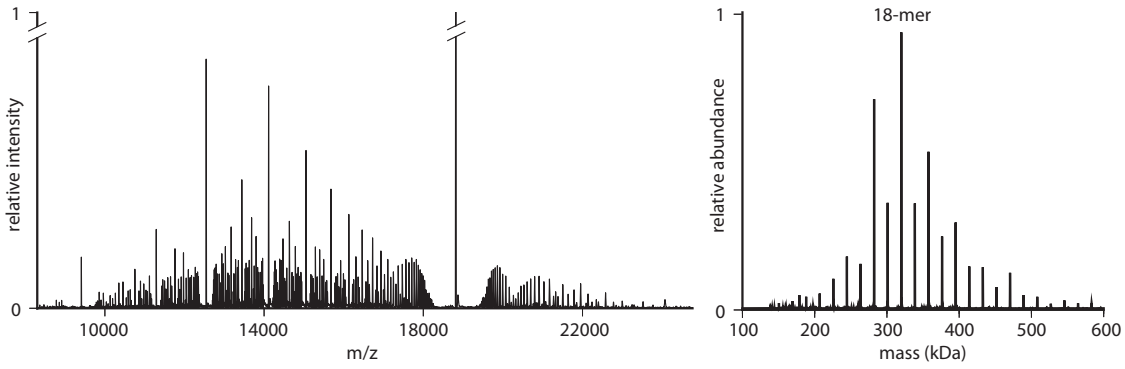


Figure 5.8: The CID spectrum (left), and corresponding CID-corrected mass distribution (right) of the frog HspB6 ortholog. The main oligomeric size seems to be an 18-mer with the whole higher-oligomer distribution spanning sizes from roughly 12-mers to 28-mers. There also seems to be an odd:even bias with preference for even oligomeric size.

Both human and fish HspB5 orthologs are polydisperse

Next, we wanted to explore the polydispersity of the HspB5 orthologs. We tried expressing numerous HspB5 orthologs, but the expression levels for most of them were too low; this might be because our expression system that uses a SUMO-His-tag may not be suitable for the HspB5 orthologs.

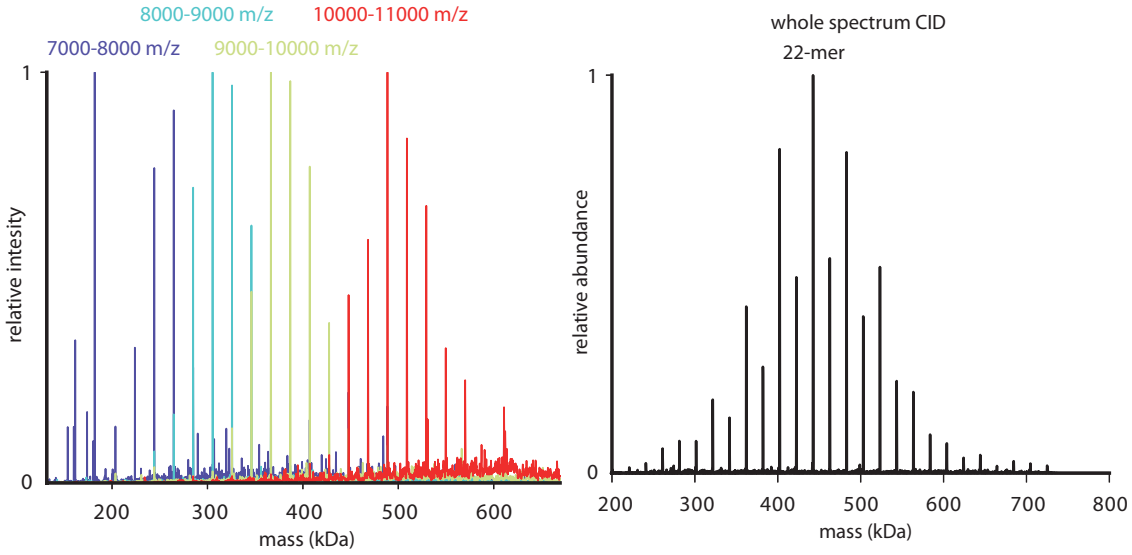
Despite this, we were able to obtain at least partial MS/MS spectra for the elephant shark HspB5 (Fig. 5.9). We were, however, not able to apply the intensity correction using a native spectrum because we were unable to obtain a good-quality native spectrum for this purpose.

Still, the CID-corrected mass spectrum show a level of polydispersity with oligomeric sizes ranging from 10-mers up to roughly 30-mers (Fig. 5.9 a)–similar to human HspB5 (Fig. 5.9 b). Interestingly, there seems to be a quite small, if any, bias towards forming either odd or even oligomers.

The fact that HspB5 stayed polydisperse in both humans and elephant shark (despite losses of polydispersity of HspB6 and HspB1, respectively), suggests that there might be a functional role for polydispersity in HspB5, as proposed by Laganowsky et al. ([26]) in addition to also acting as a molecular chaperone within the eye lens.

The phylogeny of HspB1, HspB4, HspB5 and HspB6 orthologs

As mentioned previously, we reconstructed the whole phylogenetic tree for the HspB homologs up the LCA of all these proteins spanning a range of animal classes from hemichordates to tetrapods (Fig. 5.10). First, there was a gene duplication event roughly between when cephalochordates and tunicates split off from the evolutionary line leading to extant tetrapods that resulted in the extant HspB1 and HspB5 orthologs. Interestingly, after this tunicates and hagfish both lost the HspB5



(a) Elephant shark HspB5 **(b)** Human HspB5

Figure 5.9: Oligomeric distributions for the elephant shark and human HspB5 orthologs. We were not able to obtain the full oligomeric distribution for the elephant shark protein, but by careful MS/MS measurements, we are able to show that the protein is still polydisperse with the range of the measured oligomeric sizes spanning a range from 10- up to 32-mers. There is also surprisingly low odd:even bias compared to the human HspB5 (b), meaning that we can expect the full oligomeric distrubtion of the elephant shark HspB5 to be roughly Poison-like in nature.

orthologs, whereas lamprey kept their HspB5 orthologs. It should be noted that both tunicates and hagfish do not have the eye lens, whereas lampreys do have primitive eyes–this suggests that vision in this group of animals probably evolved somewhere between where the tunicate and the hagfish-lamprey branches separated ([134]). Next, there was a gene duplication event resulting in an HspB1-like ortholog that is still present in extant fish and cartilaginous fish, although extant tetrapods have lost this ortholog. In the HspB5 branch, HspB6 orthologs separated first sometime before cartilaginous fish branched out–although both cartilaginous fish and fish then lost this ortholog. Finally, the HspB4 and HspB5 orthologs separated afterwards.

Polydispersity might not have a functional relevance outside the eye lens

As we have shown, there seem to be a slow loss of polydispersity in the HspB1 lineage and a complete loss of polydispersity in mammalian HspB6 orthologs without

a loss of chaperone-like function. On the other hand, HspB5 orthologs in both mammals and cartilaginous fish are still polydisperse, hinting at the possibility that HspB5 orthologs are polydisperse in the whole of vertebrates. This is interesting for multiple reasons. Firstly, this hints at the possibility that HspB5 orthologs have to stay polydisperse since it is functionally relevant for them. This could be because HspB5 is highly abundant in the eye lens ([24], [25]) where the protein concentration is so high that it is very likely that if HspB5 (and HspB4) was not polydisperse, it would crystallise out in the vitreous body of the lens.

In contrast, neither HspB1 nor HspB6 are expressed in the eye lens in a large extent, even though HspB1 is still one of the most expressed sHsp in other human tissues. This points to the possibility that polydispersity might not be relevant for the proteins' chaperone function.

5.3.3 The co-assembly between different HspB homologs

After exploring the evolution of polydispersity in different vertebrate species, we describe the co-assembly between different HspB homologs in the final section of this thesis. We first inspect the co-assembly between human HspB1 and HspB6 to confirm the broader applicability of the polydisperse mixing model based on chemical kinetics. We then move onto 'cross-species' mixing between HspB1, HspB5 and HspB6 from human and elephant shark. We picked the elephant shark for its evolutionary distance from humans. This can be seen for example in Fig. 5.10 b–sharks are the first vertebrate group to branch out from the path leading to tetrapods, which include mammals. As we could not express the elephant shark HspB5 in appreciable amount, as we discussed in the previous section, we could only obtain partial information on its co-assembly with other proteins. Nevertheless,

as we show, there is a reason to believe the HspB1 and HspB5 proteins don't co-assemble between the two species.

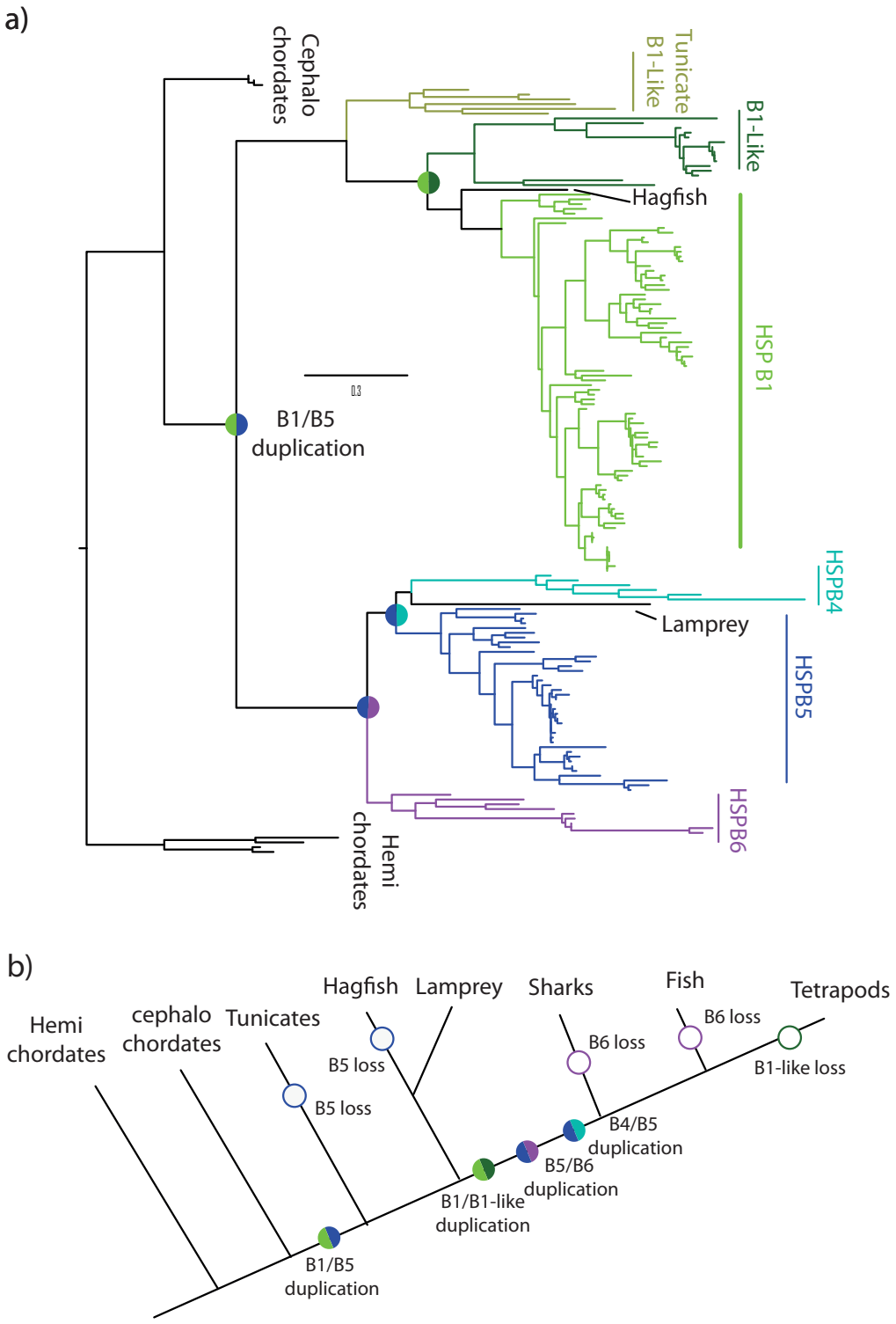


Figure 5.10: (a) Phylogenetic tree of HspB homologs and predicted gene loses and duplication events. (b) Phylogenetic tree of Eumetazoa again showing predicted gene loses and duplication events.

5.3.4 Human HspB1 and HspB6 form polydisperse, obligate n:n co-assemblies

To explore how human HspB1 and HspB6 co-assemble, we performed the same set of MS/MS and MP experiments as in Chapter 3–that is, tandem MS and MS intensity correction together with MP measurements followed by the averaging, frame and KDE pipeline to increase resolution–again for three mixing ratios. The results can be seen in Fig. 5.11.

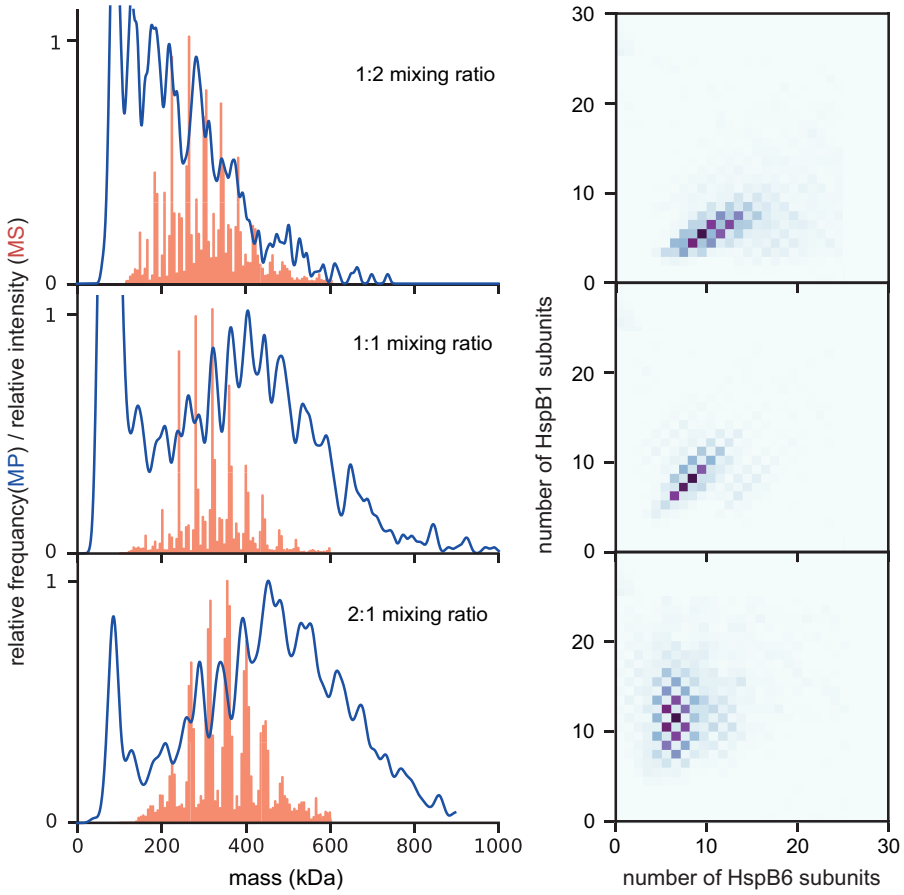


Figure 5.11: MS and MP spectra of the HspB1-HspB6 mixtures (left) and corresponding 2D histograms (right). There seem to be an increase in the average mass with a higher HspB1:HspB6 ratio. Mixing ratios in the figure are in the form of HspB1:HspB6.

Although we again aimed at roughly 1:3, 1:1, and 3:1 mixing ratios to make the distributions somewhat more comparable to the HspB-1HspB5 co-assemblies, by simple integration of the abundances of sub-stoichiometries of the mixtures

from MS/MS, the mixing ratios were closer to roughly 1:2, 1:1, and 2:1. This error is possibly due to the relatively small molar absorption coefficient of HspB6 (ϵ = 9970 $M^{-1}cm^{-1}$, compared to HspB1's ϵ of 40450 $M^{-1}cm^{-1}$) which makes concentration measurements more prone to error. Even so, the sub-stoichiometry distributions (Fig. 5.11 right) can tell us about the possible modes of co-assembly. In the case of the roughly 1:1 mixing ratio, the sub-stoichiometries indeed lie on the minor diagonal–suggestive of a preferential obligate n:n co-assembly and presumably hetero-dimerisation. This behaviour has been observed before ([131], [132]). The MS/MS envelope is, again, shifted towards lower masses compared to the MP envelope. Compared to the HspB1-HspB5 mixing experiments, the MP envelope is quite well-resolved, presumably due to the relatively low heterogeneity of this mixture with about 40 kDa separation between each n:n stoichiometry, and the relatively low abundance of odd stoichiometries.

In the case of the 1:2 (HspB1:HspB6) mixing ratio, there is a decrease in the size of the MP envelope, which roughly agrees with the MS/MS envelope. This shift is presumably caused by the inability of the HspB6 dimers alone without a crowding agent to form higher oligomers, both with itself and possibly with HspB1, too, to some extent. Interestingly, the sub-stoichiometry distribution is still 'diagonal' even though it does not lie on the minor diagonal. This might, again, be suggestive of a strong preference to form HspB1-HspB6 hetero-dimers–that is, in the case when HspB6 is in excess, all HspB1 monomers preferentially hetero-dimerise with HspB6. The remaining HspB6 subunits might still be able to join the higher oligomers to a small extent, as illustrated by both the fact that the co-assemblies lie slightly 'off' the minor diagonal in the HspB6-rich region of the 2D histogram as well as the most abundant oligomer is a 5:9-mer (HspB1:HspB6), presumably composed mostly of the $(B1B6)_5(B6B6)_2$ dimer arrangement.

Interestingly, the sub-stoichiometries formed with the 2:1 mixing ratio are not diagonal even though the HspB1-HspB6 hetero-dimers might still form preferentially. One of the possible explanations for this behaviour might be the relative propensity of the ($B1B1$) and ($B6B6$) homo-dimers to form higher oligomers–even though HspB6 is capable of forming higher oligomers under crowded conditions, it seems that in vitro without any crowding agents it is incapable of forming higher oligomers to an appreciable extent. This means that even though the (B6B6) homo-dimers might be able to be recruited into higher oligomers with HspB1, as happens with the 1:2 mixing ratio, further decrease in the HspB1:HspB6 ratio might cause a gradual decrease in the size of the n:n oligomers and an increase in the abundance of the HspB6-HspB6 dimers that do not co-assemble further.

In the case of increasing HspB1:HspB6 ratio, we see an HspB1-rich region parallel to the vertical axis with no 'diagonal-like' character. The size of the mass distribution does not seem to change much between the 2:1 and 1:1 mixing ratios (Fig. 5.11), and pure HspB1 homo-oligomers (Chapter 3, Fig. 3.12). This again might suggest an uneven contribution to the stability of the co-assemblies, similar as was the case with the HspB1-HspB5 co-assemblies.

5.3.5 The chemical-kinetical model is able to extract energy biases and shows a preference for hetero-dimerisation

We next performed an MCMC fitting as described in Chapter 4. Briefly: for the 1:2 mixing ratio, we fitted the whole measured MS/MS mass distribution as it agrees with MP mass envelope fully. We also penalised distributions that had 12 percent of their distribution above roughly 480 kDa (24-mers). For the 1:1 and 2:1 mixing ratios, we only fitted the MS/MS distributions in a region of increasing abundance.

We also set restraints on the abundances of the higher oligomers: 10 percent (32-mers) for the 1:1 mixing ratio, and 18 percent for the 2:1 mixing ratio. We also set $k^{+A}[A] = k^{+B}[B]$ for faster fitting and ran 12 MCMC chains for 10000 steps each.

The fitted hetero-dimer stabilisation energy was -26.20 ± 15.28 kJ/mol of interactions–much higher hetero-dimer stabilisation than the unbiased HspB1-HspB5 co-assemblies (Fig. 5.12 a), which was, either -1.08 ± 4.36 or -1.29 ± 4.39 depending on which normalisation method we use.

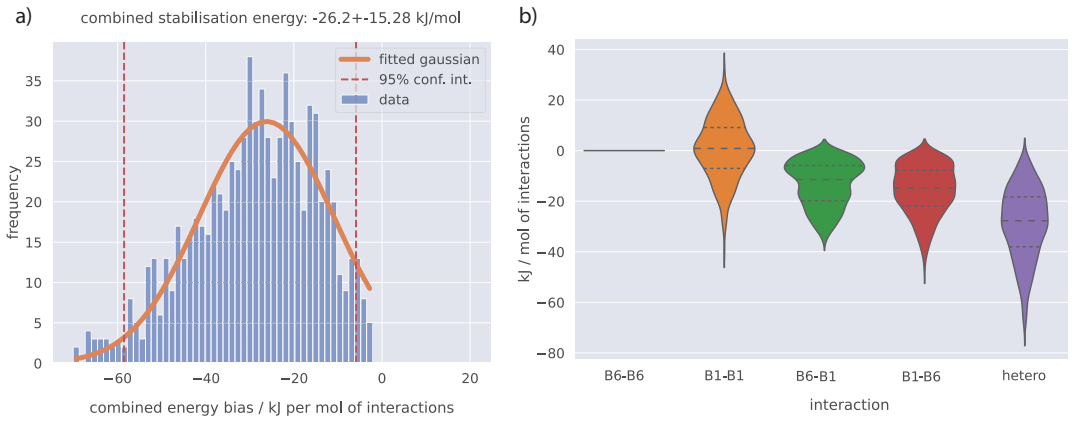


Figure 5.12: Energy biases extracted for the HspB1-HspB6 mixtures. Both hetero-dimer interactions are more favourable compared to the homo-dimer interactions. It is possible that the hetero-dimer interactions are 'directional' to an extent.

Similarly to the previous chapter, we further explored the separate stability of the different homo- and hetero-dimers (Fig. 5.12 b) relative to that of the B6-B6 homo-dimers and found the B1-B1 homo-dimer to be of equal stability, whereas both hetero-dimers (B1-B6 and B6-B1) are more stable than either of the homo-dimers. This fits with both our experimental data showing obligate, n:n co-assembly between human HspB1 and HspB6 as well as other published sources ([131], [132]). Similarly, there might, again, be some level of 'directionality' in the hetero-dimer interactions as the increasing HspB1:HspB6 ratio results in progressively larger oligomers.

5.3.6 *Callorhincus* HspB1 and human HspB5 homologs prefer to self-assemble, yet the HspB1 and HspB6 orthologs show preferential n:n co-assembly

HspB1 & HspB6

Finally, we briefly describe the cross-species co-assembly between elephant shark and human HspB1 and HspB5, and HspB1 and HspB6. These were chosen for their evolutionary distance; indeed, in vertebrates, cartilaginous fish are a clade most distant from mammals and the two clades are thought to have separated more than 460 million years ago ([135]).

Even though cartilaginous fish don't have HspB6 orthologs, the elephant shark HspB1 seems to co-assemble with human HspB6 in a similar, n:n fashion as its human HspB1 counterpart (Fig. 5.13 bottom left and right, respectively). This is interesting for two reasons: firstly, as mentioned, the evolutionary distance between the homologs was vast, and the protein clearly didn't need to co-evolve in elephant shark at least. This might be because there is a slight asymmetry in the interactions between HspB1 and HspB6, where changes to the HspB6 NTD sequence cause dramatic changes in co-assembly, whereas equivalent deletions in the HspB1 NTD either do not affect the co-assembly, or make it faster ([132]). This 'flexibility' might be one of the reasons why the elephant shark HspB1 still co-assembles with the human HspB6 in the same fashion as its human HspB1 counterpart.

HspB1 & HspB5

In the case of a cross-species co-assembly between the elephant shark and human HspB1 & HspB5, there seems to be a slight preference for self-assembly that is, however, hard to quantify. In the case of human HspB5 – elephant shark HspB1 co-assembly, there seem to be two regions: the first one that is parallel to the y-axis

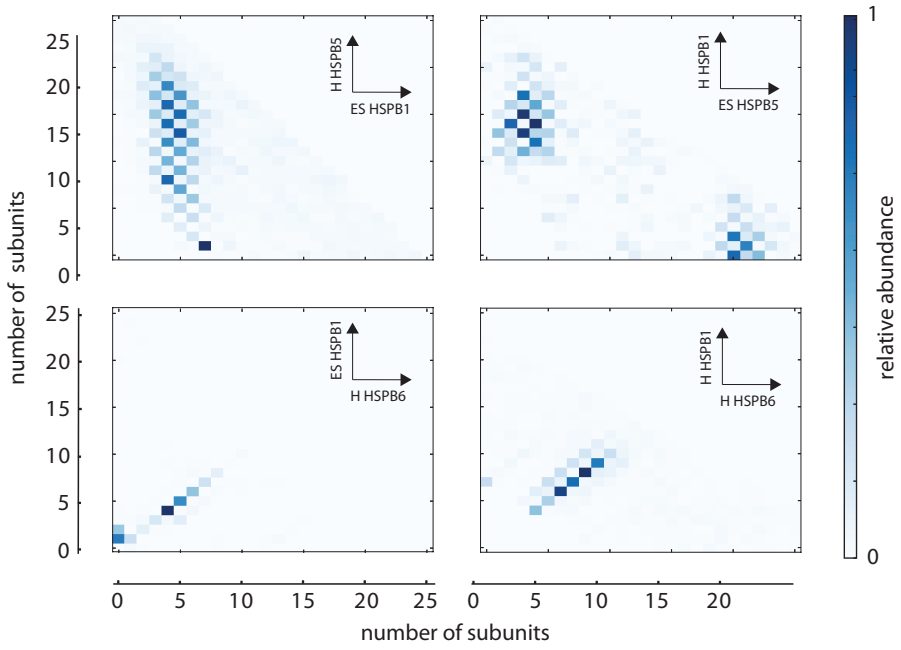


Figure 5.13: Co-assembly between HspB homologs from human and elephant shark. Whereas the elephant shark and human HspB1 co-assemble with the human HspB6 in a similar, n:n, fashion (bottom left and bottom right, respectively). The cross-species HspB1-HspB5 mixtures seem to prefer self-assembly. The fact that the cross-species HspB1-HspB5 mixtures prefer to self-assemble, contrasted to the unbiased co-assembly between the human paralogs might suggest that there is a potential functional relevance for the way the human paralogs co-assemble.

that is rich in human HspB5, and the other is the hetero-octamer that is composed of 7 elephant shark HspB1 subunits and 1 human HspB5 subunit.

For the opposite case of co-assembly between human HspB1 and elephant shark HspB5, we again see two regions; one along the y-axis that is rich in human HspB1, and one along the x-axis that is rich in elephant shark HspB5. Again, due to the low amounts and quality of the elephant shark HspB5 protein, we were only able to record one MS/MS spectrum in a narrow m/z region. We can expect that the whole distribution would be broader and would extend further up or right for the human-rich and elephant-shark-rich regions, respectively. Interestingly, it seems that for the co-assembly between human HspB5 and elephant shark HspB1, all co-assemblies have 7 or fewer elephant shark subunits.

Both of these distributions have two regions, one that is composed mostly of human HspB and one that is rich in elephant shark protein. Extracting quantitative energies for the case of mixing between human HspB5 and elephant shark is not possible yet since we don't have a good model for mixing between one polydisperse and one monodisperse protein that prefer to self-assemble due to the lack of experimental data for these systems.

5.4 Acknowledgements

I would like to thank Georg Hochberg for his phylogenetic analysis which made it possible to reconstruct the LCAs, as well as guiding the project and selecting appropriate protein sequences for the vertebrate species of interest. I would further like to thank Marco Hiscox and Christina Lee for their help expressing and purifying the proteins and Marco for conducting a subset of the CD melting experiments. I would also like to thank Wiktoria and Luke for their feedback on this chapter.

5.5 Methods

5.5.1 Protein expression and purification

HspB1 orthologs, acfrican clawed frog & human HspB6, and elephant shark HspB6

Similiarly to Chapter 3, the full-length HspB1 homologs (all shown HspB1s, elephant shark HspB5 and HspB6s) protein was encoded in the pETHSUL vector (kindly donated by the Weeks group) containing an N-terminal His-tag with a SUMO sequence for digestion. The vector was used to transform BL21(DE3) competent cells (Novagen). Single colony was chosen and grown in a starter culture (10 mL of LB) overnight at 20 °C. Next, 500 μL of the starter culture was then used to inoculate 1 L of the ZYP auto-induction medium. Cells were expressed overnight

at 24 ° with shaking at 180 rpm until OD reached values above 4, after which the temperature was lowered to 18 °C and the cells were left, again with shaking at 180 rpm, for 24 hours before they were harvested.

After harversting, the cells were resuspended in roughly 50 mL 250 mM NaCl, 50 mM phosphate, 12.5 mM imidazole, 5 mM BME buffer with two protease inhibitor cocktail tablets (Roche) added, and lysed by microfluidising. The lysed supernatant was centrifuged for 20 minutes at 20000xg (rfc), and the clear fraction was filtered before being loaded onto two 5-mL Nickel columns (GE Healthcare) connected in series and the protein was eluted using a gradient up to 500 mM imidazole. The resulting fractions were pooled and buffer exchanged overnight back into low-imidazole buffer with concurent SUMO cleavage using the Ulp1 hydrolase (expressed and purified in-house). The hydrolysis step was done at 4 °C out of prudency to ensure minimal truncation. Next day, the sample was again loaded onto the column, but this time, the flow-through was collected. The flow-through was then concentrated and loaded onto the Superdex 200/20 column. This served as a further purification step and well as a buffer exchange method. The protein was exchanged into a PBS buffer (pH 7.4) with 1 mM DTT. All proteins other than the HspB5 orthologs showed good expression levels.

Human HspB5

BL21(DE3) competent cells (Novagen) were transformed with a pET23a(+) vector (Novagen) which had an insert coding for the full-length HSpB5 protein. After DNA sequencing to ensure the transformation didn't produce any mutations, two one-litre flasks of Luria broth (LB) were inoculated with the transformed *E. Coli*. The cells were grown at 37 °C until the solution's optical density (OD) reached values between 0.6 and 0.8, after which expression was induced by adding IPTG

to the final concentration of 1 μM. The cells were then grown overnight at 20 °C, harvested, flash-frozen with liquid nitrogen and kept at -80 °C before purification.

The purification process was described previously ([110]). Briefly, the cells were lysed by microfluidising in phosphate-buffered saline (PBS) with 5 mM β-mercaptoethanol (BME) at pH 7.4 with EDTA-free protease inhibitor cocktail (Roche). The lysate was then spun down at 20000xg (RCF) for 20 minutes. The clear portion of the lysate was filtered through a 0.22 μm filter and the proteins were then purified by anion-exchange chromatography (IEX) followed by size-exclusion chromatography.

First, the clear lysate was loaded into two 5-mL HiTrap Q columns (GE Healthcare) connected in series and the protein was eluted using a salt gradient of 500-900 mM NaCl in the lysis buffer. After this, the appropriate collected fractions were concentrated and the concentrated protein solution was exchanged into the storage buffer (PBS at pH 7.4 with 1 mM DTT) by SEC using the Superdex 200/20 column. This also functioned as another, orthogonal purification step as only the purest fractions were then concentrated. The protein concentration was determined by absorbance at 280 nm using an extinction coefficient value of 9970 $M^{-1}cm^{-1}$ calculated from the protein sequence using the ProtParam website ([111]). The samples were then aliquoted and flash frozen with liquid nitrogen and stored at -80 °C.

5.5.2 Mass spectrometry

All HspB homologs were stored at -80 °C. On the day of measurement, the proteins were thawed and buffer exchanged into 200 mM ammonium acetate (pH 6.9) using BioRad biospin columns. For homo-oligomeric mesurements, the proteins were then

equilibrated at 37 °C for 60 minutes at 15 μM monomer concentration to ensure the recorded distribution in the eqilibrium distribution at that temperature.

Similarly, for the HspB1-HSpB6 hetero-oligomeric 1:2, 1:1 and 2:1 mixtures, the mixtures were equilibrated again at 37 °C for 60 minutes, but at 60 μM monomer concentration before they were diluted by a factor of three before the MS measurements.

The data was collected using in-house gold-plated capillaries on a Q Exactive^{TM} mass spectrometer in positive ion mode with source temperature of 170 °C, capillary voltage of 1.4 kV. and S-lens RF 200 %. In-source trapping was set to -200 V for hetero-oligomers, and HspB1 and HspB6 orthologs, and -180 V for the HspB5 orthologs. Ion transfer optics and voltage gradient throughout the instruments were optimized for ideal transmission as follows: injection flatapole 10V, inter-flatapole lens 8V, bent flatapole 6V, transfer multipole 4V, C-trap entrance lens 3V. The RF MP2 and MP3 voltages were overridden to remain constant at maximum value (800+V). For the measurements shown, we used nitrogen as the collison gas as a heavier gas, argon, did not seem to result in a significant improvement to the resulting spectra.

Spectra were acquired with 10 microscans to increase the signal-to-noise ratio with transient times of 64 ms, corresponding to resolution of 17,500 at m/z = 200 for a transient time of 64 ms, and AGC target of 1.0 x 106. The noise threshold parameter was set to 3 and the scan ranges used were and the spectra were recorded in the m/z region between 12000 and 30000 m/z. The tandem MS spectra were recorded in sequential selection windows of 500 m/z.

Mass deconvolution was achieved using the UniDec software ([106]). Before deconvolution, the data were not smoothed, subtracted, binned, or otherwise pre-processed. We used charge range of $10^+ - 50^+$. Further, the fitted peak FWHM

was optimized for each spectrum (generally within 0.5 and 1 Th). All other UniDec parameters were left at their default values. The deconvolution was performed in the mass range of 100 to 800 kDa. After deconvolution, n-1 and n-2 charge series were assigned (as shown in Fig. 3.4), and the spectrum was corrected for CID as shown in Chapter 3.

After deconvolution, the spectra were rescaled, noise-subtracted, and added together as described in the text. 2D binning was done using mass windows of 200 Da centered at the expected mass of the sub-stoichiometries using sequence monomer masses.

5.5.3 Circular dichroism

All CD measurements were done with proteins at 10-20 μM monomer concentration in a 20 mM phosphate buffer. The melting curves were recorded between 20 and 90 $^{\circ}C$ in 2 $^{\circ}C$ increments. The fitting was done using using models II and III from [136].

In model II, we assume a two-state transition with linear correction for pre- and post-transition changes in ellipticity with temperature. Model III describes a two-state transition between a folded and unfolded state assuming that there is a change in heat capacity of the folded and unfolded forms. We found that model III fits the data better and so the melting temperatures mentioned in the text are the fitted values to model III. The melting temperatures were defined as the temperature at which half of the protein present is unfolded.

6 Concluding remarks

6.1 How this work expands the understanding of sHsp co-assembly and their evolution

The broad objective of this work was to obtain a quantitative description of 'mixing', or co-assembly, in sHsps. As the known homologs of sHsps are both mono- and polydisperse, we develop two models of co-assembly. We briefly applied the model based on statistical mechanics to the co-assembly between mitochondrial Hsp26.5 and Hsp21 from the chloroplast and showed that there is a slight bias for self-assembly, compared to a completely unbiased mixing.

In the next chapter, we then applied two experimental techniques, mass photometry and mass spectrometry, to study highly heterogeneous, polydisperse co-assemblies between the two most expressed human sHsps–HspB1 and HspB5. Using a novel application of tandem MS together with the algorithms for CID and intensity correction, we are able to 'piece' the individual MS/MS measurements back together to obtain information about the two proteins' co-assembly at the sub-stoichiometry level of organisation.

Next, we showed how we can apply MP, with careful selection of frame averaging, filter values, and time and spatial selection to obtain mass envelopes with dimer-level of resolution in the whole mass envelope in the case of HspB1 homo-oligomers, and the higher-oligomer region of HspB5. We also show that HspB1 and HspB5 seem to interact with the glass surface in two distinct fashions which causes HspB5 to disassemble into smaller oligomeric species that we don't detect on MS, and hypothesise about the mechanism behind this behaviour. We were further able to obtain mass-unbiased mass envelopes for the HspB1-HspB5 mixtures, again with dimer-level of resolution–at least for the two HspB1-rich mixtures.

In Chapter 4, we asked ourselves how we can apply the chemical modelling to the HspB1-HspB5 mixtures to obtain 'complete' mass envelopes as well as to extract any energy biases. We find that, within the 95 percent confidence interval, there does not seem to be a statistically significant energy difference between homo- and hetero-interactions in the co-assemblies, although there might be a slight 'directionality' in the hetero-oligomer interactions. We also show that the fits seem to be a good match to the MP mass envelopes.

To explore the interaction 'directionality' further, we performed two different sets of experiments: binding experiments with MS and average-size measurements for different mixing ratios with MP. In the MS experiments, both for the aCD-aCD and aCD-peptide interactions, we found that although the aCD-aCD interaction is unbiased and the two aCDs dimerise in a purely statistical fashion, there is a slight energy preference for the interaction between the HspB1 Ctp and the HspB5 aCD, which might explain the modelling results. In the MP set of experiments, we showed that the average size of the co-assemblies does not vary linearly with the HspB1:HspB5 mixing ratio–the average size rises up to the point of the roughly 1:1 mixing ratio, after which it either stays constant, or even slightly decreases

possibly. From this, and the MS mass envelopes of homo-oligomers, we were also able to extract hetero-interaction energies close to the energies obtained by fitting the detailed MS/MS data. Further MD experiments also confirm the HspB1 Ctp - HspB5 aCD bias.

Finally, in Chapter 5, we explore the evolution of polydispersity in sHsp vertebrate homologs and find that both the HspB1 and HspB6 orthologs have lost polydispersity in at least some vertebrate lineages, and the highly polydisperse HspB1 orthologs might be slowly losing their polydispersity, both in terms of their average size, and the variance of their higher-oligomeric structure. We also note that both the elephant shark and human HspB5 have regained their polydisperse structure, which might hint at the functional role of polydispersity in the eye lens where HspB5 is highly expressed.

After studying the changes of polydispersity in the HspB1 and HspB6 orthologs, we decided to explore the co-assembly between the human paralogs. We found that they co-assemble in an n:n fashion meaning there is a high preference for hetero-dimerisation, and extracted the energy biases. We then also explored the co-assemblies between the elephant shark HspB1 and human HspB6, as well as the two most evolutionary distant HspB1 and HspB5 orthologs (shark and mammalian). We found that the HspB1-HspB6 still prefer to co-assemble in a roughly n:n fashion, whereas the HspB1-HspB5 cross-species co-assemblies tended to form two 'species-rich' regions–possibly suggestive of preferential self-assembly to a degree.

6.2 Future works and ideas that this work has made possible

6.2.1 The evolution and functional relevance of polydispersity

The surprising discovery of polydispersity reduction in HspB1 and HspB6, but not HspB5, vertebrate orthologs poses an interesting question: what exactly is the relevance of polydispersity in sHsps, and probably an even more fundamental one: If polydispersity is not functionally relevant anymore (meaning this high-entropy state is no longer needed to be maintained), what happens during the neutral evolution of these proteins? We currently know of proteins that changed their quarternary structure only on the order of just sub-units ([137], [138], [17]), so seeing the large changes of dispersity in the vertebrate sHsps homologs is quite interesting. We have estimated that, assuming ideal coverage, we only need about 10 to 12 other HspB1 and HspB5 orthologs to obtain a more fundamental understanding of these evolutionary processes.

6.2.2 The entropic contribution to the evolution of vertebrate sHsp co-assemblies

In Chapters 3, 4, and 5, we have shown that human HspB1 and HspB5, and to some extent human HspB1 and HspB6, form highly heterogeneous co-assemblies. Some such high-entropy states can be generated for example by gene duplication, where, initially, the protein paralogs will co-assemble without any preference for self- or co-assembly. From here, the two proteins can start favouring self-assembly, which often correlates with a gain of new functions for the two proteins ([82]) on one hand–such as we see with some of the elephant shark-human HspB1-HspB5 co-assemblies, or for the two plant proteins, Hsp21 and Hspb26.5. On the other

hand, the two proteins can start 'moving' towards co-assembling 'completely', n:n, such as we see with the HspB1-HspB6 co-assemblies.

Similarly, for the case of monodisperse co-assembly, such as we've seen in Chapter 2, we know from literature ([82]) about a pair of sHsps that 'completely' self-assemble–they don't form any co-assemblies that the authors could detect. From our own data, we now know of a pair of plant sHsps (Hsp21 and Hsp26.5) that partially co-assemble, although they are still somewhat biased towards self- rather than co-assembly, but we can expect that right after a gene duplication event, these sHsps would have co-assembled without any preference. We are yet to see any obligate, n:n, co-assembly for plants proteins, but with the great number of plans sHsps, it is feasible some pairs do co-assemble in this fashion.

Even though our understanding of sHsp co-assembly is still somewhat limited, here we have shown that we are able to extract energy biases for co-assembling proteins, both mono- and poly-disperse. We have further shown and analysed the co-assemblies between human HspB1-HspB5 and HspB1-HspB6 pairs that cover the assembly space between co-assembly without preference for self- or complete n:n co-assembly as well as the case of complete n:n co-assembly. In the final chapter, we have also shown an example of two protein pairs that prefer to self-assemble.

From this, we believe there is enough information in the evolution of the two pairs, HspB1-HspB5 and HspB1-HspB6, to cover the complete spectrum of polydisperse co-assembly. We can then extract associated energy differences, and study how high-entropy states affect neutral the evolution of these co-assemblies.

References

[1] Emmanuel Jaspard et al. “sHSPdb: A database for the analysis of small Heat Shock Proteins”. In: *BMC Plant Biology* 16.1 (2016), pp. 1–11.

[2] A. Pacheco et al. “Small heat-shock protein Hsp12 contributes to yeast tolerance to freezing stress”. In: *Microbiology* 155.6 (2009), pp. 2021–2028.

[3] Elizabeth R. Waters et al. “Plant small heat shock proteins – evolutionary and functional diversity”. In: *New Phytologist* 227.1 (2020), pp. 24–37.

[4] Evgeny V. Mymrikov et al. “The chaperone activity and substrate spectrum of human small heat shock proteins”. In: *Journal of Biological Chemistry* 292.2 (2017), pp. 672–684.

[5] John Den Engelsman et al. “The Small Heat-Shock Proteins HSPB2 and HSPB3 Form Well-defined Heterooligomers in a Unique 3 to 1 Subunit Ratio”. In: *Journal of Molecular Biology* 393.5 (2009), pp. 1022–1032.

[6] Evgeny V. Mymrikov et al. “Heterooligomeric complexes of human small heatshock proteins”. In: *Cell Stress and Chaperones* 17.2 (2012), pp. 157–169.

[7] Yuki Sugiyama et al. “Muscle develops a specific form of small heat shock protein complex composed of MKBP/HSPB2 and HSPB3 during myogenic differentiation”. In: *Journal of Biological Chemistry* 275.2 (2000), pp. 1095–1104.

[8] Andrew W. Senior et al. “Improved protein structure prediction using potentials from deep learning”. In: *Nature* 577.7792 (2020), pp. 706–710.

[9] John Jumper et al. “Highly accurate protein structure prediction with AlphaFold”. In: *Nature* 596.May (2021).

[10] M. Michael Gromiha et al. “Protein structural bioinformatics: An overview”. In: *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*. Vol. 1-3. Elsevier, Jan. 2018, pp. 445–459.

[11] J. Drenth. *Principles of Protein X-Ray Crystallography*. 2007.

[12] Ewen Callaway. “Revolutionary cryo-EM is taking over structural biology”. In: *Nature* 578.7794 (2020), p. 201.

[13] K. Wuthrich. “Protein structure determination in solution by NMR spectroscopy”. In: *Journal of Biological Chemistry* 265.36 (1990), pp. 22059–22062.

[14] Sofia K Forslund et al. “Evolution of Protein Domain Architectures”. In: *Evolutionary Genomics: Statistical and Computational Methods*. Ed. by Maria Anisimova. New York, NY: Springer New York, 2019, pp. 469–504.

[15] William Lucas. “Viral Capsids and Envelopes: Structure and Function”. In: *Encyclopedia of Life Sciences* (2010), pp. 1–7.

[16] Lena Danielli et al. “Quantifying the distribution of protein oligomerization degree reflects cellular information capacity”. In: *Scientific Reports* 10.1 (2020), pp. 1–10.

[17] Nicholas J. Fraser et al. “Evolution of Protein Quaternary Structure in Response to Selective Pressure for Increased Thermostability”. In: *Journal of Molecular Biology* 428.11 (2016), pp. 2359–2371.

[18] Emmanuel D. Levy et al. “3D complex: A structural classification of protein complexes”. In: *PLoS Computational Biology* 2.11 (2006), pp. 1395–1406.

[19] Carol V. Robinson et al. “The molecular sociology of the cell”. In: *Nature* 450.7172 (2007), pp. 973–982.

[20] Pierre C. Havugimana et al. “A census of human soluble protein complexes”. In: *Cell* 150.5 (2012), pp. 1068–1081.

[21] Renwick C.J. Dobson et al. “The crystal structure of three site-directed mutants of Escherichia coli dihydrodipicolinate synthase: Further evidence for a catalytic triad”. In: *Journal of Molecular Biology* 338.2 (2004), pp. 329–339.

[22] Tina Perica et al. “The emergence of protein complexes: Quaternary structure, dynamics and allostery”. In: *Biochemical Society Transactions* 40.3 (2012), pp. 475–491.

[23] J.D. Watson et al. “Structure of small viruses”. In: *Nature* 177.4506 (1956), pp. 473–475.

[24] Olga M. Selivanova et al. “Structural and functional peculiarities of α-crystallin”. In: *Biology* 9.4 (2020).

[25] B. K. Pierscionek. “The refractive index along the optic axis of the bovine lens”. In: *Eye (Basingstoke)* 9.6 (1995), pp. 776–782.

[26] Arthur Laganowsky et al. “Crystal structures of truncated alphaA and alphaB crystallins reveal structural mechanisms of polydispersity important for eye lens function.” In: *Protein science : a publication of the Protein Society* 19.5 (2010), pp. 1031–1043.

[27] Marco Hiscox. “Investigating the structure and function of small heat-shock proteins in vertebrates”. PhD thesis. 2019.

[28] Kate L. Moreau et al. “Protein Misfolding and Aggregation in Cataract Disease and Prospects for Prevention”. In: *Trends in Molecular Medicine* 18.5 (2012), pp. 273–282.

[29] Marc A. Sprague-Piercy et al. "α-Crystallins in the Vertebrate Eye Lens: Complex Oligomers and Molecular Chaperones". In: *Annual Review of Physical Chemistry* 72 (2020), pp. 143–163.

[30] M. Bagchi et al. "Associated proteins of lens adherens junction". In: *Journal of Cellular Biochemistry* 86.4 (2002), pp. 700–703.

[31] William M. Jacobs et al. "Oligomers of Heat-Shock Proteins: Structures That Don't Imply Function". In: *PLoS Computational Biology* 12.2 (2016).

[32] Florian Stengel et al. "Quaternary dynamics and plasticity underlie small heat shock protein chaperone function". In: *Proceedings of the National Academy of Sciences of the United States of America* 107.5 (2010), pp. 2007–2012.

[33] Justin L.P. Benesch et al. "Small heat shock protein activity is regulated by variable oligomeric substructure". In: *Journal of Biological Chemistry* 283.42 (2008), pp. 28513–28517.

[34] Andrew J. Baldwin et al. "alphaB-Crystallin Polydispersity Is a Consequence of Unbiased Quaternary Dynamics". In: *Journal of Molecular Biology* 413.2 (2011), pp. 297–309.

[35] Frank Sobott et al. "Subunit Exchange of Multimeric Protein Complexes". In: *The Journal of Biological Chemistry* 277 (41) (2002), pp. 38921–38929.

[36] Alexander J. Painter et al. "Real-Time Monitoring of Protein Complexes Reveals their Quaternary Organization and Dynamics". In: *Chemistry and Biology* 15.3 (2008), pp. 246–253.

[37] Florian Stengel et al. "Dissecting Heterogeneous Molecular Chaperone Complexes Using a Mass Spectrum Deconvolution Approach". In: *Chemistry & Biology* 19.5 (2012), pp. 599–607.

[38] Martin Haslbeck et al. “A first line of stress defense: Small heat shock proteins and their function in protein homeostasis”. In: *Journal of Molecular Biology* 427.7 (2015), pp. 1537–1548.

[39] Thomas Kriehuber et al. “Independent evolution of the core domain and its flanking sequences in small heat shock proteins”. In: *FASEB Journal* 24.10 (2010), pp. 3633–3642.

[40] Antonio De Maio et al. “Extracellular heat shock proteins: a new location, a new function”. In: *Shock* 40.4 (2013), pp. 239–246.

[41] N De Miguel et al. “Differential Subcellular Localization of Members of the Toxoplasma gondii Small Heat Shock Protein Family”. In: *Eukaryotic Cell* 4.12 (2005), p. 1990.

[42] Elizabeth R. Waters et al. “Evolutionary analysis of the small heat shock proteins in five complete algal genomes”. In: *Journal of Molecular Evolution* 65.2 (2007), pp. 162–174.

[43] Erik Franck et al. “Evolutionary Diversity of Vertebrate Small Heat Shock Proteins”. In: ().

[44] J J Thomson. “Charge carried by the Cathode Rays .” In: *Philosophical Magazine* 44.293 (1897), pp. 1–24.

[45] John R. Yates. “A century of mass spectrometry: From atoms to proteomes”. In: *Nature Methods* 8.8 (2011), pp. 633–637.

[46] J. R. Chapman. *Protein and Peptide Analysis by Mass Spectrometry*. 1996.

[47] Kristina Lorenzen et al. “Native mass spectrometry as a tool in structural biology”. In: *Current Protocols in Protein Science* 62.Supplement 62 (Nov. 2010), pp. 1–17.

[48] G. R. Hilton et al. “Two decades of studying non-covalent biomolecular assemblies by means of electrospray ionization mass spectrometry”. In: *Journal of The Royal Society Interface* 9.70 (2012), pp. 801–816.

[49] Aneika C Leney et al. “Native Mass Spectrometry: What is in the Name?” In: *Journal of American Society for Mass Spectrometry* 28 (2017), pp. 5–13.

[50] Oliver Peetz et al. “LILBID and nESI: Different Native Mass Spectrometry Techniques as Tools in Structural Biology”. In: *Journal of the American Society for Mass Spectrometry* 30.1 (2019), pp. 181–191.

[51] Marie Barth et al. “Native mass spectrometry—A valuable tool in structural biology”. In: *Journal of Mass Spectrometry* 55.10 (2020), pp. 1–17.

[52] John B Fenn et al. “Electrospray Ionization for Mass Spectrometry of Large Biomolecules”. In: *Science* 246 (1989), pp. 64–71.

[53] Jason S Page et al. “Ionization and Transmission Efficiency in an Electrospray Ionization-Mass Spectrometry Interface”. In: *Journal of American Society for Mass Spectrometry* (2007), pp. 1582–1590.

[54] Charlotte Uetrecht et al. “Ion mobility mass spectrometry of proteins and protein assemblies”. In: *Chemical Society Reviews* 39.5 (2010), pp. 1633–1655.

[55] Kevin Giles et al. “A Cyclic Ion Mobility-Mass Spectrometry System”. In: *Analytical Chemistry* 91.13 (2019), pp. 8564–8573.

[56] Anas El-Aneed et al. “Mass spectrometry, review of the basics: Electrospray, MALDI, and commonly used mass analyzers”. In: *Applied Spectroscopy Reviews* 44.3 (2009), pp. 210–230.

[57] Sara Rosati et al. “Exploring an orbitrap analyzer for the characterization of intact antibodies by native mass spectrometry”. In: *Angewandte Chemie - International Edition* 51.52 (2012), pp. 12992–12996.

[58] Michael L. Poltash et al. “Fourier Transform-Ion Mobility-Orbitrap Mass Spectrometer: A Next-Generation Instrument for Native Mass Spectrometry”. In: *Analytical Chemistry* 90.17 (2018), pp. 10472–10478.

[59] Emmanuel D. Levy et al. “Assembly reflects evolution of protein complexes”. In: *Nature* 453.7199 (2008), pp. 1262–1265.

[60] Dennis R. Benjamin et al. “Mass spectrometry of ribosomes and ribosomal subunits”. In: *Proceedings of the National Academy of Sciences of the United States of America* 95.13 (1998), pp. 7391–7395.

[61] Yang Song et al. “Refining the structural model of a heterohexameric protein complex: Surface induced dissociation and ion mobility provide key connectivity and topology information”. In: *ACS Central Science* 1.9 (2015), pp. 477–487.

[62] Gili Ben-Nissan et al. “The application of ion-mobility mass spectrometry for structure/function investigation of protein complexes”. In: *Current Opinion in Chemical Biology* 42 (2018), pp. 25–33.

[63] J. Andrew Aquilina et al. “Polydispersity of a mammalian chaperone: Mass spectrometry reveals the population of oligomers in αB-crystallin”. In: *Proceedings of the National Academy of Sciences* 100.19 (2003), pp. 10611–10616.

[64] Jani Reddy Bolla et al. “Membrane protein-lipid interactions probed using mass spectrometry”. In: *Annual Review of Biochemistry* 88 (2019), pp. 85–111.

[65] Dror S. Chorev et al. “Protein assemblies ejected directly from native membranes yield complexes for mass spectrometry”. In: *Science* 362.6416 (2018), pp. 829–834.

[66] Marek Smoluch et al. “A brief history of mass spectrometry”. In: *Mass Spectrometry: An Applied Approach* (2019), pp. 5–8.

[67] Ayman El-Faramawy et al. “Efficiency of nano-electrospray ionization”. In: *Journal of the American Society for Mass Spectrometry* 16.10 (2005), pp. 1702–1707.

[68] P. Agarwal et al. “A review on analyzers for mass spectrometry”. In: *International Journal of Pharma and Bio Sciences* 8.4 (2017).

[69] Kyle L. Fort et al. “Expanding the structural analysis capabilities on an Orbitrap-based mass spectrometer for large macromolecular complexes”. In: *Analyst* 143.1 (2018), pp. 100–105.

[70] Victoria L. Morton et al. “Insight into virus capsid assembly from non-covalent mass spectroemtry”. In: *Mass Spectrometry Reviews* 27 (2008), pp. 575–595.

[71] Lintao Wang et al. “Detecting structural changes in viral capsids by hydrogen exchange and mass spectrometry”. In: *Protein Science* 10.6 (2001), pp. 1234–1243.

[72] Charlotte Uetrecht et al. “High-resolution mass spectrometry of viral assemblies: Molecular composition and stability of dimorphic hepatitis B virus capsids”. In: *Proceedings of the National Academy of Sciences of the United States of America* 105.27 (2008), pp. 9216–9220.

[73] Jessica Z. Bereszczak et al. “Sizing up large protein complexes by electrospray ionisation-based electrophoretic mobility and native mass spectrometry: Morphology selective binding of Fabs to hepatitis B virus capsids”. In: *Analytical and Bioanalytical Chemistry* 406.5 (2014), pp. 1437–1446.

[74] Carla Schmidt et al. “Comparative cross-linking and mass spectrometry of an intact F-type ATPase suggest a role for phosphorylation”. In: *Nature Communications* 4.1985 (2013), pp. 1–11.

[75] Kallol Gupta et al. “Identifying key membrane protein lipid interactions using mass spectrometry”. In: *Nature Protocols* 13.5 (2018), pp. 1106–1120.

[76] Adam Fineberg. “Interferometric scattering microscopy: from bacterial dynamics to molecular assembly”. PhD thesis. 2019.

[77] Daniel Cole et al. “Label-Free Single-Molecule Imaging with Numerical-Aperture-Shaped Interferometric Scattering Microscopy”. In: *ACS Photonics* 4.2 (2017), pp. 211–216.

[78] Gavin Young. “Weighing Single Molecules with Interferometric Scattering Microscopy”. PhD thesis. 2019.

[79] Norma J. Greenfield. “Using circular dichroism spectra to estimate protein secondary structure”. In: *Nature Protocols* 1.6 (2007), pp. 2876–2890.

[80] Sharon M. Kelly et al. “How to study proteins by circular dichroism”. In: *Biochimica et Biophysica Acta - Proteins and Proteomics* 1751.2 (2005), pp. 119–139.

[81] J. W. Gibbs. “On the Fundamental Formula of Statistical Mechanics, with Applications to Astronomy and Thermodynamics”. In: *Proceedings of the American Association for the Advancement of Science* 33 (1884), pp. 57–58.

[82] Georg K.A. Hochberg et al. “Structural principles that enable oligomeric small heat-shock protein paralogs to evolve distinct functions”. In: *Science* 359 (2018), pp. 930–935.

[83] Nobel Prize Outreach AB 2021. *All Nobel Prizes in Chemistry.*

[84] S Arrhenius. “On the reaction rate of the inversion of non-refined sugar upon souring”. In: *Z. Phys. Chem.* 4 (1889), pp. 226–248.

[85] S. Gottlieb et al. “Time Discretization Techniques”. In: *Handbook of Numerical Analysis* 17 (2016), pp. 549–583.

[86] Alexander J. Dear et al. “Statistical Mechanics of Globular Oligomer Formation by Protein Molecules”. In: *Journal of Physical Chemistry B* 122.49 (2018), pp. 11721–11730.

[87] Marija Iljina et al. “Quantifying Co-Oligomer Formation by α-Synuclein”. In: *ACS Nano* 12 (2018), pp. 10855–10866.

[88] Kaiming Zhang et al. “A novel mechanism for small heat shock proteins to function as molecular chaperones”. In: *Scientific Reports* 5 (2015), pp. 1–8.

[89] Spraha Bhandari et al. “Dodecameric structure of a small heat shock protein from Mycobacterium marinum M”. In: *Proteins: Structure, Function and Bioinformatics* 87.5 (2019), pp. 365–379.

[90] P. A. Frey. “Low-Barrier Hydrogen Bonds”. In: *Encyclopedia of Biological Chemistry: Second Edition*. Academic Press, 2013, pp. 756–759.

[91] Fériel Skouri-Panet et al. “Structural and functional specificity of small heat shock protein HspB1 and HspB4, two cellular partners of HspB5: Role of the in vitro hetero-complex formation in chaperone activity”. In: *Biochimie* 94.4 (2012), pp. 975–984.

[92] Aparna Vidyasagar et al. “Heat shock protein 27 (HSP27): biomarker of disease and therapeutic target”. In: *Fibrogenesis & Tissue Repain* 5.7 (2012), pp. 1–7.

[93] André-Patrick Arrigo et al. “Hsp27 Consolidates Intracellular Redox Homeostasis by Upholding Glutathione in Its Reduced Form and by Decreasing Iron Intracellular Levels”. In: *Antioxidants & Redox Signaling* 7.3 & 4 (2005), pp. 414–424.

[94] Raman Bakthisaran et al. “Small heat shock proteins: Role in cellular functions and pathology”. In: *Biochimica et Biophysica Acta - Proteins and Proteomics* 1854.4 (2015), pp. 291–319.

[95] Junna Hayashi et al. “The multifaceted nature of αB-crystallin”. In: *Cell Stress and Chaperones* 25.4 (2020), pp. 639–654.

[96] L. K. Muranova et al. “αB-Crystallin Phosphorylation: Advances and Problems”. In: *Biochemistry (Moscow)* 83.10 (2018), pp. 1196–1206.

[97] Satoru Kase et al. “αB-crystallin regulation of angiogenesis by modulation of VEGF”. In: *Blood* 115.16 (2010), pp. 3398–3406.

[98] Marta Maksimiuk et al. “αB-crystallin as a promising targetin pathological conditions – A review”. In: *Annals of Agricultural and Environmental Medicine* 27.3 (2020), pp. 326–334.

[99] Eiryo Sha et al. “Functional and structural characterization of HspB1/Hsp27 from Chinese hamster ovary cells”. In: *FEBS Open Bio* 9.10 (2019), pp. 1826–1834.

[100] Hidenori Ito et al. “Phosphorylation-induced Change of the Oligomerization State of αB-crystallin”. In: *Journal of Biological Chemistry* 276.7 (2001), pp. 5346–5352.

[101] Blagojce Jovcevski et al. “Phosphomimics destabilize Hsp27 oligomeric assemblies and enhance chaperone activity”. In: *Chemistry and Biology* 22.2 (2015), pp. 186–195.

[102] Yuki Sugiyama et al. “Muscle Develops a Specific Form of Small Heat Shock Protein Complex Composed of MKBP / HSPB2 and HSPB3 during Myogenic Differentiation”. In: *The Journal of Biological Chemistry* 275.2 (2000), pp. 1095–1104.

[103] Evgeny V. Mymrikov et al. “Regulation of small heat-shock proteins by hetero-oligomer formation”. In: *Journal of Biological Chemistry* 295.1 (2020), pp. 158–169.

[104] Blagojce Jovcevski et al. “Phosphomimics Destabilize Hsp27 Oligomeric Assemblies and Enhance Chaperone Activity”. In: *Chemistry & Biology* 22.2 (2015), pp. 186–195.

[105] J. Andrew Aquilina et al. “Structural and functional aspects of hetero-oligomers formed by the small heat shock proteins αB-crystallin and HSP27”. In: *Journal of Biological Chemistry* 288.19 (2013), pp. 13602–13609.

[106] Michael T. Marty et al. “Bayesian deconvolution of mass and ion mobility spectra: From binary interactions to polydisperse ensembles”. In: *Analytical Chemistry* 87.8 (2015), pp. 4370–4376.

[107] Justin L.P. Benesch et al. “Tandem Mass Spectrometry Reveals the Quaternary Organization of Macromolecular Assemblies”. In: *Chemistry and Biology* 13.6 (2006), pp. 597–605.

[108] Pawasuth Saengdee et al. “Surface modification of silicon dioxide, silicon nitride and titanium oxynitride for lactate dehydrogenase immobilization”. In: *Biosensors and Bioelectronics* 67 (2015), pp. 134–138.

[109] Miranda P Collier. “Small heat shock protein interactions with in vivo partners”. PhD thesis. 2018.

[110] Joseph Horwitz et al. *Methods in enzymology*. 1998, pp. 365–383.

[111] Elisabeth Gasteiger et al. “The Proteomics Protocols Handbook”. In: *The Proteomics Protocols Handbook* (2005), pp. 571–608.

[112] Gavin Young et al. “Quantitative mass imaging of single molecules in solution”. In: *Science* 360 (2017), pp. 423–427.

[113] Evgeny V. Mymrikov et al. “Regulation of small heat-shock proteins by hetero-oligomer formation”. In: *Journal of Biological Chemistry* 295.1 (2020), pp. 158–169.

[114] Maria V. Sudnitsyna et al. “The Role of Intrinsically Disordered Regions in the Structure and Functioning of Small Heat Shock Proteins”. In: *Current Protein & Peptide Science* 13.1 (2012), pp. 76–85.

[115] Wilbert C. Boelens. “Structural aspects of the human small heat shock proteins related to their functional activities”. In: *Cell Stress and Chaperones* 25.4 (2020), pp. 581–591.

[116] Daniel Foreman-Mackey et al. “emcee : The MCMC Hammer ”. In: *Publications of the Astronomical Society of the Pacific* 125.925 (2013), pp. 306–312.

[117] Kyeong Kyu Kim et al. “Crystal structure of a small heat-shock protein”. In: *Nature* 394.6693 (1998), pp. 595–599.

[118] Tetsuya Abe et al. “StHsp14.0, a small heat shock protein of Sulfolobus tokodaii strain 7, protects denatured proteins from aggregation in the partially dissociated conformation”. In: *Journal of Biochemistry* 150.4 (2011), pp. 403–409.

[119] Rob L.M. Van Montfort et al. “Crystal structure and assembly of a eukaryotic small heat shock protein”. In: *Nature Structural Biology* 8.12 (2001), pp. 1025–1030.

[120] Yuya Hanazono et al. “Nonequivalence observed for the 16-meric structure of a small heat shock protein, SpHsp16.0, from Schizosaccharomyces pombe”. In: *Structure* 21.2 (2013), pp. 220–228.

[121] Evgeny V. Mymrikov et al. “The chaperone activity and substrate spectrum of human small heat shock proteins”. In: *Journal of Biological Chemistry* 292.2 (2017), pp. 672–684.

[122] Di Wu et al. “The N terminus of the small heat shock protein HSPB7 drives its polyQ aggregation-suppressing activity”. In: *Journal of Biological Chemistry* 294.25 (2019), pp. 9985–9994.

[123] Lydia K. Muranova et al. “Quaternary structure and hetero-oligomerization of recombinant human small heat shock protein HspB7 (cvHsp)”. In: *International Journal of Molecular Sciences* 22.15 (2021).

[124] B. K. Pierscionek et al. “The refractive index along the optic axis of the bovine lens”. In: *Journal of The American Optometric Association* 66 (1995), pp. 739–743.

[125] M. Bagchi et al. “Heat shock proteins of adult and embryonic human ocular lenses”. In: *Journal of Cellular Biochemistry* 84.2 (2002), pp. 278–284.

[126] Arthur C. Beall et al. “Cyclic nucleotide-dependent vasorelaxation is associated with the phosphorylation of a small heat shock-related protein”. In: *Journal of Biological Chemistry* 272.17 (1997), pp. 11283–11287.

[127] Catherine M. Dreiza et al. “The small heat shock protein, HSPB6, in muscle function and disease”. In: *Cell Stress and Chaperones* 15.1 (2010), pp. 1–11.

[128] Fazhao Li et al. “Study of HSPB6: Insights into the Properties of the Multifunctional Protective Agent”. In: *Cellular Physiology and Biochemistry* 44.1 (2017), pp. 314–332.

[129] Olesya V. Bukach et al. “Some properties of human small heat shock protein Hsp20 (HspB6)”. In: *European Journal of Biochemistry* 271.2 (2004), pp. 291–302.

[130] Kanefusa Kato et al. “Purification and characterization of a 20-kDa protein that is highly homologous to αB crystallin”. In: *Journal of Biological Chemistry* 269.21 (1994), pp. 15302–15309.

[131] Michelle Heirbaut et al. “The preferential heterodimerization of human small heat shock proteins HSPB1 and HSPB6 is dictated by the N-terminal domain”. In: *Archives of Biochemistry and Biophysics* 610 (2016), pp. 41–50.

[132] Michelle Heirbaut et al. “Specific sequences in the N-terminal domain of human small heat-shock protein HSPB6 dictate preferential hetero-oligomerization with the orthologue HSPB1”. In: *Journal of Biological Chemistry* 292.24 (June 2017), pp. 9944–9957.

[133] Lawrence A Kelley et al. “The Phyre2 web portal for protein modeling, prediction and analysis”. In: *Nature Protocols* 10.6 (2015), pp. 845–858.

[134] Daichi G. Suzuki et al. “The stepwise development of the lamprey visual system and its evolutionary implications”. In: *Biological Reviews* 93.3 (2018), pp. 1461–1477.

[135] Byrappa Venkatesh et al. “Elephant shark genome provides unique insights into gnathostome evolution”. In: *Nature* 505.7482 (2014), pp. 174–179.

[136] Norma J. Greenfield. “Using circular dichroism collected as a function of temperature to determine the thermodynamics of protein unfolding and binding interactions”. In: *Nature Protocols* 1.6 (2007), pp. 2527–2535.

[137] Bernard Labedan et al. “Using Quaternary Structures to Assess the Evolutionary History of Proteins: The Case of the Aspartate Carbamoyltransferase”. In: *Molecular Biology and Evolution* 21.2 (2004), pp. 364–373.

[138] Michael D.W. Griffin et al. “Evolution of Quaternary Structure in a Homotetrameric Enzyme”. In: *Journal of Molecular Biology* 380.4 (2008), pp. 691–703.