SOLUTION-PHASE DYNAMICS OF THE HEPATITIS B VIRUS CAPSID: KINETICS-BASED ASSAYS FOR STUDY OF SUPRAMOLECULAR COMPLEXES.

by

Jonathan Kyle Hilmer

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

Doctor of Philosophy

in

Biochemistry

MONTANA STATE UNIVERSITY
Bozeman, Montana

November, 2009

© Copyright

by

Jonathan Kyle Hilmer

2009

All Rights Reserved

APPROVAL

of a dissertation submitted by

Jonathan Kyle Hilmer

This dissertation has been read by each member of the dissertation committee and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the Division of Graduate Education.

Dr. Brian Bothner

Approved for the Department of Chemistry and Biochemistry

Dr. David Singel

Approved for the Division of Graduate Education

Dr. Carl A. Fox

STATEMENT OF PERMISSION TO USE

In presenting this dissertation in partial fulfillment of the requirements for a doctoral degree at Montana State University, I agree that the Library shall make it available to borrowers under rules of the Library. I further agree that copying of this dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for extensive copying or reproduction of this dissertation should be referred to ProQuest Information and Learning, 300 North Zeeb Road, Ann Arbor, Michigan 48106, to whom I have granted "the exclusive right to reproduce and distribute my dissertation in and from microform along with the non-exclusive right to reproduce and distribute my abstract in any format in whole or in part."

Jonathan Kyle Hilmer

November, 2009

DEDICATION

I dedicate this dissertation to my wife Kimberly, for her infinite patience and support.

ACKNOWLEDGEMENTS

I would like to thank all the faculty, staff, and students of Montana State University with whom I have worked during my time here. Every one of them has contributed to my development in the sciences, and none of this would have been possible without them.

I would also like to personally thank Dr. Brian Bothner, who introduced me to viral dynamics and mass spectrometry. I have been blessed with a history of teachers who have inspired wonder in the intricate workings of the world, and he continues their grand tradition.

TABLE OF CONTENTS

1. INTRODUCTION - STRUCTURAL DYNAMICS OF VIRAL CAPSIDS IN SOLUTION 1

Introduction 1
Quaternary Dynamics 2
Maturation Associated Dynamics in a Bacteriophage 3
Maturation Associated Dynamics in Small RNA Viruses 4
Structural Transitions 6
Solution-Phase Equilibrium Dynamics 7
Virus Particles are Dynamic 7
Proteolysis and Mass Analysis 8
Methods for Studying Viruses in Solution 11
Spectroscopy 12
NMR 15
Raman 16
Small Angle X-ray Scattering (SAXS) 17
Small Angle Neutron Scattering (SANS) 18
Computation 19
Labeling Experiments 21
Chemical Reactivity 21
Hydrogen Exchange 22
Kinetic Hydrolysis 25
Summary 28

2. KINETIC HYDROLYSIS OF CP149: LARGE-SCALE DYNAMICS AS A FUNCTION OF TEMPERATURE 29

HBV as a Model Capsid 29
Structure and Assembly Studies of Cp149 32
Implications of Dynamics 39
Experimental Techniques and Planning 40
Inherent Problems of the Cp149 System 40
Detection Assays for Proteolysis 41
Experimental Conditions 42
Kinetic Hydrolysis Modeling 43
Determination of k_{cleave} 49
SDS-PAGE Analysis of Kinetic Hydrolysis 54
Restriction of Reaction Progression 54
Inhibition of Proteolysis 56
Protocol Variations and Specifics 59
SDS-PAGE Results 60

TABLE OF CONTENTS – CONTINUED

Calculations and Discussion of SDS-PAGE Data 62

3. KINETIC HYDROLYSIS AND THE INFLUENCES OF ASSEMBLY EFFECTORS 66

Introduction 66
Analysis Methods: HPLC-MS 67
RP and SEC Chromatography 68
Carry-Over 70
Automated SEC-HPLC-MS 72
Hardware 72
Protocols 75
Data Processing 78
Shared Parameter Kinetic Model of Proteolysis 81
Structure to Model 83
Model Features 84
Data Fitting 87
Results and Discussion 88
Influences of HAP12 89
Influence of Zinc 92
Conclusions 93

4. HX ASSAYS OF ASSEMBLY EFFECTORS: SMALL-SCALE DYNAMICS ... 97

Introduction and Background 97
Back-Exchange 97
Theory 98
Chromatography Methods 101
Protocol 103
Back-Exchange Tests 105
Results 106
Discussion 110
Conclusions 113

5. HBV STRUCTURE AND DYNAMICS 116

Structure of the HBV Core Protein Dimer 116
Implied Dynamics of HBV Biology 119
Viral Life Cycle 119
Protein Phosphorylation 121
Capsid Protein Conformations 123

TABLE OF CONTENTS – CONTINUED

Structure and Biology Conclusions 125
Unified Model of Kinetic Hydrolysis and Hydrogen-Deuterium Exchange 128

6. PHYSICAL SIGNAL MODULATION 135

Introduction and Background 135
Abbreviations and Conventions 136
Effects of Data Sampling Interval on Mass Accuracy and Precision 136
Reduction of the Data Interval Δx 144
Physical Signal Modulation Increases Data Sampling and Reduces Noise .. 145
Experimental 146
Materials 146
Instrumentation 146
Spectra Simulations 147
Spectra Calculation 147
Physical Signal Modulation 149
Data Processing 149
Results and Discussion 150
Centroid Calculations of Simulated Spectra 150
Centroid Calculations of Experimental Data 153
Centroid Errors and Lock-Mass Corrections 157
Physical Signal Modulation 159
Improvements in Δx 162
Conclusions 163

REFERENCES CITED 166

APPENDICES 180

APPENDIX A: Supplementary biophysical research tools 181

LIST OF TABLES

Table Page

1.1 Benefits and drawbacks of techniques to study viral dynamics. 14

2.1 Kinetic and thermodynamic parameters from the IRTPP cleavage site. ... 63

LIST OF FIGURES

Figure Page

1.1 Viral particles exhibit multiple forms of dynamics in solution. 9
1.2 Utility of various tools for study of viral dynamics. 13
1.3 Kinetic hydrolysis of the HBV capsid protein Cp149. 27
2.1 Structures of Cp149 capsid and dimer 31
2.2 Assembly and disassembly displayed marked hysteresis 34
2.3 A family of HAP compounds affect HBV assembly 37
2.4 HAPs induce aberrant assembly structures 37
2.5 Classic fluorescence assay of the IRTPP model peptide 50
2.6 Single-stage fluorescence assay 51
2.7 SDS-PAGE of kinetic hydrolysis reactions 61
2.8 Observed proteolysis trends of Cp149 follow expected patterns 62
2.9 Kinetics and thermodynamics of Cp149 dynamics 64
3.1 Cp149 carryover on RP column 71
3.2 SEC provides minimal carryover and easy correction 73
3.3 Schematic of the Cp149 proteolytic model 85
3.4 HAP12 and $Zn^{(2+)}$ have substantial effects of Cp149 dynamics 90
3.5 Clustering analysis reveals regulated but unknown masses 95
3.6 Identification of masses from clusters 96
4.1 Rapid chromatography for HX 104
4.2 Intact protein mass additions as a result of HX 107
4.3 HX mass shifts of Cp149 in response to assembly effectors 109
5.1 Crystal structure of the core protein dimer 118
5.2 Model of HBV core protein dynamics 130
5.3 Summary of proteolysis measurements of dynamics 132

LIST OF FIGURES – CONTINUED

Figure Page

6.1 Data sampling of peaks gives discrete data sets with random centroiding errors. 140

6.2 Accuracy error in centroid calculation as a function of $\Delta m/z$ 152

6.3 Centroid peak-calculations show artificial peak shifts. 156

6.4 Physical signal modulation prior to spectral alignment shows coherent noise and highly predictable behavior. 161

6.5 Comparisons between raw spectra and PSM-processed spectra. 164

A.1 Overview of the PSM software code 187

ABSTRACT

Viruses are the most abundant form of life on the planet. Many forms are pathogenic and represent a major threat to human health, but viruses can also be useful nanoscale tools: as adjuvants, gene therapy agents, antimicrobials, or functionalized nanoscale building blocks.

Viruses have historically been viewed as static and rigid delivery vehicles, but over the last few decades they have been recognized as flexible structures. Their structural dynamics are a crucial element of their functionality, and they represent a substantial target for antiviral strategies. To overcome the inherent problems of characterizing the biophysics of supramolecular complexes, we have developed a set of kinetic assays to probe capsid motion at several different amplitudes. The first assay, kinetic hydrolysis, works via the differential cleavage of folded versus unfolded proteins, and reports on large-scale conformational changes. The second assay, hydrogen-deuterium exchange, is a probe of small-amplitude dynamics. Both of these assays were used to study the solution-phase dynamics of the hepatitis B virus (HBV) capsid under the influence of assembly effectors and temperature.

The results of these assays indicate that the HBV capsid adopts multiple conformations in response to the external environment. The dimeric subunit becomes primed for assembly via an entropically-driven process, but once assembled the capsid has reduced dependence on hydrophobic contacts. Depending on the assembly state, the subunit protein has varying response to assembly effectors, with changes to both small-amplitude and large-amplitude motions.

The sum of the assay results indicate that the HBV capsid protein is capable of rotational translocations of the alpha-helices, while maintaining most of the secondary structure. Concerted structural shifts are implied, consistent with an allosteric model, which helps to explain previously observed allostery of capsid assembly and response to drugs.

INTRODUCTION - STRUCTURAL DYNAMICS OF VIRAL CAPSIDS IN SOLUTION

Introduction

A great deal of what we know about viruses has come from the analysis of capsid structure. Early structural information came from electron micrographs and X-ray diffraction patterns which displayed the shape and general quaternary organization of particles. Then, with the maturation of X-ray crystallography, the elegance and intricacies of viral capsids were revealed. Icosahedral capsids represent some of the most stunning structural models produced, and even nonscientists are intrigued by their symmetrical beauty. The symmetry and size of icosahedral particles also made them good subjects for cryo-electron microscopy (cryo-EM) and numerous technical advances were driven by structural virology. In one respect, the structural models may have been too convincing, because many virologists were persuaded to think of the protein capsid only as the rigid shell depicted in the static images. However, even as the structural models were shaping the way virologists and structural biologists thought about the particles, biochemical evidence was accumulating to suggest that, in solution, there was more to these structures than met the eye. Capsid proteins are responsible for an array of functions critical for completion of a virus lifecycle. These include particle assembly, intracellular transport, genome protection and release, and, in the case of non-enveloped viruses, receptor binding. It is now clear that protein dynamics have an essential role in each of these steps. The most obvious indication that capsids are active structures comes from the dramatic protein rearrangements which can occur after assembly and before the particle becomes fully infective. These maturation associated changes have been caught in sequential still-life models for a

number of viruses. Each step in the progression represents a distinct capsid form that can be isolated and often has unique physical properties. A second more subtle, yet equally important, form of dynamics exists in particles at each step along this progression. Unlike the large-scale maturation induced changes that can be kinetically trapped, the second mode of dynamics is a solution-phase equilibrium process and can not be captured by the structural models. Equilibrium dynamics are important for each of the capsid protein functions mentioned above and change throughout the maturation process. The two categories of dynamics differ in the physical characteristics of the dynamic motion, as well as the techniques that can be used to study them. This chapter will focus on viruses with icosahedral capsids, beginning with discussions of quaternary dynamics and solution-phase equilibrium dynamics. It will conclude with an overview of techniques that have been used to study virus particles in solution, the information that can be obtained from such experiments, and future directions.

Quaternary Dynamics

Large scale quaternary rearrangement of subunits in icosahedral capsids is generally associated with maturation events or swelling and contraction induced by solution conditions [1]. These changes can involve major alterations of capsid size or geometry, but are not necessarily accompanied by large changes in secondary or tertiary elements. The capsid as a whole undergoes a symmetric transformation, with radial translocation and/or subunit rotation, creating the net effect of a larger or smaller capsid that still retains its overall symmetry and general features (Figure 1.1). In a biological context, these are generally one-way events triggered by packaging of nucleic acids [2], trafficking solution conditions [3], [4], or receptor binding [5]. Although it

may be possible in some cases to reverse the process [4], the steps are not necessarily populated as an equilibrium in solution. In many cases, the one-way nature of quaternary transitions serve as a regulatory gateway for structural maturation that coincides with key events in the viral lifecycle. Because the particles can become kinetically trapped in a particular form, quaternary dynamics have been well-characterized using classic structural techniques [1]. This has permitted detailed comparison of the pre- and post-transition structures, and in some cases the transition itself has been directly observable.

Maturation Associated Dynamics in a Bacteriophage

Icosahedral capsids often have a spherical form after assembly, adopting their final quasi-equivalent form upon maturation. One of the most interesting and best detailed characterizations of this process involves the bacteriophage HK97. HK97 is a dsDNA λ-like coliphage with a T=7 capsid possessing a portal structure in place of one pentamer [6]. Progression from the initial Prohead form to the final mature Head II has been followed in vitro by co-expressing the major capsid protein GP5 and the viral protease GP4. In the mature form of the capsid, GP5 proteins are topologically linked together, creating a protein chain-mail [7]. Maturation begins with digestion of the N-terminal 103 amino acids by the viral protease. The subsequent quaternary rearrangement involves subunit rotation and radial expansion (with the diameter going from 450 to 650 Å) [3, 8]. This dramatic rearrangement is also accompanied by a thinning of the capsid cross-section, and a transition from a nearly spherical shape to a much more angular polyhedron. Once expanded, the protein shell uses autocatalytic cross-linking between Lys and Asn side-chains to create hexamers and pentamers that are concatenated together, but not directly cross-linked. The final product is a remarkably thin, yet robust protein shell. Similar Prohead to Head

transitions have been described for P22, Phi6, Phi 29, T7 [1], and is not limited to bacteriophages, as members of the Adenoviridae and Herpesviridae [9] undergo similar transitions. However, HK97 is the only particle known to use a cross-linked chain-mail, and the precise choreography required to generate the final structure makes this a jewel in structural virology. The mature capsid is a rigid structure, capable of withstanding the estimated 60 atmospheres of pressure exerted by the packaged DNA [6, 7]. As such, the dynamics which were important for assembly and maturation have been quenched, but because HK97 is a bacteriophage with an active portal assembly, the mature capsid does not need to participate in receptor binding or other events which might require conformational flexibility.

Maturation Associated Dynamics in Small RNA Viruses

A second recurring theme in maturation associated dynamics is the use of an autocatalytic proteolysis event to trigger maturation. For small RNA viruses of the Picornaviridae, Nodaviridae and Tetraviridae, assembly of the procapsid positions the subunits such that auto-hydrolysis of the protein chain occurs. Cleavage of the subunit acts as a molecular switch, preventing a reversal of the assembly process allowing access to local energy minima not populated by the procapsid. This event has been described with some detail for Flock House Virus (FHV) [10] and is a requirement for generation of infective particles [11]. Cleavage of the alpha capsid protein in FHV produces a 363 amino acid beta-protein, which comprises the capsid shell and cellular receptor(s), and the 44 residue gamma-peptide which is situated inside the capsid shell next to the RNA in structural models. Both in vivo and in vitro experiments with FHV have shown that the gamma-peptide (found in noda and tetraviruses) is involved in membrane penetration/disruption [12], analogous to the functional role of VP4 in picornaviruses [13].

The tetraviruses provide an interesting example of the relationship between quaternary dynamics and autohydrolysis. Nudarelia capensis omega virus (NωV), is the best studied member of this family of T=4 viruses that only infect members of the Lepidoptera. In the case of the tetraviruses, pH can be used to induce a transition from the procapsid to a smaller mature capsid: the structures of both forms of NωV have been determined at moderate resolution [14, 15, 16]. In addition to a radial contraction of 16%, subunit rotation and tertiary changes in an internal helical region occur in the transition. These rearrangements can be reversed by simply raising the pH, so long as the autohydrolytic cleavage has not proceeded beyond 15% of the subunits [16]. The driving force for structural rearrangement is electrostatic and it appears that pH may be the trigger for maturation in vivo as well, based on recent data showing that infection by NωV induces apoptosis and a decrease in intracellular pH in insect midgut cells [17]. As mentioned above, NωV uses a capsid protein processing event that is similar to FHV in which the C-terminal region is hydrolyzed by an asparagine side-chain-mediated attack on the peptide backbone [10]. A mutant form of NωV in which the catalytic function has been abrogated can repeatedly transition between forms, conclusively demonstrating that hydrolysis disconnects the driving force for the procapsid-capsid transition [18]. As a model system, NωV is significant because it was the first virus for which the conformational change between a capsid intermediate and the mature form were monitored in real-time. This was initially accomplished using solution small angle X-ray scattering [4]. Subsequently, a number of biophysical techniques have been applied to characterize both forms and the transition, including fluorescence [19], chemical reactivity [20], proteolysis [21], and FT-IR [21]. Considering the size and complexity of a T=4 capsid, the 240 subunits react surprisingly rapidly upon a pH reduction from 7.5 to 5.0, completing

the rearrangement within 100 ms. Hydrolysis occurs much more slowly, having a half-life of hours [16].

Structural Transitions

One of the unifying themes of quaternary dynamic transitions is their usage as a delineator between distinct structural states. These states often have dramatic differences in structural stability, and the end product of the transition is highly tailored for its environment. In the case of HK97, the end product is a highly robust capsid capable of packaging large amounts of DNA. Viruses which lack a portal assembly require a capsid with the functional and structural flexibility to dock with the host and release the genome using only the capsid shell subunits. Such functionality can be conferred by making the end product of maturation a metastable state: a local energy minimum robust enough to serve as a delivery vehicle and yet primed to release its contents upon the right environmental signal. Two examples of mature metastable particles are poliovirus [22] and NωV [21]. In many cases, metastability can be inferred from available data on the infection process without biophysical confirmation. One such case involves the puzzling case of parvoviruses, a group of small single-stranded DNA viruses with a non-enveloped T=1 icosahedral capsid. Parvoviruses all use receptor mediated endocytosis for internalization. Lacking an envelope that would allow membrane fusion, they gain entry into the cytoplasm by deploying a phospholipase domain [23]. Structural models clearly indicate that this domain resides on the inside of the capsid [24]. Release of the 130 amino acid domain occurs after receptor binding and endosomal acidification without particle disruption [25]. Heating can be used as a surrogate trigger, again without particle disruption. Analyses of the structural models reveal no pores or channels that could accommodate such a large domain. Mutational analysis [23, 26] and post release cryo-EM data identify regions

that affect translocation and are altered by the process, respectively; however, where and how a folded globular domain of ∼15 kDa extricates itself from the capsid interior is still a mystery.

Solution-Phase Equilibrium Dynamics

While maturation events largely involve quaternary rearrangements, sometimes with tertiary structural triggers, solution-phase dynamics is characterized by rapid equilibrium motion with localized perturbations in the tertiary structure of the subunits (Figure 1.1). These motions have been described as breathing [27], but the motion may involve only a subset of the subunits or capsid population, and there is no evidence that it is a symmetric transition. Due to the rapid equilibrium, it is not possible to isolate populations of just one state, and crystal packing forces or reduced temperatures quench these motions, severely limiting the applicability of X-ray crystallography and cryo-EM as viable study tools. Instead, a variety of solution-phase techniques have been applied to detect and quantitate equilibrium dynamics. The technical limitations and difficulty in performing these assays has hampered a full characterization of equilibrium structural dynamics in capsids.

Virus Particles are Dynamic

Dynamic protein regions in assembled particles were first identified in NMR experiments on the plant virus Cowpea Chlorotic Mottle Virus (CCMV) [28]. The observed dynamics were on the timescale of 1-10 nanoseconds, and were associated with the N-terminus of the capsid. The number of dynamic side chains dramatically increased in capsids without RNA, and it was proposed that RNA induced the formation of internal alpha-helices. This suggested to the authors that the N-terminal domain must be

important for assembly of infectious particles. A different form of rapid dynamics was observed for picornaviruses. In poliovirus, antibodies raised against intact particles actually recognized a domain that was on the internal surface of the static structural model [13, 22]. Further studies on poliovirus using antibodies demonstrated that the externalization of internal domains on VP1 and VP4 could be reversed and involved regions important for cell entry [29, 30, 31]. Antiviral drugs that increase the thermal stability of capsids, such as the hydrophobic WIN compounds, are known to prevent transition out of the metastable phase in picornaviruses [32]. These observations for poliovirus and other picornaviruses led to the concept of the mature particle as a metastable structure, as discussed above. Later computational studies, using molecular dynamics, suggested that the increased stability had an entropic basis, diminishing the entropy gain of uncoating [33, 34].

Proteolysis and Mass Analysis

Initially, the extent of the dynamics responsible for the reversible exposure of internal domains was not fully appreciated. A striking experiment that changed this involved the use of limited proteolysis and mass spectrometry. It had recently been demonstrated that this combination of a standard biochemical technique with a powerful analytical platform could be used to investigate capsid protein dynamics [35]. Time-dependent analysis of a protease reaction using mass spectrometry provided precise identification of the cleavage sites and qualitative kinetic data on the solution-phase dynamics of a capsid. Researchers in the lab of Gary Siuzdak at The Scripps Research Institute showed that the dynamic motion of the internal VP4 protein in human rhinovirus made it more cleavage-accessible than regions on the exterior of the capsid. Importantly, by using proteases as a probe, the whole capsid was screened for dynamics without bias towards any specific site. When carried out in the presence

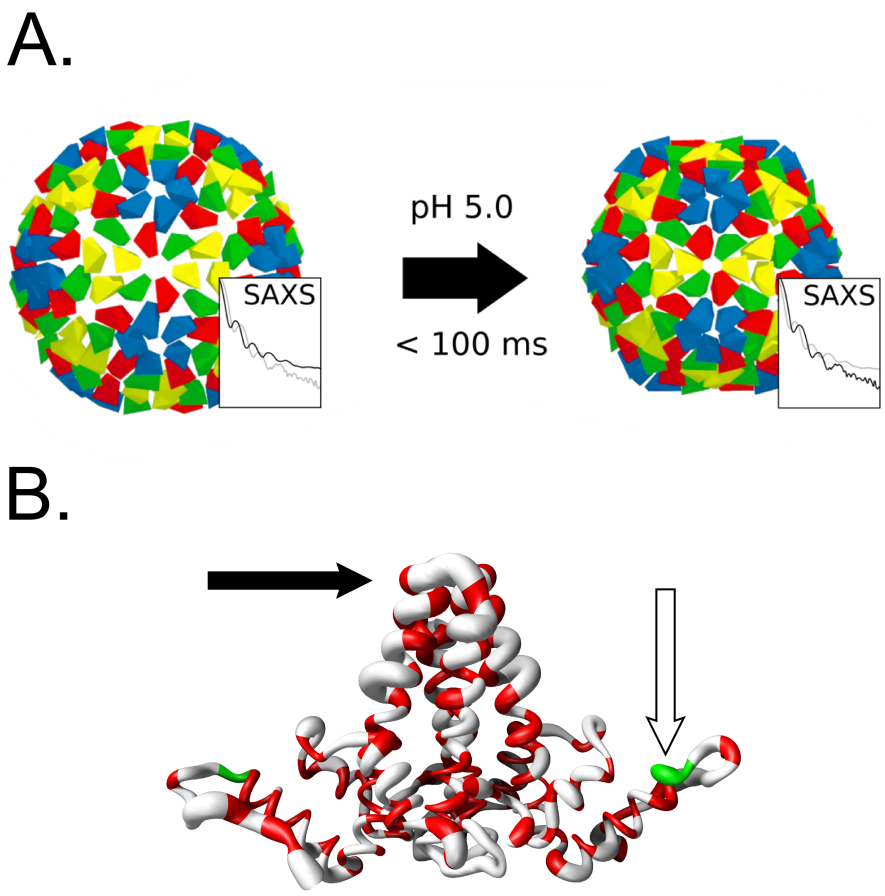


Figure 1.1: Viral particles exhibit multiple forms of dynamics in solution. A.) Maturation of the Nudaurelia capensis omega virus VLP. The contraction and shifting of the subunits reduces the diameter by 16% and occurs very rapidly, within 100 ms. The transition was studied by SAXS (data shown in insets as the scattering vector), which revealed substantial changes in the radial density distribution. Adapted from Canady, 2001. B.) The HBV capsid protein Cp149 undergoes equilibrium dynamics. The crystal structure of the dimer is shown (extracted from the crystal structure obtained of the whole capsid). The protein backbone is shown with thickness scaled according to B-factors. Red regions indicate possible protease cleavage locations, while the green region (indicated with white arrow) shows the only cleavage location observed when the protein is in the dimer form. The top of the tower (black arrow) has substantial structural rearrangements between capsid and dimer conformations (A. Zlotnick, personal communication).

of the WIN stabilizing compounds, the proteolysis experiments showed a loss of the dynamic breathing motion [27]. These results definitively established that icosahedral capsids exist as an ensemble of conformations in solution, only one of which is captured in the structural models.

The application of proteolysis and mass spectrometry to the study of capsid protein dynamics was initially unintentional. Intrigued by the potential of identifying antigenic peptides on the surface of virus particles, Bothner, Johnson, and Siuzdak exposed FHV, a member of the Nodaviridae, to proteases and identified the released peptides using MALDI and electrospray mass spectrometry. Surprisingly, the very first peptides to be generated mapped to the interior of the capsid, positioned next to the RNA [35]. Following extensive control experiments to assure that a subpopulation of disrupted particles were not the source of the released peptides, it was accepted that nanometer range, reversible protein motion was responsible for the exposure of internal domains on the capsid surface and subsequent protease mediated cleavage. This indicates that FHV particles are present as an ensemble of conformers in solution and is consistent with the fact that transient exposure of the internally located amphipathic gamma-peptide onto the particle surface mediates interaction with host cell membranes [12].

In addition to their contribution to understanding the maturation process in icosahedral capsids, tetraviruses have proven to be an interesting model system for the study of solution-phase properties. As described above, their T=4 capsids undergo an autocatalytic cleavage during maturation that is very similar to what occurs in FHV. The original structural model of NωV had distinct similarities to FHV with respect to the arrangement of gamma-peptides around the five-fold axes. It was therefore reasoned that the gamma-peptide would also be highly dynamic and exposed to the capsid surface. However, the proteolytic susceptibility and chemical reactivity were

much less than that seen in FHV [21]. Subsequent refinement of the X-ray structure revealed an additional set of helices contributed by the N-termini of the protein, creating a previously undetected interaction with the gamma-peptides at the 5-fold axes [14]. In order for the gamma-peptides to transiently sample external positions at the 5-fold axes, a strong set of interactions would need to be broken. The work with NωV and Helicoverpa armigera stunt virus (HaSV) demonstrate the power of solution-based approaches in cases where structural data is not available or may lead to incorrect conclusions.

Methods for Studying Viruses in Solution

Equilibrium-type dynamic motions within viral capsids are inherently difficult to study, as it is not possible to kinetically trap conformational isoforms for static structural studies. When a population can be found in a nearly-homogeneous state, it will typically be the low-energy form that dominates, while it is the higher-energy, activated form which is often of most interest as the functional species. Experimentally, methods need to be specific enough to resolve the difference between the two states and sensitive enough to provide accurate measurements despite a heavy population bias for one form over the other (Figure 1.2). While these challenges have by no means been met in entirety, a wide variety of approaches have proven successful for detecting capsid protein dynamics in solution.

Due to their large size and high degree of symmetry, the application of standard biophysical approaches for studying protein dynamics in viral capsids has been challenging. For example, NMR, electron paramagnetic resonance (EPR), and Förster resonance energy transfer (FRET) are standard methods used to investigate protein dynamics in mono and multimeric proteins. With respect to the large supramolecular

complexes, which virus particles are, only NMR has made a significant contribution. However, this has not stopped researchers from trying, and steady advances are being made using both standard biophysical techniques and creative new approaches (Table 1.1). Regardless of the approach, certain precautions must be taken to insure the validity of the results. First most, is the assurance of a homogeneous population of intact particles. Depending on the experiment, even a few percent of unassembled protein could seriously bias the results. Establishing sample quality can be a non-trivial task, requiring size exclusion chromatography, light scattering, and gradient centrifugation, along with a well-established understanding of capsid stability. Meeting these requirements is a necessity that may rule out some experimental techniques, due to protein concentration limitations or solution conditions. Even when a capsid is stable under the required conditions, the experiment itself may have destabilizing effects. Small angle X-ray scattering (SAXS), resonance-based, and covalent labeling experiments can be disruptive, and have the potential to alter the dynamics, if not destabilize the capsid itself.

Spectroscopy

Historically NMR, Raman, and small angle scattering techniques have been the dominant spectroscopic approaches to study virus capsids in solution. This should not be taken to imply that these techniques are the only viable ones: FRET, EPR, and other methods all have potential advantages, limited primarily by the ability to produce appropriately labeled capsids. Considering the challenges of detecting small sub-populations of highly dynamic regions, additional sensitive techniques such as paramagnetic relaxation enhancement should not be overlooked as potential strategies for examining solution-phase dynamics.

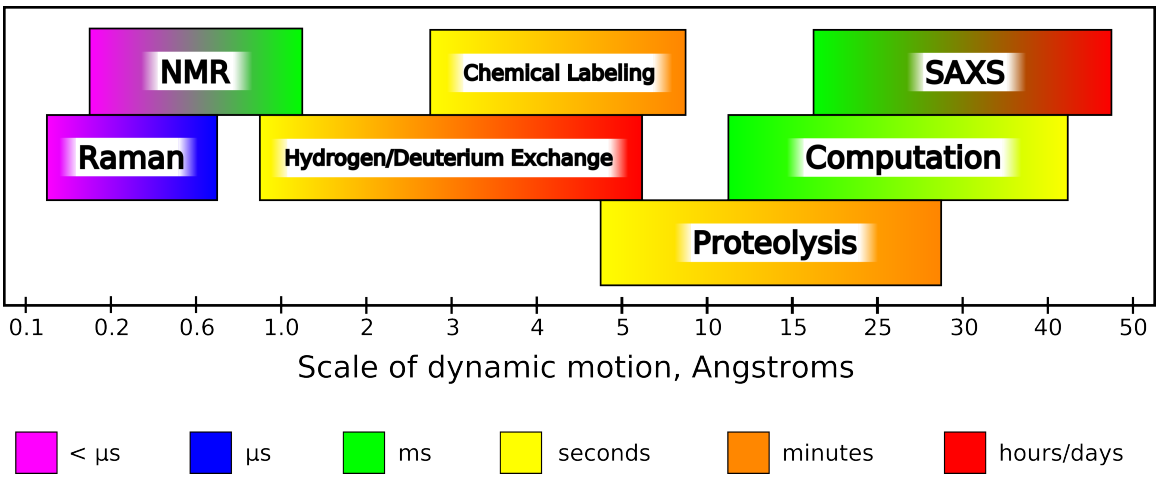


Figure 1.2: Utility of various tools for study of viral dynamics. The structural motion of protein complexes can take place on a full range of time scales (nanoseconds to hours) and with varying amplitude (static to nanometer or micron). Every technique has a slightly different utility for measuring a particular timescale and amplitude, although in many cases the limitations are imposed by current state-of-the-art in instrumentation, and as such tend to be extended with time. The horizontal axis indicates the scale of the dynamic motion that can be probed by each technique in nanometers. Some methods, such as Raman, may be combined with other tools such as HX to change their usefulness. Each method also measures a particular rate of dynamics (see color key). The "computation" marker indicated here refers to more coarse-grained simulations. All-atom simulations are currently limited to short timescales, although they provide information for the full range of amplitude that can be sampled in that time.

Table 1.1: Benefits and drawbacks of techniques to study viral dynamics.

Method	Information	Benefits	Drawbacks
NMR	Structure. Detection of HX. Time scale and range of motion data.	Can be highly specific. Bias towards dynamic regions.	Peak broadening with particle size. Difficulty assigning residues. Requires concentrated samples
Raman	Secondary structural elements. Detection of HX.	Uses dilute protein. Can probe low-frequency motions.	Difficult to assign signal location. Low signal/noise.
FRET	Distance constraints between groups.	Highly specific. Accurate distance measurements. Fluorescence lifetime kinetics.	May require site-directed mutations. Signal assignment is difficult.
HX	Backbone dynamics at resolution of several residues.	High spatial resolution. Time scale data is possible. Native proteins and protein mixtures, low concentrations.	Quantitative experiments are technically difficult. Difficult to assign sites for protein isoforms of varying length.
SAXS	Low-resolution density measurements. Real-time data.	Native proteins. Approximate size and shape. Rapid data collection.	Requires specialized instrumentation. Heterogeneous mixtures are a problem.
SANS	Very low-resolution density measurements. HX detection. RNA/protein/lipid differentiation.	Specificity to different components of virion. Low resolution structures.	Requires very specialized instrumentation. Mixtures of complexes are a problems.
Chemical reactivity	Local chemical environment.	Highly specific. Many probes available.	Limited to availability of reactive side chains.
Computation	Limited only by computational resources.	Avoids physical limitations of experimental setup. Field with exponential growth.	Size of viral particles pushes the limits of computational capacity. Requires known structure. Results quality limited by initial assumptions.
Proteolysis	Backbone dynamics at resolution of several residues.	Provides detailed kinetics. Probes larger motions. Does not require known structure. Tolerant of solution conditions.	Protein sequence limits possible enzymes. Simultaneous cleavages complicate kinetic interpretation.

NMR

As discussed previously, NMR has been instrumental in showing that regions of viral capsids have dynamic character in solution [28]. The primary advantage of NMR over other analysis methods is the resolution and content of the data: it is theoretically possible to determine the position and motion of the protein backbone with single-residue precision. However, limits in current instrumentation make this impractical for intact viral capsids, due in part to slow tumbling in solution which dramatically speeds relaxation, crowded spectra because of the large number of atoms, and the necessity for isotope-enriched protein concentrations in the millimolar range. NMR does have an advantage in that it selectively detects dynamic regions. Because of the lack of signal from well-ordered regions in a slow-tumbling capsid, any regions with a large degree of localized motion generate a very unique signature that immediately stands out from the background. With successful assignment, the signals from dynamic regions can be tracked in response to solution conditions, maturation state, or other perturbations. The early experiments with CCMV used this approach to examine the dynamics at the N-terminus which were quenched in the presence of packaged RNA. More recently, the latest advances in NMR technology have been applied to similar problems to detect dynamic regions in the course of HK97 maturation [36].

Icosahedral capsids such as CCMV or HK97 continue to present technical challenges to NMR analysis, but another category of viral capsid is much more amenable to such experiments. Filamentous capsids, which are composed of a very large copy number of fairly small monomeric subunits, can be spatially aligned along their axis, either via magnetic field or with solid-state crystals. Pulsing schemes designed to take advantage of the resultant dipole-dipole interaction can generate high-resolution structures, without the need for particle tumbling and regardless of the large par-

ticle size [37]. One of the interesting observations from such work is the extreme lack of dynamics in some filamentous capsids: in the case of the Pf1 bacteriophage, order parameters for alpha-carbons approached the upper limit of 1.0, indicating a very static structure. For comparison, crystalline glycine needs to be cooled to -45°C to reach such a limit. However, despite the overall high degree of order, some of the side-chains displayed rapid mobility. Residues deep on the interior of the virion were highly dynamic, despite being in proximity to the packaged DNA. This finding contrasts opposite observations for many icosahedral viruses, in which the presence of nucleic acids decrease the dynamics of residues in the local proximity, and sometimes generally throughout the capsid [38, 39]. As with all experiments for detecting dynamic motion, NMR spectroscopy imposes a particular set of criteria for interpreting the results. The rate dynamic motion required to detect a NMR peak varies depending on the protein and the exact experiment, but in general it involves motion in the microsecond range or faster. While structural transitions that complete in milliseconds could be considered rapid in the context of maturation and large scale equilibrium associated motions, they would be on the slow side for most NMR experiments.

Raman

Raman spectroscopy provides several key features that are invaluable to the study of viral capsids in solution. By probing the vibrations of specific bonds, Raman reports on the local environment for each particular signal. Unlike NMR, Raman is far more tolerant of dilute protein solutions, and it does not suffer from the scattering effects of large assemblies as does circular dichroism. Raman spectra can be predicted for specific secondary structure elements allowing approximate assignment of observed spectra to the corresponding helices or sheets, and though the single-

dimension spectra can get very crowded with overlapping signals, digital difference comparisons between two conditions can be used to isolate unique signals. Side-chains can also generate unique peaks, especially for aromatic amino acids and cysteine.

As a specific probe of secondary structure and local environment, Raman spectroscopy has been used extensively to study the effects of capsid assembly and maturation events [40, 41, 42]. For the P22 capsid, maturation from the procapsid to the capsid form produces a substantial number of changes in the side-chain signals without altering the distribution of secondary structure [41]. This suggests that the maturation involved quaternary translation of relatively intact domains in the capsid expansion. Due to its sensitivity to bond vibration, Raman is also sensitive to hydrogen-deuterium exchange, making it a very specific detection system for time-resolved proton exchange reactions, discussed below. As with NMR, direct Raman measurements depend on the sum population in solution, and the timescale of Raman dynamics is very fast (on the order of bond motion), which necessitates the use of a coupled kinetic probe such as deuterium exchange to detect slower motions.

Small Angle X-ray Scattering (SAXS)

Both Raman and NMR spectroscopy can be highly specific for certain regions in a virus capsid and both can give information on the local environment of the signal in question. Unfortunately, it can also be very time-consuming to collect enough spectra to gain sufficient signal-to-noise. In contrast to these methods, SAXS represents an orthogonal approach. The solution-phase complement to X-ray diffraction, SAXS measures the distribution of X-rays scattered at low angles ($< 10°$) from the incident beam. This provides information on the distribution of density within the sample directly from solution. The resulting reconstruction is low resolution ($\sim$25-100Å), but can be collected in milliseconds. This rapid data collection allows SAXS to be

used in a time-dependent manner to study maturation events and capture transient intermediate structures [4]. This ability was critical in the study of HK97 expansion between prohead II and head I: data collected at one minute intervals demonstrated a biphasic transition with two isosbestic points on an intensity/resolution plot [8]. At the current time the resolution limits for SAXS restrict its ability to detect smaller localized and equilibrium dynamics, but continuing improvements to synchrotron light sources are expanding the scope of this technique. Smaller scale light sources with high brilliance are also now available for local installation.

Small Angle Neutron Scattering (SANS)

The general application of small angle scattering can also be used with a collimated neutron beam, known as small angle neutron scattering (SANS). In addition to being less destructive than SAXS, this technique is of particular interest to the study of viral particles due to the sensitivity of SANS to hydrogen, which is nearly silent in SAXS experiments. Hydrogen nuclei interact efficiently with the incident neutrons, giving the solvent a substantial and distinctive scattering length density, which can be controlled according to the ratio of H_2O to D_2O in the solvent. Protein in the sample also scatters, but with a different length density as a function of the average elemental composition. Likewise, nucleic acids and lipids both have unique scattering signatures. Because these scattering length densities have different responses to the presence of deuterium, by performing measurements in a series of H_2O/D_2O ratios, the individual contributions from proteins, nucleic acids, and lipids can be discerned. This technique is known as contrast variation, and it has been successfully applied to solution-phase structural studies of the icosahedral MS2 bacteriophage [43]. Kuzmanovic et al found that in solution, the RNA is compacted within a radius of 83 Å, despite the capsid itself extended to a radius of 115 Å: the area within the inner radius of capsid was

composed of 81% water. Separate studies were performed to investigate the effects of the A protein, a 44 kDa RNA-binding protein which is present as only a single copy within the capsid shell [44]. Recombinant viruses without the A protein formed thicker shells (∼34 Å vs 24 Å) compared to normal MS2 viruses, but the outer diameter of the viruses and capsids remained nearly the same. The values obtained from these solution-phase measurements are subtly different from measurements of MS2 crystals structures: the crystal structures retain the same protein shell thickness, but with a slightly reduced overall diameter, presumably due to crystallization conditions or packing forces. More importantly, when comparing crystal structures obtained with the A protein to those without, no difference can be seen, which illustrates the value of the solution-phase studies, despite the reduced resolution of such methods.

Computation

By far the most recent development in the study of capsid dynamics is the application of computational simulations and network and graph theory. Only recently has computing hardware advanced to the point where meaningful results can be obtained for systems of any magnitude: in 2006 an all-atom simulation of Satellite Tobacco Mosaic Virus (STMV) for 13 ns represented the first such work of its kind [45]. The results of that simulation indicated that the RNA core could be self-stable without the protein capsid, which matched previous experimental results. Likewise, the calculations indicated that the empty capsid would be unstable, whereas the RNA-filled would not, in agreement with experimental observations. Computational developments have proceeded rapidly since then: less than a year after that initial virus simulation, an all-atom simulation of the 70S ribosome was performed with approximately double the number of atoms [46]. Despite technological progress, the short timescales and limited size of all-atom simulations remain a bottleneck for prac-

tical simulations of viral systems. To circumvent this limitation, one approach is to reduce the complexity of the system in silico, a technique known as coarse-graining. There are several schemes to achieve this goal: the protein structure and its network of connections can be represented by a series of point masses. In the most common form of simulation, known as normal mode analysis, the connections are simulated as a series of springs, and the lowest-mode harmonic distortions represent the large-amplitude, low-frequency motions possible in the capsid structure. These harmonic modes have been compared to the experimentally-observed quaternary dynamics of HK97, CCMV, and other viruses with relatively good agreement [47, 48]. Normal mode analysis (NMA) has also been used to distort high-resolution crystal structures to fit into a lower density cryoEM map, while maintaining the lowest-energy conformation possible [49]. Computational tools like NMA have invaluable use as a modeling framework upon which to put experimental results. One of the most intriguing developments is the combination of coarse-grained simulations and graph theory to understand unique spectroscopic features and shared characteristics of highly symmetric viral capsids [50, 51]. Elastic wave theory and an amorphous isotropic bond polarizability model have been used to predict the extreme low-frequency Raman spectra for M13 [52]. The results predicted an axial distortion that was found to be in good agreement with experimentally-collected data. Another study used group theory to analyze the Caspar-Klug viruses from the VIPER database with weighted subunit interactions [53]. Surprisingly, a low-energy plateau that extended through 24 modes before dramatically increasing was common across capsid architectures. One explanation is that the plateau provides a theoretical basis for the metastable, dynamic structures that have been repeatedly observed in a variety of experimental studies.

Labeling Experiments

Spectroscopic approaches used to study viruses in solution have the advantage of being relatively non-destructive, but regardless of the precise technique there are recurring problems of sensitivity, signal assignment, and a lack of control over the scale of dynamics being measured. Alternatively, the high-energy dynamic state of a viral capsid can be differentiated from the ground state by means of a permanent covalent modification. The general scheme involves the careful labeling or cleavage of the capsid under kinetically-controlled conditions. After the labeling phase has been completed, an analysis step (or multiple steps) is employed to quantitate the course of the reaction. The three variants of this approach are discussed in detail below.

Chemical Reactivity: Chemical cross-linking is the process of reacting a small chemical probe to an engineered or native reactive group on the body of a viral capsid. In the case of a cross-linking reagent with two reactive termini, the distance between two amino acid side chains can be inferred from the length of the linker region. A selection of cross-linking reagents ranging in size from 8 to 25 Å, with a variety of reactive groups, are commercially available. Key uses of cross-linking are to map subunit interfaces when the structure in not known and to investigate solution dynamics when structural models are available. Analysis normally involves the use of proteolysis and mass spectrometry to identify the specific residues involved. The P22 and HIV-1 CA systems have been probed in this manner. With HIV-1 the linker was only found bridged in a single position at Lys70-Lys182, indicating that the N domain of one subunit is in close contact with the C domain of another [54]. For the P22 system, the use of different length cross-linking reagents allowed inter-subunit distances between two lysine residues (Lys183-Lys183) to be determined in relation

to a known intra-subunit lysine distance (Lys175-Lys183) [55]. Chemical labeling experiments which measure the rate of site-directed labeling reactions are another way to probe capsid dynamics. In place of distance constraints within the structure, this method provides information on the chemical environment and solvent exposure at a particular site. Using this technique, Bothner, Taylor, and others probed the reactivity of the specific capsid regions including the internal gamma-peptides of tetra and nodaviruses and were able to show distinct differences between T=4 procapsids and capsids and T=4 and T=3 mature capsids, consistent with other data indicating that the solution properties are different [20, 21]. Similar techniques were used to show that FHV virus like particles (containing heterologous cellular RNA) and wild type particles have dramatically different dynamic properties in solution even though they are crystallographically identical [38]. In addition to the rate of chemical modification, the maximum stoichiometry of labeling also provides information regarding the structure of capsids in solution. The maturation of the NωV capsid is a very rapid and dramatic contraction, with a decrease in diameter of 16% within 100 ms [4]. This conformational change involves a transition from a highly fenestrated structure to a much more tightly packed capsid. Labeling experiments on the two forms of the protein with fluorescein derivatives showed that the maximum extent of labeling was far greater for the expanded procapsid than the capsid: up to $\sim$800 lysine or cysteine (depending on reactive chemistry used) were labeled on the procapsid [20]. In comparison, the capsid form of the protein could only react with $\sim$100 dye molecules per capsid.

Hydrogen Exchange: A powerful technique for investigating protein dynamics is hydrogen-deuterium exchange (HX). This approach makes use of the varying rates

of exchange for amide or side-chain protons with solvent, depending on their local environment and participation in hydrogen bonding. Upon dilution or buffer exchange of a protein (containing normal environmental ratios of $^{1}H/^{2}H$) into $^{2}H_2O$ (D_2O), the protein protons will begin to equilibrate with their surroundings. In practice, the rate of exchange varies from nearly instantaneous to effectively zero over the course of months, depending on the stability and solvent exposure at a specific site. Detection of the exchange can be accomplished via spectroscopic methods such as NMR or Raman, and this approach has been used to study capsid subunit interactions, folding, and conformational changes [40, 41, 56]. However, recent application of HX has focused heavily on mass spectrometry to detect the mass shift of deuterium incorporation. After exchange, the proteins are cleaved with pepsin to generate peptides so that changes can be assigned to a localized area. The number of sites that have exchanged is measured as the shift in mass difference before and after exposure to D_2O. This approach requires small quantities of material and is less subject to the size scaling issues discussed above for NMR. The only real limitation to the study of large complexes using HX-MS is that as the number of different proteins in a sample increases, it becomes more difficult to assign exchange data to a specific site. Virus capsids, with their use of multiple copies of the same protein abrogate this issue, at least with respect to low T-number capsids. However, HX is an intrinsically reversible process, and the phenomenon of back-exchange (^{2}H to ^{1}H) during proteolysis, separation, and detection can complicate data analysis: in some cases more than 70% of the deuterons are lost in this second phase of analysis. Back-exchange can be minimized by controlling pH, reducing temperature, and minimizing the time of analysis, but these complications make intact-protein HX-MS an attractive option due to the lack of post-labeling processing for experiments which do not necessitate specific assignment of exchanging regions. Due to the flexibility with which HX experiments can be con-

ducted, there is a wealth of excellent studies using this approach, many of which have contributed greatly to our understanding of viral capsid structures. Our discussion of this topic is only a brief selection to highlight some of the features of this experimental method: for more examples of HX applied to viruses see [57, 58, 58, 59, 60].

As opposed to spectroscopic studies, one of the key features of HX-MS is the ability to work at low concentrations and in the presence of a heterogeneous mixture of proteins. This tolerance to mixtures of proteins was used to probe the dynamics of 29 scaffolding protein gp7 in the free and procapsid-bound forms [61]. A combination of intact protein and proteolyzed fragments were analyzed, allowing characterization of the H-L-H domain which interacts with the coat protein. Furthermore, HX-MS provided confirmation for the multiple concentric layers of gp7 implied by cryo-EM reconstructions: the cooperative HX reached a saturation limit in amino acids 20-31. This implies that within the procapsid, gp7 exists in bound and unbound forms, and only the bound population was destabilized.

Hydrogen-deuterium exchange experiments conducted on HRV14 provided a good example of the application of HX data to a system with a known structure, placing dynamics measurements directly into the context of classic static structures [60]. As with the studies on ϕ29 gp7, post-exchange proteolysis was used to improve the resolution of assignments: $\sim$80% of the protein sequence was measured via 90 unique peptides, and rates of exchange were tracked for all peptides. To achieve this level of coverage only 20 pmol of protein per sample was required, which is a benefit of the sensitivity of HX-MS. In contrast to the exchange curves for VP1, VP2, and VP3, which had varying regions of protection and dynamic motion, VP4 was almost uniformly dynamic and highly exchanged after only 2.5 minutes. This data supports other experimental data such as limited proteolysis, which indicates that VP4 is a very dynamic region [27]. However, the dynamics of VP4 as indicated by HX-MS

have a subtle nuance: because the threshold for deprotection in HX is so low, it is impossible to evaluate the full limit of the dynamic motion. Limited proteolysis of the same region provides evidence that the dynamic behavior of VP4 is not just limited to solvent exposure on the interior of the capsid, but rather that it undergoes substantial deformations.

Kinetic Hydrolysis: Quantitative measurements of protein dynamics are critical if we are to understand the thermodynamic nature of conformational changes, how they are coordinated within the supramolecular structure of a capsid, and how they relate to the lifecycle of a particular virus. To date, only a single example exists in which equilibrium and rates of conversion have been determined for a megadalton complex. The hepatitis B virus capsid has been instrumental as a model system to understand viral capsid assembly and has served as the foundation for substantial theoretical models of supramolecular complex assembly. A member of Hepadnaviridae, the core protein forms an unusual dimer composed almost entirely of alpha-helices. The dimer cannot be separated without denaturing conditions, and 120 copies assemble to form the T=4 capsid. Due to a well-established strong hysteresis to disassembly, intact capsids can be studied under the same conditions as the dimer. This feature was exploited recently by researchers in the Bothner lab to quantitatively measure dynamics of dimer and capsid forms using kinetically-controlled proteolysis (kinetic hydrolysis) [62]. Because the assembled and unassembled protein can be studied under the same solution conditions, dynamics associated with protein dimers can be separated from emergent dynamic properties resulting from assembly of the capsid. Using SDS-PAGE to measure rates of intact protein degradation and peptide mass mapping to identify the sites of hydrolysis, it was shown that the C-terminal region of

the capsid protein Cp149 was dynamic in both forms. By performing assays across a range of protease concentrations, kinetic curves of HBV digestion were obtained which detailed both the rate of exposure from the closed to the open conformation, as well as the equilibrium between those states in solution. This approach works because while the specificity of a protease is a function of the protease, the rate of hydrolysis at a particular site is highly dependent upon the accessibility by the protease, as mediated by the local backbone dynamics. Based on a 2-state model using the structure from X-ray crystallography as the low energy state, the transition has a lifetime of approximately two seconds and $\sim$3 subunits per capsid are in the open (or high energy) conformation at any time. By docking enzymes to the HBV surface, it was estimated that a translocation of > 13 angstoms from the location in the crystal structure is required to reach the open, cleavage-accessible conformation [62]. A surprise finding was that the protein in the dimer and capsid forms was thermodynamically distinct (Figure 1.3). While no substantial differences were determined for opening rate, the equilibrium between open and closed forms had an opposite temperature dependence when comparing dimer and capsid. The dynamic site is also in close proximity to the binding location of HAP compounds that have demonstrated antiviral activity [63]. Together these results have interesting implications for the role of dynamics in HBV assembly and the specific targeting of dynamic regions with antiviral agents. The surprising behavior of the HBV system illustrates the power of solution-phase measurements of dynamic protein motion. Future applications of this technique to other systems will undoubtedly illuminate trends for viral capsids in general and serve to better connect our understanding of structure and function.

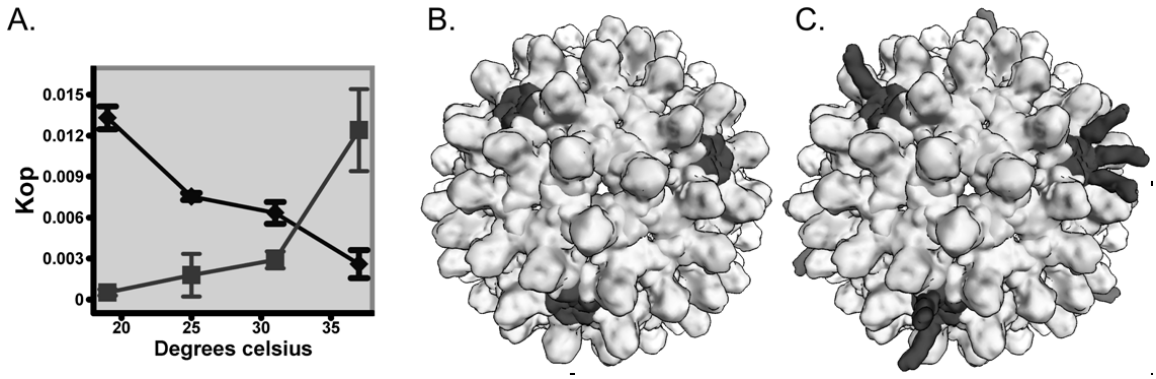


Figure 1.3: Kinetic hydrolysis of the HBV capsid protein Cp149. Controlled proteolysis revealed structural motion at the C-terminus, which clusters around the fivefold and quasi-sixfold centers of symmetry on the capsid. A.) Equilibrium (open vs closed) constants as a function of temperature: the decrease in opening for dimer (diamonds) at high temperatures rules out capsid formation and suggests an entropic stabilization of an assembly-active state. The capsid data (squares) shows an increase in dynamics with temperature. B.) The capsid in the "closed" state, suggested from the crystal structure. The C-terminal helix is shown in black. C.) Artist rendition of a possible "open" state. Substantial unfolding is needed to expose the observed cleavage site to protease. It is currently unknown if the centers of symmetry unfold cooperatively or not.

Summary

Biochemical and biophysical investigations of virus particles in solution are critical to understanding their functional properties. The array of functional demands that are placed on capsid proteins requires numerous structural calisthenics to be performed, often in the context of a delicate balance between assembly and disassembly. Together with structural models, information on the location and extent of capsid dynamics provides a basis for linking structure to function with greater detail than either approach alone. A number of significant questions remain to be addressed including, the role of mutations that alter dynamics on viral fitness and whether dynamics is a symmetric or asymmetric property. The later point has implications for receptor binding, cell entry, and genome release all of which remain poorly characterized in non-enveloped viruses. The significance of this information goes beyond the basic biology of viruses. Dynamic regions are a relatively untapped target for antiviral therapy and viruses are excellent model systems for studying allostery and dynamics in supramolecular complexes. Viruses are currently being used and developed as bio-inspired nanomaterials, with applications from gene delivery to nanowires. Scientists are now seeking next generation nanomaterials that can actively respond to various stimuli and a thorough understanding of the dynamic properties of capsids will be important.

KINETIC HYDROLYSIS OF CP149: LARGE-SCALE DYNAMICS AS A FUNCTION OF TEMPERATURE

HBV as a Model Capsid

Of all the icosahedral viral particles, perhaps none have a greater breadth of available assembly data than the hepatitis B viral capsid. However, the exact details of its structural transitions and motions have remained elusive. Structural dynamics have been shown to play critical roles in viral capsid function, and to obtain a better understanding of these factors for the HBV capsid we have applied several kinetic-based tools to probe the structural motions.

There are several reasons why the HBV capsid is ideal for such a study. First, HBV is extremely interesting in its own right. Hepatitis B is a major human pathogen, despite the availability of effective vaccines. More than 350 million people have chronic HBV infections, which leads to hepatocellular carcinoma and cirrhosis. The capsid itself forms both T=3 and T=4 icosahedral capsids, with 120 copies of a homodimer in the T=4 and 90 copies in the T=3 form. The T=4 icosahedral particle is approximately 350 angstroms in diameter [64]. The homodimer consists of two helical monomers; the dimer unit is extremely stable and cannot be separated without denaturing conditions. The monomer protein (HBcAg) has 183 amino acids and can be partitioned into an assembly domain, from residues 1-149, and an additional 34 residues composing an RNA-binding domain at the C-terminus. The native virus also produces a variant of the protein (HBeAg) which has 10 additional amino acids on the N-terminus, and terminates at the end of the capsid assembly domain, residue 149. The assembly domain (Cp149) is sufficient form a capsid which contains all structural features observed to date [65], and such capsids are indistinguishable from

native virions [66]. The virion normally packages RNA, which is reverse-transcripted into dsDNA only after the capsid is assembled. Capsids can be produced in *Escherichia coli* and do not require nucleic acids or scaffolding proteins for formation: capsid assembly proceeds spontaneously upon translation [67]. However, RNA does enhance the capsid assembly rate [68, 69].

Unlike some other capsids, HBV Cp149 can be disassembled and reassembled with sufficient yields to allow kinetic characterization of the assembly process. Combined with the ability to produce high-purity capsids from a recombinant expression system, this has made the HBV capsid a model system for the study of capsid assembly. The assembly process has been monitored in response to pH, temperature, ionic strength, dimer concentration, and assembly-affecting small molecule compounds [70, 71, 72, 73, 74, 75]. In general, capsid assembly is driven by low pH, elevated temperatures, and increased ionic strength. To date there are no known specific negative assembly effectors, although strong assembly activators can have a negative effect on overall capsid formation (discussed below).

The last characteristic which makes HBV/Cp149 a good target for biophysical characterization is its hysteresis to disassembly. The network of low-energy interactions between dimer units forms a highly multivalent network with an overall strong assembly equilibrium: the equilibrium constant is

$$K_{\text{capsid}} = [\text{Cp149}_\text{c}]/[\text{Cp149}_2]^{120} \tag{2.1}$$

where Cp149_c and Cp149_2 represent molar units of capsid and dimer, respectively. This model accurately predicts a pseudo-critical concentration of dimer which is necessary to achieve any level of capsid formation. At the same time, there is a kinetic trap to removing the first dimer from an intact capsid, which causes the hysteresis effect. By carefully choosing solvent conditions and dimer concentration below the

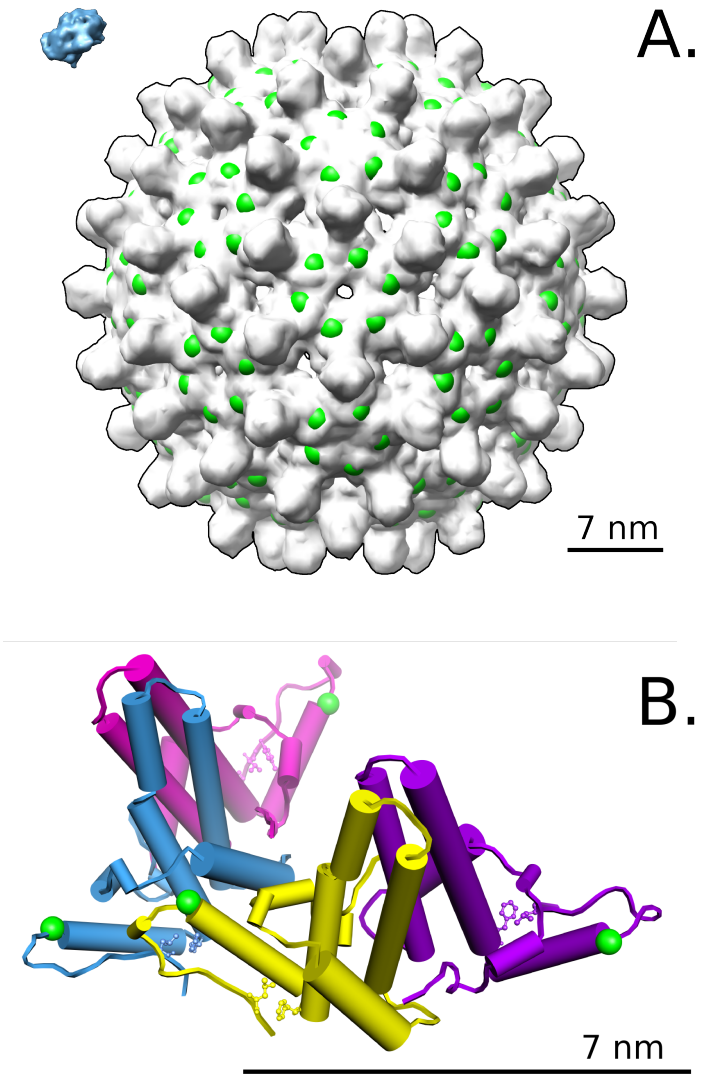


Figure 2.1: Structures of Cp149 capsid and dimer. A) The Cp149 capsid, based upon the crystal structure. The primary observed cleavage location is shown as green spheres, and is not fully solvent-exposed. The other cleavage site identified in effector studies is located near the tops of the towers. The structure of typsin is shown for comparison. B) The structure of an individual asymmetric unit. Dimer-dimer contact in the capsid is localized around the C-terminal helix. The 127/128 cleavage site is shown with green spheres: there are four possible symmetry-separated variants of this cleavage site: three similar variants from the quasi-sixfold axis and another at the 5-fold.

pseudo-critical point, but within the concentration limits for hysteresis, solutions of dimer can be maintained for extended times without capsid formation, and likewise capsids will not dissociate into dimers. This permits experiments to be conducted with perfectly matched solvent conditions, giving directly comparable results between dimer and capsid.

Structure and Assembly Studies of Cp149

The structure of the Cp149 T=4 capsid has been available for several years now, and it is supported by a host of cryo-EM structures of native capsids, sequence mutants, covalently-labeled capsids, and capsids with bound antibodies [76, 66, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86]. Based on the crystal structures and low-resolution density data, the overall structure of the dimer subunit is maintained without significant structural variability, and consists of five discrete helices. The fourth helix, between residues 79 and 110, contains kink at approximately position 92 and could be considered two separate helices. Helices 3 and 4 together protrude from the capsid surface in a helix-turn-helix motif and form a vertical bundle (sometimes termed the tower) with the corresponding helices 3 and 4 from the other half of the homodimer. Helix 1 is positioned between the tower and helix 5, which angles away from the body of the dimer. At the C-terminal apex of helix 5 the peptide backbone has a turn, and in the crystal structure the remaining 22 amino acids form a loop which folds under helix 5 and is stabilized in that position via hydrophobic contacts. Additional hydrophobic contacts around helix 5 form the basis for dimer-dimer interactions on the surface of the capsid. This area of dimer-dimer contact is relatively small, and the capsid overall is highly fenestrated [87]: the intra-dimer contacts encircle pores at the 5-fold a quasi 6-fold (true 2-fold) axes.

To date, the exact mechanism of capsid assembly is unknown. This question has been partially addressed by studying the thermodynamics and kinetics of capsid assembly, which points to a kinetic-controlled pathway and a network of weak interactions to form, and later maintain, the correct capsid conformation. Studies conducted at elevated pH, which slows overall capsid formation, found that there was a third-order dependence for assembly progression, which points to a trimeric nucleus of dimers as an early rate-limiting step in the capsid assembly pathway, and the rate of capsid assembly overall followed second-order behavior indicative of rapid capsid "extension" [88]. On the basis of equilibrium calculations for capsid assembly at varying temperatures, the interactions which drive capsid extension are very weak, on the scale of -13 kJ M^{-1}, but the very large number of dimer-dimer contacts (240) maintains the overall capsid stability [70]. Mathematically, the combination of the number of contacts, and the multivalency for each dimer, predicts that removal of dimers from an intact capsid is a kinetically-unfavorable process [89]. In order to remove a subunit, multiple low-energy contacts must be broken simultaneously, and once released, there is a very high localized concentration of subunit which ran reform the intact capsid. This produces a hysteresis between assembly and disassembly, and such effects have been experimentally observed, as seen in figure 2.2. Assembly reactions do not proceed strongly to completion until dimer concentration is above 20 μM, but once capsids have formed they remain stable for days with concentrations a low as 5 μM equivalent dimer.

The capsid assembly experiments described thus far have used ionic strength, temperature, and pH to regulate capsid assembly: these factors have substantial impact on the thermodynamic parameters for assembly. However, small molecules have been observed to have an allosteric effect upon assembly, with the capacity to very rapidly cause assembly reactions to reach completion [71, 74]. A series of mono-

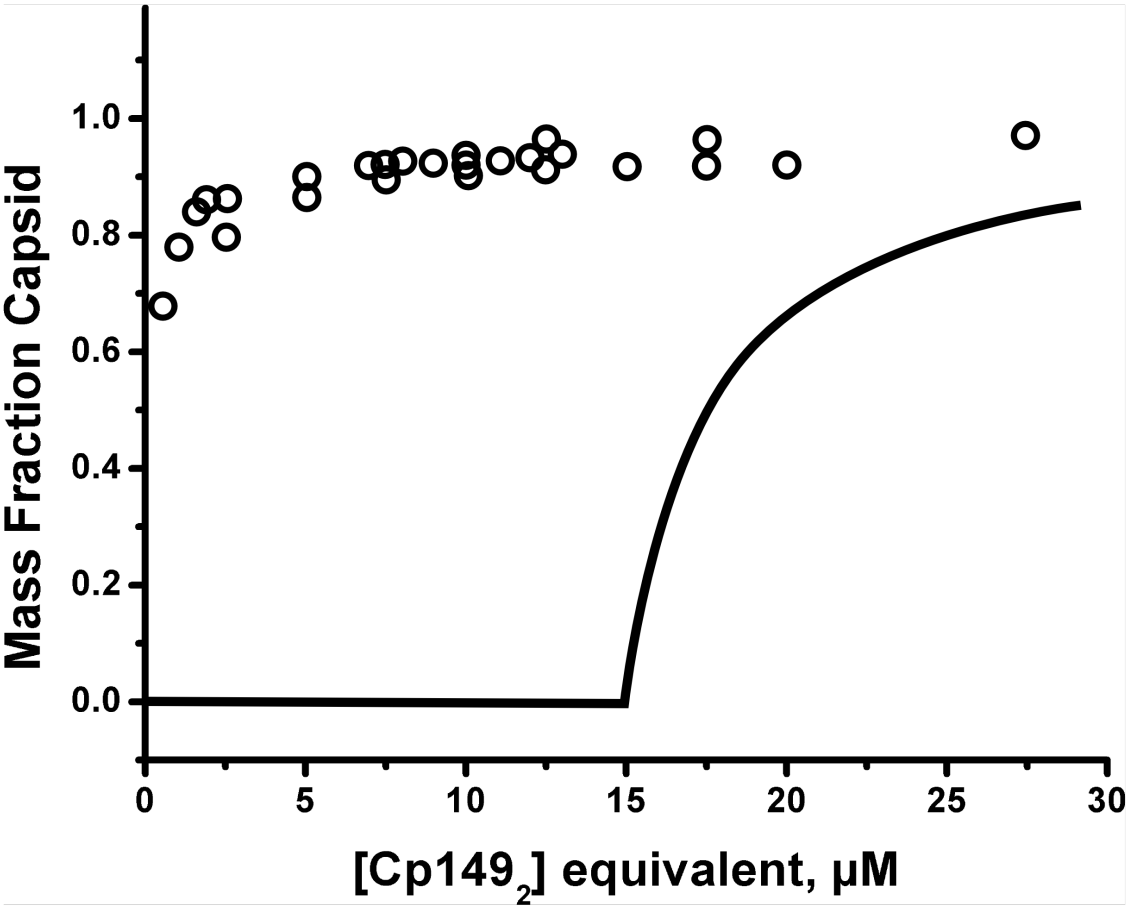


Figure 2.2: Assembly and disassembly displayed marked hysteresis. Purified capsids were diluted to the indicated concentration, left for 5 days at 21 °C, then assayed via SEC. The assembly isotherm for these conditions is shown as the solid line, while SEC assay results are shown as circles. Adapted from [89].

and di-valent cations have been tested for efficacy as regulators of capsid assembly, and all showed some effect, but massive quantities (millimolar to molar) were required to drive assembly, and even then the yields were low. In contrast, zinc produces very strong assembly at micromolar concentrations, much lower than would be predicted for strictly bulk charge-based interactions [71]. Zinc is also unique in that it has the capacity to prevent efficient capsid formation when present in high concentrations: rather than intact capsids, a heterogeneous population of capsid intermediates is formed. This effect is consistent with the nucleation model: if the solution can be over-populated with nucleation sites, all free dimer can be consumed before capsids can be completely formed. One possible mechanism for this phenomenon is to assume that zinc somehow de-regulates the initial rate-limiting step of nucleation formation, which allows excessively large flux into the assembly pathway. To date there is no confirmed binding site for zinc interaction with the Cp149 dimer, although one possible binding site involves E8, H51, and H104 from one monomer, with H47 from the other monomer. The H104 residue is of some interest due suggestion via kinetic hydrolysis experiments (chapter 3) that dynamics in the 83-127 region may be greatly increased in the presence of zinc.

Zinc is not the only molecule known to have allosteric-type effects upon Cp149 capsid assembly. Originally discovered by Bayer [90], a class of molecules known as heteroaryldihydropyrimidines (HAP) have been shown to reduce viral load in vivo, and vivo assays of assembly have shown them to be potent assembly activators [73, 74]. These compounds have tight binding affinity to Cp149 (low micromolar) and can rapidly drive assembly into aberrant structures. A moderate-resolution crystal structure was obtained for Cp149 in the presence of HAP-1, which showed it bound in the dimer-dimer hydrophobic contact region [76]. The overall capsid structure was relatively unchanged at the protein level, but the inclusion of HAP into the dimer-

dimer contact region acted as a wedge, causing a quaternary rotation of the dimer. This distorted the 5-fold and 6-fold centers of symmetry by making the C-terminal region of the monomer protrude outwards at the 5-fold site and partially flattening the 6-fold. Interestingly, the observed HAP binding site was at the 6-fold, but did not display full occupancy: one of the three possible locations was free of HAP, a second had only partial density, and the third was the region of highest density. In response to the torsion induced on the dimer structure, the tower of the dimer was distorted slightly sideways from the normal orientation.

Due to the resolution of the density set, it is impossible to definitively place HAP-1 into the region of observed density. However, the two possible orientations of HAP-1 in the binding pocket both place the methyl group, coming off the 6 position of the central dihydropyrimidine ring, emerging from the binding pocket of one dimer and protruding into the dimer-dimer contact region. To investigate the effects of the functional group of the HAP molecule upon capsid assembly, a series of modified compounds were synthesized (figure 2.3) [74]. These compounds each produced unique results, with the specific type of HAP derivative tending to produce different forms of aberrant assembled complexes (figure 2.4). However, since the application of HAPs as antiviral agents depends on their ability to interfere with a kinetic-limited process, rather than their affinity or ability to induce particular aberrant structures, additional research is needed to fully characterize the interactions and mechanism of HAP binding with Cp149.

In contrast to the wealth of knowledge available for Cp149 capsid structure, very little is known about the structure of the dimer in solution. For many years the only data available was inferred from antibody binding studies, which implied that certain sequences on the protein have differential recognition depending on assembly state [91]. More recently, NMR studies have been completed on the Cp149 dimer

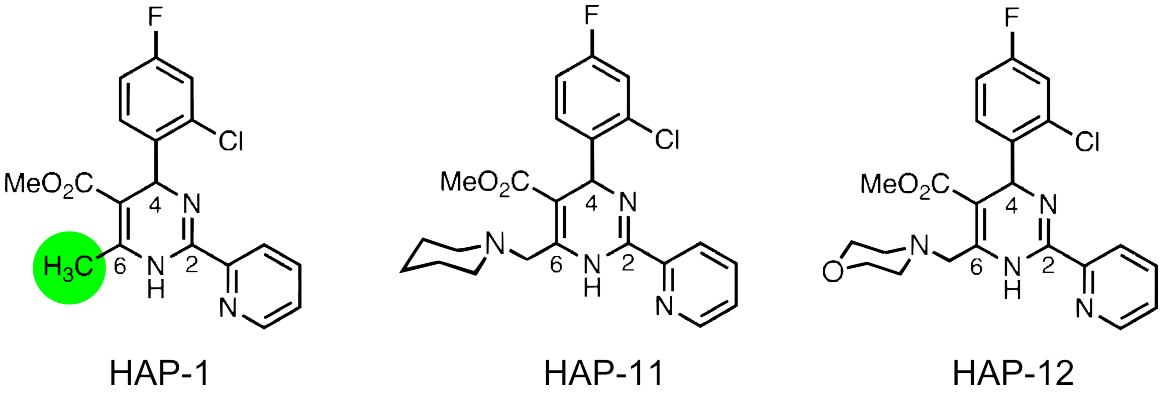


Figure 2.3: A family of HAP compounds affect HBV assembly. The basic structure of HAP (heteroaryldihydropyrimidines) consists of three aromatic rings: these rings bind into a hydrophobic pocket on the capsid. Varying substituents at position 6 cause different impacts on Cp149 assembly.

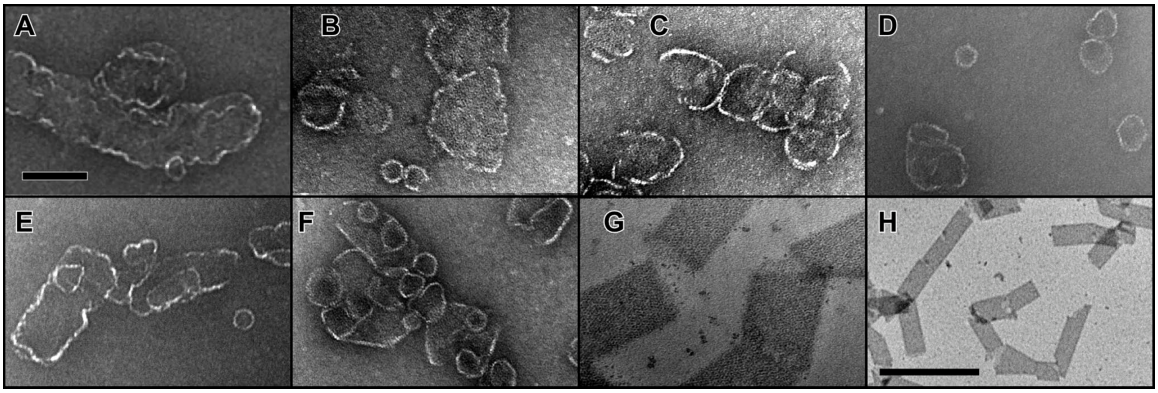


Figure 2.4: HAPs induce aberrant assembly structures. The exact form of the resultant complex depends upon the HAP variant and concentration. Assembly protocol involved 30 μM HAP , 20 μM Cp149 monomer, 37 °C, 24 hours: HAP13 (A), HAP11 (B0, HAP12 (C), HAP4 (D), HAP14 (E), HAP7 (F), HAP18 (G), and HAP18 (H). The scale bar is 100 nm. From [74]

.

in high pH conditions, which was sufficient to prevent assembly [65]. Using TROSY NMR, 117 of the 137 expected cross-peaks were detected, suggesting that a reasonable number of backbone locations remain disordered. Many of these unidentified peaks were mapped to the regions between helices, but the C-terminus following residue 127 was very poorly represented in the detected peaks. In spite of the incomplete dataset, the overall structure as detected with NMR was very similar to the capsid structure: a high degree of symmetry is suggested with minimal differences between the monomers.

A very different picture is presented by the crystal structure of Cp149-Y132F dimer, which is the first atomic-level structure obtained for the HBV dimer (A. Zlotnick, in preparation). The mutation at position 132 serves to remove a major contributor to buried hydrophobic contact area: the tyrosine account for up to 10% of the total contact in the region between dimers on the capsid surface, and the Y132F mutant is assembly-deficient as a result. This residue is peripheral to the main body of the dimer, and is not expected to induce any substantial alterations to the structure of the dimer itself. The overall Cp149-Y132F dimer structure is a close match to the Cp149 capsid structure, but only at the secondary structure level. Full alignment of the two structures is impossible because the dimer displays significant shifts of large portions of structure, the most prominent being a loss of symmetry in the tower, with one subunit twisting away from close contact at the apex. The C-terminal helix also cannot be closely aligned between the capsid and dimer structures: the dimer structure shows it extending away from the location found in the capsid conformation. Despite the difference between the dimer and capsid structures, there is also a region of very high similarity, known as the chassis. This region, composed of residues 1-9, 26-62, and 94-109, forms a stable core around which the variable regions rotate. The rotational motion observed in the dimer structure may be facilitated by

a series of glycine and proline residues, found at positions G10, G63, G94, G111, and P25. Glycines 63 and 94 are of particular interest, since they form the break in the tower helices and act as a fulcrum for the asymmetric tower rotation in one monomer.

Implications of Dynamics

Data regarding the structural activity and function of the HBV capsid derives from many sources: kinetic and thermodynamic assembly experiments, drug effectors, HBV lifecycle observations, epitope characteristics, and crystal structures. Taken as a whole, this data all points to one inescapable conclusion: the capsid protein is a unique and highly active structure, capable of interacting with its environment in a multitude of ways. Some observations seem intuitive, such as the way that HAP compounds bind in an intra-dimer interface and leverage the capsid subunits into a distorted conformation. Other data remains unexplained, such as the effects of NaCl upon driving capsid assembly, or the assembly activation provided by zinc, when the putative binding site is part of the static chassis from the dimer crystal structure. However, the sum of these observations, especially when combined with the recent dimer crystal structure, all point to allosteric regulation via significant structural alterations. Probing these structural dynamics is therefor a natural approach to achieve a more detailed and complete understanding of the hepatitis B capsid, and in the process develop tools for routine assay of the dynamics of other supramolecular complexes.

Experimental Techniques and Planning

Inherent Problems of the Cp149 System

The fact that Cp149 readily assembles and disassembles with relatively minor changes in environment is a very substantial impediment to conducting assays on dynamics. In practice, this is primarily a restriction on available ranges of pH and protein concentration. To prevent capsid disassembly, and to keep dimer from assembly, the pH must be approximately 7.5. Many fluorescent probes lose their fluorescence at that level, which rules them out for covalent labeling experiments. Cp149 also contains a minimal number of available lysines, which are the primary targets for such experiments. NMR can be a viable option, and has been successlly used to obtain data on the dimer, but it necessitates very high pH and cannot be used with the intact capsid. NMR would also not be amenable to low concentrations of protein or rapid data collection. By referring to Table 1.1, we can see that the most promising techniques are the ones which make use of proteolysis and mass spectrometry. Kinetics experiments with proteolysis share the same general features of covalent labeling as a probe of the local environment, but enzymes can be functional at nearly any pH. Specific activity of a protease drops off away from the optimal pH, but this can be compensated by increasing enzyme concentration. In contrast, increasing the concentration of a fluorescent label can cause inner filter effects and label precipitation. Proteases are also very diverse: there are over 34,000 different proteases in the MEROPS protease database [92], and a large number are commercially available. With the appropriate enzyme, every bond of the peptide backbone can be cleaved, which theoretically allows data to be collected at single-residue resolution. Mass spectrometry is an ideal detection and quantitation tool due to the speed of analysis (ranging from instantaneous to a few minutes per measurement), sensitivity, and the

easy of automation. Mass spectrometry is also the analysis method of choice for hydrogen-deuterium exchange experiments, which is discussed in chapter 4.

Detection Assays for Proteolysis

Protease activity can also be assayed in multiple ways. Analysis of proteolysis can be conducted via SDS-PAGE, which has a long history but has recently been advanced in quantitative form by Marqusee [93, 94]. Gel-based analysis has the advantage of being very simple in terms of technology and analysis, but obtaining high degrees of precision can be difficult. Proteolysis can also be measured with mass spectrometry, which has the advantages of providing assignments to the cleavage locations. This approach has been used to good effect by Bothner et al for semi-quantitative observations of HRV14 (chapter 1, section 1) [35]. Both of these approaches also have an attractive position with regards to technological method development. Quantitative limited proteolysis via SDS-PAGE is an established technique and is already widely accepted in the community, but application to date had been restricted to monomeric proteins and simple cleavage patterns and analysis. Likewise, mass spectrometry-based quantitation is also widely accepted, especially for small molecules, but it has not been applied in rigorous fashion to proteolysis reactions. Thus, the adaptation of these techniques to high-precision measurements of capsid proteolysis was deemed to be feasible, receive a minimum amount of reviewer skepticism, and have a high probability of generating valid and applicable results for the study of capsid dynamics. The chemistry department at Montana State University is also well-equipped to perform mass spectrometry-based quantitation due to the availability of multiple forms of MS instrumentation (MALDI-TOF, ESI-ion trap, and ESI-TOF) which have high access availability for custom experiments. The advantages and disadvantages of SDS-PAGE and MS quantitation are discussed in more detail in the following

sections, but it is important to note that they are very complimentary approaches for proteolysis measurements. SDS-PAGE can be used with minimal preparation and adaptation for any protein system, and easily tolerates protein mixtures, but analysis is a time-consuming process and lacks precision. Mass spectrometry requires careful planning for maximum effectiveness but is the more sophisticated overall technique. For these reasons proteolysis was selected as the approach for initial characterization of proteolysis.

Experimental Conditions

Since proteolysis can be used in nearly any solvent conditions, Cp149 can be characterized in buffers that provide the best biophysical relevance, as opposed to selecting experimental conditions solely based upon the necessities of the analysis. To select the ideal conditions for assays, the substantial body of literature regarding Cp149 assembly was reviewed. Relevant portions of this data is reviewed above in section 2. In brief, concentration of $Cp149_c$ were kept at or below 0.5 mg/mL, NaCl was maintained at 150 mM, and the solutions were buffered at pH 7.5. The protein concentration corresponds to approximately 15 μM concentration of dimer, which is below the pseudo-critical assembly concentration for the chosen level of ionic strength and pH. Using these conditions, experiments can be conducted between the temperatures of 19°C and 37°C. At 19°C no assembly should occur, even over extended lengths of time. This is essential because reduced protease activity at that temperature requires longer reaction times of up to 8 hours. Likewise, protease activity is substantially increased at 37°C, which allows shorter experiment times. This also is necessary because the increased assembly rates would produce a non-trivial amount of assembled Cp149 if the experiment were allowed to proceed. To help rule out the possibility of capsid formation, experiments with conditions close

to assembly-inducing limits were conducted as rapidly as possible, sometimes with no (added) delay in the sampling times: aliquots were removed from reactions as rapidly as possible. These precautions were sufficient to ensure the stability of a chosen assembly state regardless of temperature, but for stronger assembly-activating influences such as HAP compounds, experiment design required further modification. Because the HAPs efficiently reduce the pseudo-critical concentration, there is no viable alternative except to reduce the molar concentration of Cp149 in solution and avoid elevated temperatures. For this reason, experiments conducted with HAP compounds were carried out with 0.3 mg/mL Cp149 and at a temperature not exceeding 25°C. These conditions represent a best-chance compromise, and it was expected that dimer assembly could take place, with data representing unknown assembly intermediate structures. Considering that there is currently no knowledge of the presence or structures of such assembly intermediates, such data would be highly desirable and negated concerns for keeping more "pure" assembly-state conditions. Experimental details for hydrogen-deuterium exchange reactions are discussed in greater detail in chapter 4 section 4, but the much reduced length of time for HX incubation compared to proteolysis helps to avoid assembly and permits the use of HAP compounds without concern for capsid formation or disassembly.

Kinetic Hydrolysis Modeling

The technique of kinetic hydrolysis is based upon the premise that proteases require a very particular conformation of substrate peptide backbone when docked in the active site channel of the enzyme. With the exception of some specific inhibitory peptides, the probability of finding a native protein structure in that conformation is nearly zero. Therefore, in order to be cleaved by a protease, any given protein will

need to adopt a non-native localized structure. There are two components to this non-native structure requirement. First, the peptide backbone much be free of any secondary structure for several amino acids both N-terminal and C-terminal to the cleavage location. Second, steric constraints restrict the presence of interfering secondary structure which is sequence-distant but spatially-close to the cleavage site. An extended loop or turn may be protease-accessible on the basis of lacking secondary structure, but it may still be protected by nearby protein elements, and thus will be protected from protease cleavage. However, when a sufficiently long section of residues is unfolded and fully solvent-exposed, it can be cleaved with a rate that is limited only by diffusion and electrostatics of the cleavage mechanism. Any sequence in such a state is referred to as being in the "open" conformation. Likewise, the same sequence of residues is called "closed" or protected when the cleavage is blocked by local secondary structure or more distant sterics. In this manner, the presence of protease allows differentiation between two states, the open and closed, and the kinetics of cleavage allows one to probe the equilibrium and kinetics of interconversion between those states. It should be noted that the equilibrium of protein motion exists as part of the native protein structure: the protease do not actually *cause* the equilibrium. Although theoretically there is a minor component of negative enthalpy when the protein substrate docks in the protease active site, this is far outweighed by the entropic terms arising from limited protein backbone conformations and overall energetics of non-native protein structural distortions.

With a protease thus capable of differentiating between the open and closed conformations, we can model proteolysis with the following equation:

$$Closed \underset{k_{close}}{\overset{k_{open}}{\rightleftharpoons}} Open \xrightarrow{k_{cleave}} Cleaved \tag{2.2}$$

where all k_{open} is the opening rate constant, k_{close} is the closing rate constant, k_{cleave} is the actual rate constant of proteolysis, and $Cleaved$ is the protein or peptide after it has been cut by the protease. From this, we can represent the equilibrium between states:

$$K_{open} = \frac{Open}{Closed} = \frac{k_{open}}{k_{closed}} \tag{2.3}$$

The k_{cleave} factor is the observed cleavage rate constant and $k_{observed}$ is the experimentally-observed first-order rate constant. The k_{cleave} factor is also assumed to be normal first-order. However, it is not invariant: k_{cleave} is itself a function of the protease cleavage kinetics, and can be adjusted at will according to enzyme concentration. Assuming that the concentration of protein substrate in the open conformation is well below the K_M concentration for that cleavage site, then k_{cleave} can be modeled as follows:

$$k_{cleave} = \frac{k_{cat}}{K_M} \times E \tag{2.4}$$

where k_{cat} is the final constant in standard Michaelis-Menten-Henri kinetics, and E is the concentration of free enzyme. Unless the opening equilibrium is shifted primarily towards the open form, and protein concentrations are very high, the cleavage kinetics will always be substrated-limited, making this a good assumption.

To this point the kinetics of proteolysis are relatively straightforward with minimal assumptions, such as the steady-state assumption of the Michaelis-Menten equation,

but further kinetic modeling requires additional and possibly non-trivial simplifications to achieve an explicit equation. First, the open/close transitions may be sufficiently rapid (technically, $k_{cleave} << k_{open} + k_{close}$) that equilibrium between the open and closed state is maintained throughout the proteolysis. This condition is known as the EX2 limit (terminology taken from HX applications), and is trivial to achieve by way of reducing enzyme concentration to reduce k_{cleave} (Equation **??**). In such rapid-equilibrium conditions,

$$k_{observed} = k_{cleave} \times open = k_{cleave} \frac{K_{open}}{K_{op} + 1} \tag{2.5}$$

At the other extreme, enzyme concentration can be very high or the open/close exchange may be slow ($k_{cleave} >> k_{open} + k_{close}$). In this situation, *open* will be nearly zero concentration and will remain that way over the course of the proteolysis. Therefore, $\Delta open = 0$, which matches classic steady-state assumptions:

$$k_{observed} = \frac{k_{open} \times k_{cleave}}{k_{close} + k_{cleave}} \tag{2.6}$$

This form of the proteolysis equation has been used by Marqusee for quantitative analysis of proteolysis kinetics [93]. At first inspection, this hyperbolic function matches the criteria for the system. When k_{cleave} is large, $\frac{k_{open}}{k_{close}}$ becomes irrelevant and $k_{observed} = k_{cleave}$: every opening event leads to cleavage in opening-limited conditions. This situation is attractive because it permits determination of the opening rate constant experimentally without consideration for enzyme activity or peptide mimics of the active site. Using equation 2.6, if k_{cleave} is sufficiently small, it simplifies to

$$k_{observed} = k_{cleave} \frac{k_{open}}{k_{close}} = k_{cleave} K_{open} \tag{2.7}$$

However, equation 2.7 conflicts with equation 2.5. With k_{cleave} small, the rapid-equilibrium model predicts $k_{observed} = k_{cleave} K_{open}$ rather than $k_{observed} = k_{cleave} \frac{K_{open}}{K_{op}+1}$.

This is not a concern so long as K_{open} is small: "normal" equilibria for well-folded proteins should not highly populate the open conformation, so $k_{cleave}\frac{K_{open}}{K_{op}+1}$ simplifies to $k_{cleave}K_{open}$, and the conflict is resolved. When K_{open} is large, such as with highly-dynamic proteins or when adapting these models to other systems such as covalent labeling, the errors can be substantial. One possible solution is to use strictly EX2-type or EX1-type conditions, with the equations for each type of reaction, but that negates the possibility of using any mixed-type EX1/EX2 data, which could comprise the majority of the physically feasible reactions for some systems.

Alternatively, accurate EX1 and EX2 equations can be reconciled into one model with the adoption of a pseudo-k'_{close}. If k'_{close} is a variable equal to $k_{close}(1 + K_{open})$ in EX2 conditions, and equal to just k_{close} in EX1-type conditions, one equation can describe the full set of possible data. We have modeled this transition in k'_{close} with the following equation, which is arbitrary but contains some desirable features.

$$k'_{close} = K_{open}\left(K_{open}\left(1 - \frac{k_{cleave}}{\frac{k_{open}(1/2K_{open}+1)}{K_{open}} + k_{open}}\right) + 1\right)^{-1} \tag{2.8}$$

When using this equation $k_{close} = \frac{k_{open}}{K_{open}}$. Although ungainly, the use of a pseudo-k_{close} with this equation makes the $k_{close}(1+K_{open})$ to k_{close} transition scale with the overall hyperbola, matching the EX2/EX1 transition. It also avoids the need for additional parameters in the equations and should ensure that the correction factor remains accurate in the transition range.

To evaluate the necessity and accuracy of these equations, a series of kinetic simulations was performed with parameters observed in Cp149 proteolysis reactions, using the Dizzy kinetic simulation software package. Differential equations were entered in the form of a series of unidirectional chemical equations, each with an associated kinetic constant. Parameters as observed from the Cp149 proteolysis experiments were used as the basis for the simulations, and the models were tested at more extreme

limits by decreasing the k_{close} parameter. The resulting systems of equations and kinetic constants were then used as inputs to a Gillespie numeric simulation model, providing the concentration of all species as a function on time. After exporting the data, it was fit using each kinetic model variant. Three possible kinetic models were evaluated. The first, termed the "classic" model, uses a static k_{close} and assumes a very small K_{open} ($open = K_{open} \times closed$). The second model, "improved equilibrium", does not assume K_{open} is small ($open = \frac{K_{open}}{K_{open}+1} closed$). The final model ("advanced") uses the accurate K_{open} model and adds the k'_{close} correction. The opening equilibria for Cp149 was approximately 1% or less, which is much larger than has been observed in studies of other proteins. However, it is not large enough to cause substantial errors when using the classic models, which is why the 2008 HBV dynamics publication was able to use the classic model [95]. However, future studies may involve much larger equilibria, so these possibilities were evaluated by decreasing the k_{close} values used in the simulations. The errors in calculating k_{open} are relatively small: less than 2% error with K_{open} values of 0.5. On the other hand, K_{open} errors are substantial: when true $K_{open} = 0.5$, the classic model calculates a value that is approximately 20% low. The improved-equilibrium model has no improvement for calculation of k_{open}, and actually worsens K_{open} calculations (30% error). The advanced model substantially improves calculations from both k_{open} and K_{open}. Errors in k_{open} decrease from 2% to 1%, which is largely irrelevant, but K_{open} errors drop to nearly zero. Considering that the advanced model does not introduce any real complexity to the data processing (no addition of variables), it should be used for any systems which have large equilibrium constants for the opening transition. However, the complexity of the equations is an impediment to clarity and accessibility, so publication of low-K_{open} findings would benefit from the use of the simpler classic models.

Determination of k_{cleave}

In order to determine the intrinsic rate constant of proteolysis for an unstructured protein backbone in the open conformation, synthetic peptides were used to assay trypsin in the manner of Park [95]. This is a very standard assay based upon separation of a FRET pair placed on the termini of a synthetic peptide designed to mimic the observed site of proteolysis. There is a substantial body of research regarding trypsin proteolysis, and it is known to have sensitivity to the three residues C-terminal to the cleaved bond, the two residues N-terminal to the bond, and partial sensitivity to the third residue on the N-terminal side. For these reasons, as well as solubility concerns, a peptide was synthesized with the two N-terminal residues and three C-terminal residues. The third N-terminal residue (W) was omitted to help ensure solubility. On the N-terminus, an Abz group (aminobenzoic acid) provides one half of the FRET pair, and a 3-NO_3-tyrosine on the C-terminus completes the set. The internal residues (IRTPP) are a match to the R127/T128 cleavage site observed on the Cp149 protein. This FRET pair provides efficient fluorescence quenching when held in close proximity by the intact penta-peptide, but gives intensity emission at 420 nm (excitation at 320 nm) following cleavage. Assays using this model peptide were conducted with exponential-rise analysis [96], and varying substrate concentrations were tested to ensure that substrate-limiting conditions were met.

The assay for intrinsic cleavage rate constant is, by itself, unremarkable, but several interesting conclusions regarding trypsin were determined. First, very specific enzyme activities may not be necessary. There is a very large body of research regarding trypsin and trypsin kinetics, and judicious compilation of the data present in existing publications can permit accurate predictions for the kinetics of cleavage at a particular cite, at a given temperature. For the IRTPP cleavage site in the

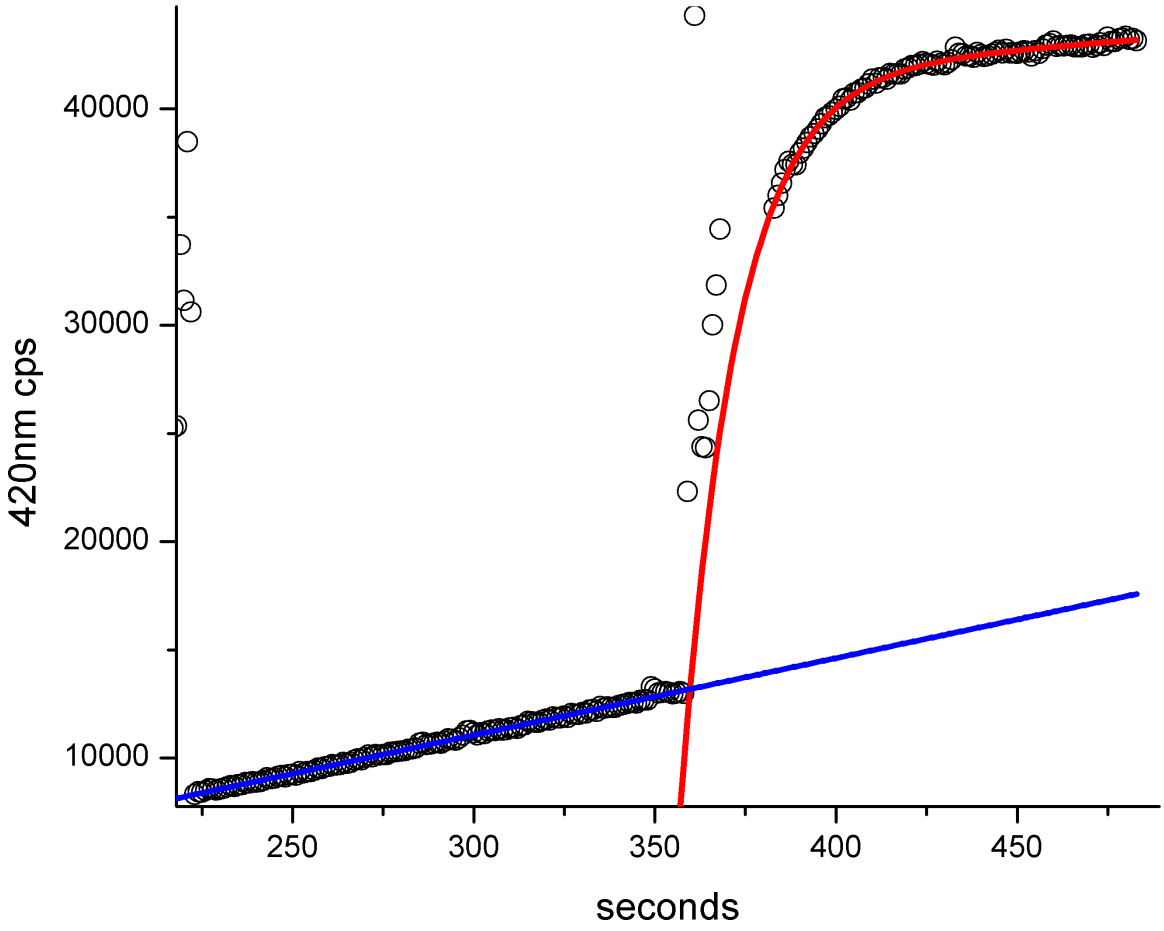


Figure 2.5: Classic fluorescence assay of the IRTPP model peptide. Background fluorescence is measured in a stirred cuvette prior to the addition of trypsin. After addition of trypsin, a nearly-linear initial rate is observed, due to the low concentration of enzyme. To determine the intensity of fluorescence with full cleavage, an additional aliquot of trypsin is added to drive the reaction to completion. The ratio of initial linear rate to overall fluorescence change is used to define the enzyme activity for that concentration of peptide and enzyme.

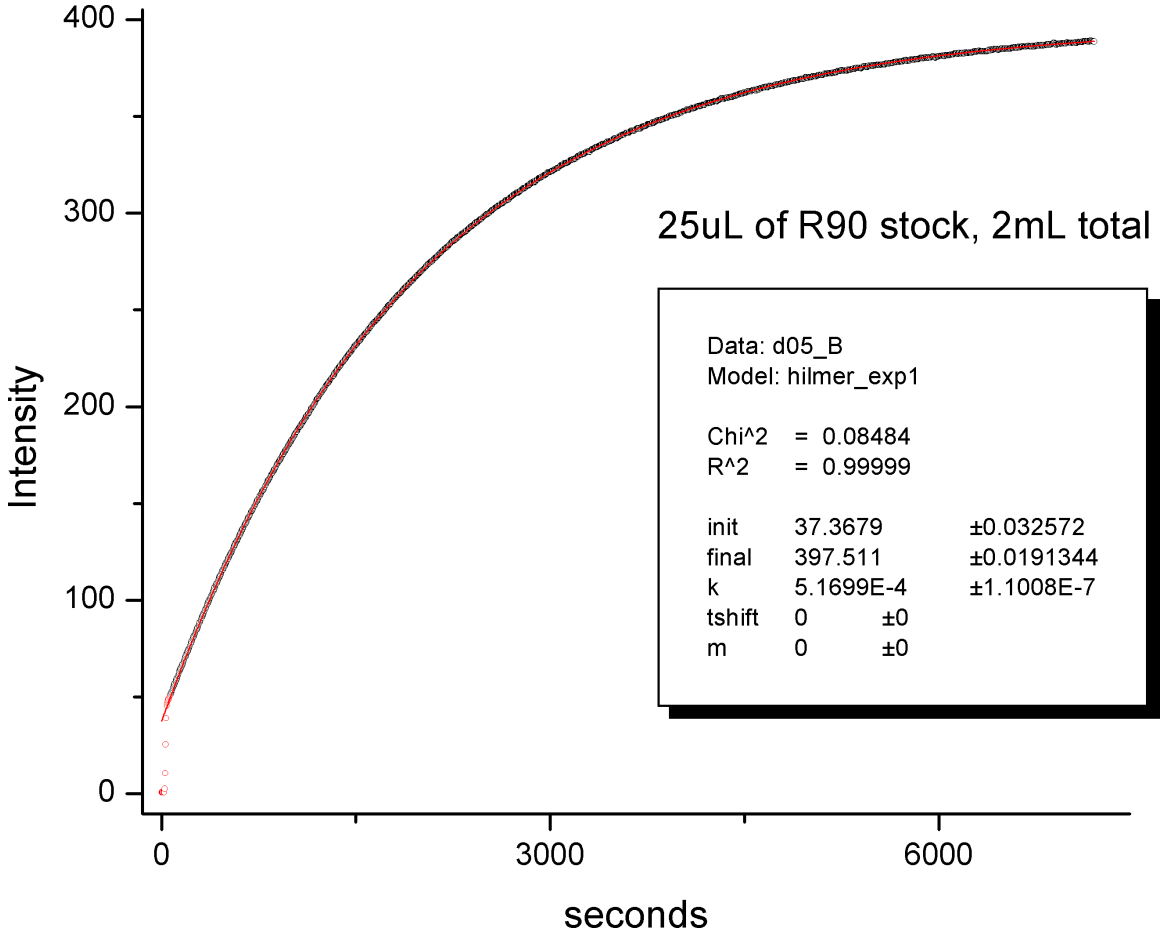


Figure 2.6: Single-stage fluorescence assay. Following the determination that trypsin was very stable with time, assays were allowed to reach completion without the addition of a second aliquot. Although more time was required to complete the assays, the overall trends were better fits to the model.

Cp149 protein, the $k_{cat}{K_M}^{-1}$ factor was predicted to within 20% of the true value. Also, the temperature dependence of cleavage at that site (measured for temperatures between 19°C and 37°C) was predicted within 10% of the true value. Some enzymes, such as thermolysin, have been systematically studied to an even greater extent than trypsin, and mathematical models are available to predict cleavage kinetics for an arbitrary peptide sequence. As data on such enzymes is accumulated with time, systematic models of the cleavage kinetics will likely make specific assays unnecessary. Second, several preconceptions about trypsin proteolysis were not born out by the assay results. Self-proteolysis is a major concern for trypsin, and high-grade trypsin is typically modified with the native arginine and lysine residues (locations of normal trypsin cleavage) removed via mutation. This form of trypsin was used for kinetic hydrolysis experiments, and it also was treated with TPCK (L-1-tosylamido-2-phenylethyl chloromethyl ketone) to remove chymotrypsin activity. Stocks of trypsin were maintained in pH 3.0 solution to further reduce self-proteolysis, since trypsin requires higher pH values (optimal about pH 8.0) for activity. The combination of K/R removal, TPCK treatment, and low pH were found to be extremely effective. Enzyme activity assays were performed over the course of a full day, and enzyme stocks were maintained on ice during that time. There was no evidence of decreasing enzyme activity occurring with time for at least 8 hours when handled in this manner. Although there were no formal assays conducted, circumstantial evidence suggests that substantial activity is retained for much longer. For example, enzyme stocks which were kept on ice for several days were used as the previously-mentioned second addition of enzyme used to drive fluorescence assays to completion, and they were very effective in that role. The effects of freeze-thaw cycles were not studied: all enzyme stocks were prepared in large batches and split into small aliquots before freezing. No enzyme aliquots were ever re-frozen or used for longer than 8 hours after

thawing. Based on the stability of trypsin when kept at low pH and at 0°C, it may be that best trypsin activity would be seen in stocks which are never frozen, but rather stored in liquid solution. The final interesting point regarding trypsin assays is their simplicity and sensitivity. The fluorescent assays of the Abz-IRTPP-3-NO_3-Tyr peptide gave ample fluorescent signal counts even at sub-micromolar concentration. This opens up the possibility of performing dark substrates assays, in which the kinetic signal arising from a standard substrate, such as a FRET peptide, changes in response to the presence of a non-detectable substrate, such as the Cp149 R127/T128 protein backbone [97]. However, the IRTPP peptide has characteristics which make it unsuitable for such an assay. By referring to equation 2.14, it can be computed that in order to be a good reporter for a dark substrate, the reference substrate needs to have not only a very high k_{cat} value, but also a very high K_M. This allows the reference substrate to be present in very low concentrations (aided by the sensitivity of the FRET assay), with a maximum decrease in signal upon addition of the dark substrate. High sensitivity in this fashion is required because the dark substrate will likely provide minimal concentrations of available substrate (arising only from the open conformation), and the K_M value for proteolysis is typically high (micromolar to millimolar range). The $\frac{S_n}{K_{M,n}}$ value from the dark substrate is thus very low, and has minimal impact on the reporter substrate. Dark substrate experiments were conducted with common trypsin substrates such as BAEE and BAPNA, which produce UV absorbance upon proteolysis, but the resulting signal changes were too low to observe. Considering the wealth of knowledge for protease cleavage kinetics, and the ability to engineer specific sequences with highly sensitive FRET tags, dark substrate experiments have promise for future study.

SDS-PAGE Analysis of Kinetic Hydrolysis

Restriction of Reaction Progression

Previous studies using kinetic hydrolysis to probe protein dynamics have focused on monomeric model proteins [98, 93, 94]. In order to adapt the methods to study HBV Cp149 dimer and capsid, some modifications were required. First, there is a concern that the proteolytic cleavages will destabilize the capsid, leading to aberrant disassembly. To alleviate such concerns, proteolysis was restricted to only the first 30% per-site of possible cleavages: 70% of the protein is completely intact without any cleavages. From a biophysical perspective, this limitation is arbitrary (though chosen to be instinctively reasonable to reviewers), but it works well within the previously-discussed limitations on assay timescale. Extended cleavage, such as 20% or less remaining intact protein, would take too much time with moderate levels of protease, and provide opportunity for assembly state changes by the end of the assay. A second reason for limiting proteolysis is to keep all cleavages within one population of subunits, if the capsid has substantially different dynamics for monomer units at different centers of symmetry. The HBV capsid has T=4 icosahedral symmetry with 240 monomers. As a consequence, there are 6 5-fold rotational centers of symmetry, giving 12 localized 5-fold centers on the surface (one per intersection of the rotational center with the capsid surface). Likewise, there are 15 2-fold, which produce 30 localized quasi-sixfold centers on the surface (technically 2-fold, but they are a relatively minor distortion from 6-fold when considering only the local subunits). Therefor, there are 60 monomers present at 5-fold symmetry centers and 180 monomers at the quasi-6-fold. If the 6-fold and 5-fold centers displayed radically different dynamics, kinetic assays could pass a transition once all monomers at a particular center were proteolyzed. For example, the 60 5-fold monomers represent 33% of the total

monomers, and 6-fold monomers represent the other ∼66%. If all proteolysis takes place at the 5-fold (this bias is not unreasonable considering that HAP-1 binding displays a heavy preference for the 6-fold centers), any cleavages past the 33% level would dramatically slow due to the exhaustion of readily-available substrate. Thus, a 20% cleavage limit is within the smallest sub-population limit imposed by quasi-equivalence symmetry. Although proteolysis reactions with dimer could theoretically be allowed to go to completion, in practice many reactions were stopped at or previous to the 30% progression level, due to previously-discussed concerns of dimer assembly at longer reaction times.

In addition to biophysical reasons to restrict reaction time, there are some subtleties of enzyme kinetics that also make it desirable to use only the initial portion of the reaction. The kinetics of proteolysis become substantially more complicated in the presence of many different substrates, which is the end result as proteolysis disassembles a largely folded and protected protein structure into a series of unstructured peptides. If we assume a standard Michaelis-Menten kinetic scheme where the enzyme forms a complex with substrate (ES_n) prior to turnover, then enzyme can be sequestered in many different forms with the various substrates available:

$$E_{free} = E_{total} - (ES_1 + ES_2 + ES_3..) = E_{total} - \sum_n ES_n \tag{2.9}$$

$$ES_n = E_{free}\frac{S_n}{K_{M,n}} \tag{2.10}$$

$$E_{free} = E_{total} - E_{free}\sum_n \frac{S_n}{K_{M,n}} \tag{2.11}$$

$$E_{free} = E_{total}(1 + \sum_n \frac{S_n}{K_{M,n}})^{-1} \tag{2.12}$$

$$ES_n = \frac{E_{total} S_n}{K_{M,n} + K_{M,n} \sum_n \frac{S_n}{K_{M,n}}} \tag{2.13}$$

The rate of the enzyme will be $k_{cat} ES_n$, so substitution gives

$$rate = \frac{k_{cat} E_{total} S_n}{K_{M,n} + K_{M,n} \sum_n \frac{S_n}{K_{M,n}}} \tag{2.14}$$

Note how this compares to the normal Michaelis-Menten equation:

$$rate = \frac{k_{cat} E_{total} S_n}{K_{M,n} + S_n} \tag{2.15}$$

Thus, when there is only one substrate, the more general multi-substrate form of the equation simplifies to the classic form. However, as substrates accumulate relative to their respective K_M values, the reaction rate of any particular substrate slows. This slowdown is a general function of enzyme kinetics, and does not take into account factors such as substrate inhibition, competitive inhibition, etc, all of which will further decrease effective enzyme activity. These factors are multiplied exponentially as large proteins unfold and produce peptides, so there is strong motivation to use only the initial reaction progression for calculation of proteolysis activity. Quantitation using mass spectrometry is a good approach to negate these concerns, since peptide production can be carefully tracked, but when using a gel-based quantitation method there are no such safeguards.

Inhibition of Proteolysis

For all of these reasons discussed previously (assembly state stability, quasi-equivalence, capsid integrity, and enzyme kinetics), kinetic hydrolysis assays should be conducted with the shortest timescales feasible and with the minimal amount of protease digestion. These factors are a necessity when dealing with HBV proteolysis, but they are also substantial complications to the practicality of conducting such

experiments. The use of short sampling intervals magnifies any errors associated with the timing reaction sampling, and stopping the reaction at precise times becomes a very serious concern. Previous kinetic hydrolysis studies have made use of enzyme inhibitors to quench the proteolysis, such as EDTA to chelate Ca_{2+} away from thermolysin, phosphoramidon as an irreversible thermolysin inhibitor, or PMSF (phenylmethylsulfonylfluoride) as an irreversible trypsin inhibitor [93, 99]. However, justification for the use of these inhibitors is extremely weak. Thermolysin contains two Ca_{2+} binding sites, neither of which is in the active site of the enzyme. It is nearly dogma that calcium is required for activity, such that EDTA is used as an "inhibitor" for many proteases in addition to thermolysin, but this has not been verified in detailed enzymatic assays of thermolysin. Instead, the calcium serves a structural stabilization role, and it is not known how effectively EDTA removes the calcium from their peripheral locations, or the zinc (a catalytic active-site metal) from the active site. Phosphoramidon is very effective as an irreversible thermolysin inhibitor, but it can take 10 or more minutes to achieve full inhibition after it has been added to solution. PMSF has very similar characteristics for the inhibition of trypsin. For enzyme assays which are conducted over hours or days, a 10 minute delay to reach full enzyme inactivation is insignificant, but when samples are removed at 3 minute intervals the delay in inhibition is unacceptable.

To achieve protease inactivation as close to instantaneously as possible, two schemes were selected. For mass spectrometry samples, aliquots from the reaction in progress were added to a chilled vial prepared in advance with a small quantity of acid and appropriate enzyme inhibitor. After mixing, the samples were immediately frozen. The resulting low-pH ($\sim$3), high concentration of inhibitor, and cold temperature combined to minimize the remaining enzyme activity. Such samples were analyzed via mass spectrometry as rapidly as possible upon thawing. Because

mass spectrometry samples prepared in this method were used strictly for peptide and protein-fragment identification, and not quantitation, small amounts of residual protein activity are acceptable. However, samples prepared for gel-based analysis used a very different protocol. For every sample, approximately 3 μL of reaction were removed and diluted into 12 μL of normal SDS-PAGE loading buffer. This mixture was rapidly vortexed and heated via digital thermocycler for 5 minutes at 100°C. Following the boiling step, samples were cooled on ice for an instant to cool and produce condensation, centrifuged for a moment to collect all condensation, vortexed, and frozen. Samples prepared in this way were observed to "freeze" and produce SDS crystals nearly instantly upon the final freezing step. To prepare the frozen samples for SDS-PAGE, the still-frozen samples were placed into a preheated thermocycler at 100°C and boiled for 5 minutes, then cooled to 25°C, centrifuged, vortexed, and maintained at temperature prior to gel loading. These additional precautions provided several benefits. First, the extreme heating and cooling steps in the presence of gel loading buffer assist in quenching protease activity, even for thermolysin. Second, Cp149 was observed to run partially as a dimer on SDS-PAGE if more "gentle" conditions were used: combined with using fresh DTT for the loading buffer in every experiment, the above protocol caused all protein to run as a monomer in the gels. Finally, the boiling and cooling of samples just before loading gels produces a very consistent viscosity of the protein/loading-buffer solution. This consistency greatly aids in reproducibility of pipetting for gel loading. Initial experiments were conducted with small concentrations of phosphorylase B diluted into the loading buffer prior to conducting the proteolysis experiments. The resulting bands were used as a control in the final gels, to evaluate protocol reproducibility and to provide a correction factor if needed for the intensity of the Cp149 band. It was found that such controls were not needed, and they were discontinued for the remaining majority of experiments. However, the

final accuracy and precision of this SDS-PAGE sampling handling protocol are highly dependent upon operator proficiency, and controls are highly recommended unless shown otherwise.

Protocol Variations and Specifics

The substantial complications imposed by the physical constraints of the HBV system, described above, prevent one single assay protocol from being used to test all conditions. The exact timing of samples and experimental preparation are varied as necessary to meet the necessary requirements. However, each assay follows the same general trend. Each reaction begins with a prepared buffer solution containing NaCl and Tris at pH 7.5: the concentrations of both are slightly elevated so that the final reaction (diluted by addition of protein and trypsin) will contain exactly 150 mM NaCl and 100 mM Tris. Protein stock is then added to the buffer solution and the mixture is rapidly vortexed. Trypsin stock is added immediately after the mixing step to begin the reaction. Timing and pipetting are critical issues for this process, due to the sensitivity of assembly to ionic strength. Prior to addition of trypsin the buffer+protein solution may contain over 300 mM NaCl, which would begin capsid assembly if left for any length of time. The addition of enzyme marks the start of the reaction, and samples were removed at appropriate intervals using the previously-described sampling method. The sampling timepoints were selected in advance to give even coverage of the 30%-completion proteolysis limit. Because some reactions (elevated temperatures, high-K_{open} protein) proceed very rapidly, and samples must be collected as quickly as possible, triplicate sampling from one reaction is not possible, due to the not-insignificant time lag between replicate samples. To deal with this, triplicates were sampled from three nearly-simultaneous reactions conducted as experimental replicates.

No exceptional SDS-PAGE protocols are necessary for running samples. Both gradient and linear gels give acceptable results, since the bands of interest are closely spaced at approximately 16 kDa. However, the running time was reduced to the minimum (approximately 25 minutes) in order to achieve vertical band separation: this is necessary to achieve sufficient horizontal separation between different lanes. Both Coomassie and Sypro Ruby stains provided good results for protein detection. Once the gel images had been acquired with the appropriate scanner, densitometry and analysis was conducted with ImageJ.

SDS-PAGE Results

Initial proteolysis experiments were conducted with both trypsin and thermolysin. The general cleavage pattern was the same for both: the intact protein degrades in a predictable first-order trend, and a single product band appeared with approximately 3 kDa smaller size (Figure 2.7). However, careful gel processing with Sypro Ruby revealed that thermolysin generated a poorly-resolved product band that overlapped the intact protein band. Considering the very small mass shift implied by this doublet, it was assigned to the hydrophobic-rich C-terminus, which has a substantial number of possible thermolysin cleavage sites. The cleavage product of trypsin reactions was identified using parallel reactions processed for mass spectrometry analysis, and multiple analysis methods confirmed the identity as the C-terminal trypsin fragment beginning at T128 [95]. Additional cleavage products were only seen via SDS-PAGE and MS after longer and more aggressive proteolysis.

The kinetics of interconversion between the intact protein band and the product followed typical first-order kinetics (Figure 2.7). The prototypic trends observed help to confirm that no additional cleavages are occurring during the initial 30%

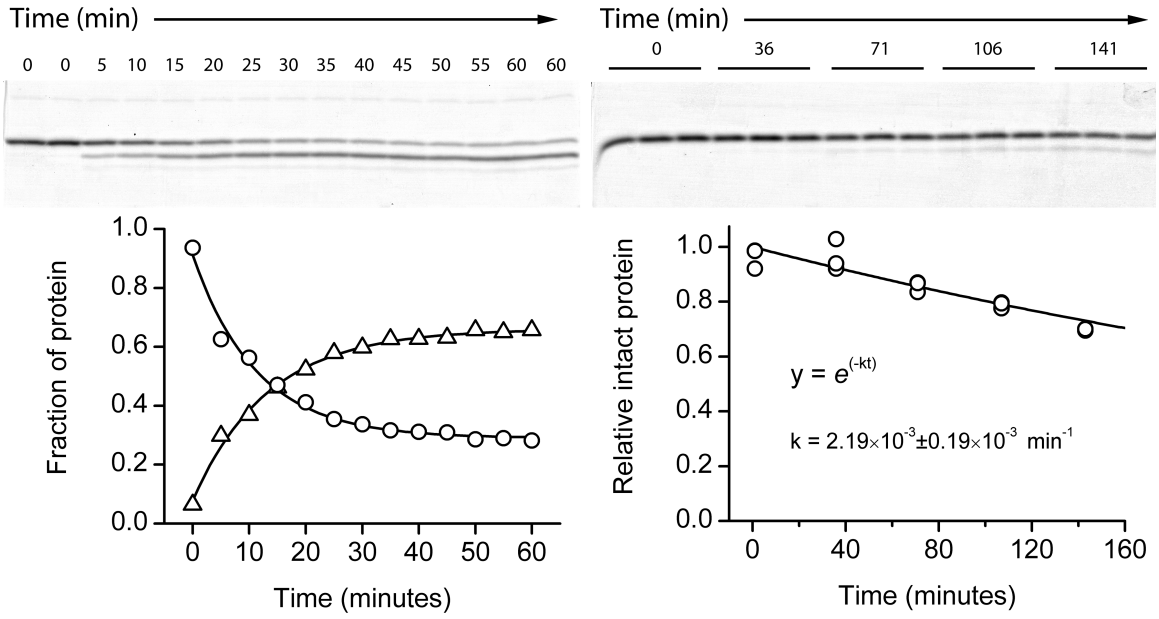


Figure 2.7: SDS-PAGE of kinetic hydrolysis reactions. The left panels show extended proteolysis of dimer: such experiments were conducted to verify the overall proteolysis trends. The very faint highest band is trypsin. The most intense band which disappears with time is intact Cp149. As Cp149 is degraded, a product band appears. The right panels show how observed rate constants were calculated. Replicate samples were used for every time point, and single-exponential fits produced the observed rate constant.

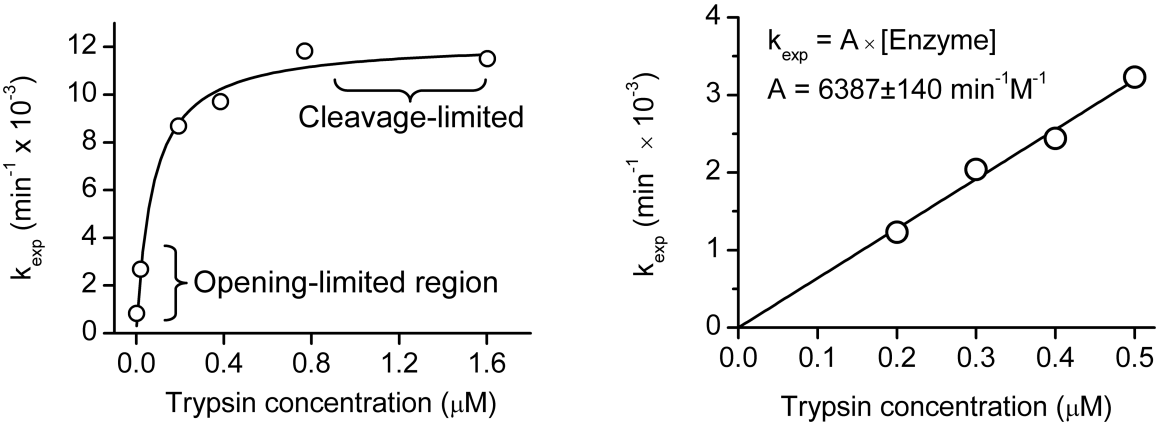


Figure 2.8: Observed proteolysis trends of Cp149 follow expected patterns. With sufficient concentration of enzyme, the observed rate constant of proteolysis reaches an upper limit (left), and the overall hyperbola trend matches the predictions of the two-state model. When enzyme concentrations were very low, the observed rate constants displayed a linear trend, which is expected for EX2-type conditions (right).

proteolysis region, and further supports the use of the two-state kinetic model for proteolysis (equation 2.2. The data from each gel was fit to a first-order exponential decay equation, $I_t = I_0 e^{-k_{observed}t}$. The resulting $k_{observed}$ values were collected from multiple experiments to generate a hyperbolic trend with respect to enzyme concentration (figure 2.8, left). The trend displayed the expected cleavage-limited (EX2) and opening-limited (EX1) conditions, and such plots and fits were used to generate K_{open} and k_{open} kinetic constants. However, when possible, higher-precision EX2-only plots were used to get K_{open} values (figure 2.8, right).

Calculations and Discussion of SDS-PAGE Data

The overall results from the SDS-PAGE kinetic hydrolysis experiments are summarized in table 2.1 and Figure 2.9. The values for k_{open} increase steadily with temperature for both dimer and capsid, showing no substantial difference between

Table 2.1: Kinetic and thermodynamic parameters from the IRTPP cleavage site.

Temperature (°C)	19	25	31	7
Peptide k_{cat}/K_M, ($\text{min}^{-1}\text{M}^{-1}$)	$7.66 \pm 0.23 \times 10^6$	$8.92 \pm 0.04 \times 10^6$	$10.6 \pm 0.26 \times 10^6$	$12.5 \pm 0.45 \times 10^6$
Dimer k_{open}, (min^{-1})	$3.53 \pm 0.38 \times 10^{-3}$	$5.08 \pm 0.51 \times 10^{-3}$	$7.60 \pm 0.70 \times 10^{-3}$	$1.19 \pm 0.13 \times 10^{-2}$
Capsid k_{open}, (min^{-1})	$2.44 \pm 0.24 \times 10^{-3}$	$4.57 \pm 0.47 \times 10^{-3}$	$1.13 \pm 0.10 \times 10^{-2}$	$1.29 \pm 0.07 \times 10^{-2}$
Dimer K_{open}	$1.33 \pm 0.83 \times 10^{-2}$	$7.54 \pm 0.26 \times 10^{-3}$	$6.34 \pm 0.80 \times 10^{-3}$	$2.59 \pm 1.03 \times 10^{-3}$
Capsid K_{open}	$5.12 \pm 2.28 \times 10^{-3}$	$1.78 \pm 1.56 \times 10^{-3}$	$2.90 \pm 0.62 \times 10^{-3}$	$1.24 \pm 0.30 \times 10^{-2}$
Dimer opening ΔG° (kJ mol^{-1})	11.59	12.11	12.79	15.35
Capsid opening ΔG° (kJ mol^{-1})	18.40	15.68	14.77	11.31

the two forms. In contrast, the K_{open} equilibrium constant between the open and closed conformations is very different: on an absolute scale dimer and capsid are similar, with approximately 1% of the protein populating a cleavage-accessible conformation, but the trends scale completely differently with temperature. For dimer, the equilibrium shifts to a more protease-protected conformation as temperature is increased. Van't Hoff plots for the temperature dependence predict ΔH = +54.5 kJ mol^{-1} and ΔS = +224 J $\text{K}^{-1}\text{mol}^{-1}$ for the closing transition in dimer. These values share trends and are within an order of magnitude for parameters calculated via capsid assembly assays [70], which suggests that the folding process may share features with the implied assembly-activated state predicted from assembly reactions. In contrast, the capsid does not display the same entropic-stabilization features seen in dimer: ΔH = -125.9 kJ mol^{-1} and ΔS = -368 J $\text{K}^{-1}\text{mol}^{-1}$. This may suggest that in capsid, the closed state can be stabilized in an unknown conformation, perhaps more extended than the hydrophobic-stabilized turn that exists in the capsid crystal structure. Regardless, kinetic hydrolysis experiments show that the dimer and capsid forms of Cp149 have different structures, either overall or localized to the cleavage location.

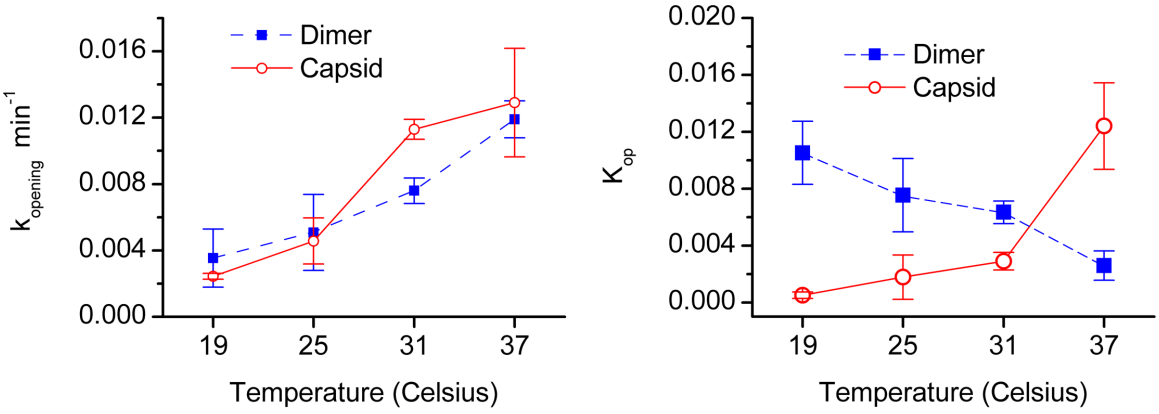


Figure 2.9: Kinetics and thermodynamics of Cp149 dynamics. The opening rate constant is very similar for both dimer and capsid, and increases with temperature. However, the opening equilibrium shifts to a more folded conformation for dimer (blue squares) as temperature is increased. The trend for capsid (red circles) is the opposite, with a higher population in the open conformation at higher temperatures.

The scale of dynamics observed for the Cp149 protein, in both dimer and capsid forms, is surprisingly high. Similar kinetic hydrolysis studies have found opening equilibria to be more than 200-fold lower than the values measured here, which range from 5×10^{-1} to 0.0124. In fact, the highest exposure rate constants are observed for capsid, which sequesters a very large amount buried hydrophobic contacts through the many dimer-dimer interactions on the capsid surface. The dimer also presents very substantial steric hindrances for protease access to the cleavage site. Manual docking models indicate that the C-terminal helix may unfold substantially to achieve the 12 Å translocation required to dock the R127/T128 peptide bond from the native structure location into the closest-approach trypsin active site. The outdated concept of a viral capsid as a static cage is clearly not applicable in this situation: the network of weak interactions which stabilizes the capsid as a whole may create a very large net stability, but they are not an impediment to rapid equilibrium motions in solution, despite relatively large displacement amplitudes. Although it is circumstantial, these

results may also provide evidence for structural transitions associated with capsid assembly and disassembly. The dimer solutions at 37 °C are extremely close to the pseudo-critical concentration for capsid assembly: dimer in these conditions is pushing the limits on the progression to becoming capsid. Likewise, capsid at 19 °C is very close to destabilization and disassembly. For both these conditions, K_{open} values are minimized, which implies that the assembly-active structure may consist of a tightly folded dimer subunit. Such a condition would fit in well with the the kinetic models of assembly: minimized extended or intercalated dimer-dimer interactions would provide the fewest kinetic traps for aberrant assembly structures, and would also provide the smallest energy barrier for capsid dissolution.

KINETIC HYDROLYSIS AND THE INFLUENCES OF ASSEMBLY EFFECTORS

Introduction

The initial kinetic hydrolysis study of Cp149 showed that the technique was a viable option for supramolecular complexes and that the HBV capsid displays substantial dynamics which can be regulated as a function of environmental effectors. In light of the assembly data collected by Zlotnick et al in response to assembly effectors such as zinc and HAP compounds, we proceeded with the application of kinetic hydrolysis in an attempt to help deduce the detailed mechanism and thermodynamics of capsid assembly. SDS-PAGE analysis of Cp149 dynamics showed that the C-terminal helix displayed varying dynamic behavior, and the capsid structure suggested that substantial unfolding would be required to reach the open conformation at this cleavage site. Perhaps not coincidentally, the putative HAP binding location (as inferred from the capsid crystal structure co-crystallized with HAP-1) falls in the inter-dimer interface between the C-terminal helices on adjacent dimers around the quasi-sixfold centers of symmetry. Thus, kinetic hydrolysis should serve as a reporter for stabilization or destabilization effects in that region. In addition to the HAP compounds, zinc was tested due to its known activity as an assembly activator. However, unlike the heteroaryldihydropyrimidines, the binding location for Zn^{2+} is not yet known: one of the objectives of testing kinetic hydrolysis in the presence of zinc was to provide data that might suggest its mechanism of action. The final results of the kinetic hydrolysis experiments do confirm that both HAP and zinc have very measurable effects on cleavage at the C-terminus, but both types of effector

induce substantial shifts in the dynamics that can best be described by long-distance allosteric effects.

Analysis Methods: HPLC-MS

SDS-PAGE is an absolutely essential tool, and provides a robust analysis method for kinetic hydrolysis, but for further kinetic hydrolysis studies the efficacy of SDS-PAGE was evaluated before proceeding with the experiments. Polyacrylamide gels are extremely thorough and reliable; if a protein is present it will be detected, and the effective mass range spans sub-kilodalton peptides up to hundreds of thousands of kilodaltons. Protein band staining is very reliable, relatively insensitive to amino acid composition, and scales predictably with protein size. However, while SDS-PAGE provides great robustness, it lacks precision. Protein migration through the acrylamide matrix is in imprecise function of protein size, which makes all mass determination estimates. Separation between bands is heavily influenced by variable factors such as heat, gel running potential, lane loading, etc. It is also nearly impossible to achieve a high degree of automation, and manual steps of the SDS-PAGE protocol can introduce very substantial errors. For these reasons, SDS-PAGE analysis of kinetic hydrolysis reactions is sufficient but far from ideal. Since mass spectrometry is required for identifying proteolysis fragments, regardless of SDS-PAGE application, a natural extension is to conduct all analyses strictly via liquid chromatography mass spectrometry (HPLC-MS). The use of mass spectrometry as a quantitative tool is becoming a widely-accepted analytical technique, and HPLC-MS provides many benefits for kinetic hydrolysis assays. Most importantly, electrospray ionization mass spectrometry is a very robust and reliable process for generating ions, which provides excellent measurement reproducibility. Quantitative signal reproducibility is further

enhanced when the ESI interface if preceded by automated HPLC, which provides for high-precision sample introduction and analyte transfer into ESI-compatible solvent conditions prior to delivery to the MS source.

RP and SEC Chromatography

In addition to the general benefits of HPLC-MS, automated chromatography with ESI simplifies the kinetic assays of proteolysis. Besides precision, the primary concern for proteolysis assays is the ability to quench protease activity. This can be achieved with heating and inhibitors, but solvent composition changes via chromatography are both fast and efficient. The two forms of chromatography used for Cp149 assays are reverse-phase and size-exclusion. Reverse-phase chromatography (RP) involves passing an aqueous solution over a hydrophobic stationary matrix. In the presence of water, proteins and peptides have high affinity for the stationary matrix and remain bound to it. This effectively immobilizes the protein of interest and the enzyme, which halts the proteolysis process. After the proteins and peptides are bound to the column, it can be washed to remove salts, then eluted with solvents composed of higher concentrations of organics such as acetonitrile, methanol, or isopropanol, which displace the hydrophobic protein/matrix interface. The second form of chromatography, size exclusion, is based upon varying migration rates of different-sized particles through a porous matrix. Both types of chromatography are highly compatible with ESI, though their different characteristics make them applicable for different situations.

Reverse-phase chromatography benefits greatly from the ability to create columns with extremely low backpressure, but capable of tolerating very high pressure. This permits very high flow rates without exceeding pressure limits, which are typically between 200 and 400 bar. With rapid flow rates, RP gradients can be extremely

rapid, which decreases the time necessary for each analytical run. The column/-analyte interaction also permits elution peaks with better resolution and separation compared to size-exclusion: SEC tends to blur peaks as elution time increases, and separation between analytes of different sizes decreases with smaller analyte size. As a consequence, peptides tend to be poorly resolved from intact proteins, protein fragments, small molecules, and salts. This is a major concern when performing searches for all peptide fragments, since peptide spectra tend to be complex and lack clear elution trends. Reverse-phase chromatography provides better separation from buffer components such as salts, more control over the elution time. For these reasons, HPLC-MS experiments carried out in concert with SDS-PAGE analysis, to detect proteolysis products, made use of extended RP gradients on high-retention columns.

During the course of the experiments using SDS-PAGE analysis, HPLC-MS methods were in constant development to eventually serve as replacements for the gel-based approach. Especially for smaller protein fragments and peptides, analyte handling (chromatography behavior, quantitative reproducibility, signal detection, etc) and instrument response were excellent even with very preliminary tests. Larger protein molecules such as intact protein produced slightly more irregular performance, but efforts to improve protocols were found to have measurable positive effects for such analytes. This suggested that with further optimization, HPLC-MS analyses had substantial room for improvement, and would be a viable replacement for the SDS-PAGE methods. Following the conclusion of the temperature-dependence study (chapter 2), mass spectrometry became the primary analysis method for kinetic hydrolysis reactions.

Carry-Over

For electrospray ionization of proteins and peptides, the most common chromatography method is reverse-phase. However, reverse-phase chromatography suffers greatly as a result of analyte carry-over, in which analyte from one run cannot be fully washed out of the column in a reasonable timeframe. Conventional wisdom holds that carry-over is a result of manufacturing defect or improper column stationary phase selection, and can be eliminated by using low-retention stationary-phase. Reverse-phase columns with short carbon chains, such as C4, are predicted to be free of carry-over from protein analytes. However, we have found that viral capsids do not follow this idealized behavior. In order to ensure that low-concentration peptides are detected, large quantities of total reaction mixture (1-10 μg total protein) must be injected onto the column. As a consequence, the intact protein displays very substantial carry-over when using normal RP solvents, such as acetonitrile. Because the carry-over is presented in the following run at the time of normal protein elution, it eliminates the ability to make basic quantitation. The carry-over phenomenon was investigated in depth with Cp149 to determine optimal RP-HPLC-MS conditions, and it was found sub-microgram quantities of protein per injection were sufficient to produce significant carry-over (figure 3.1). Since the carry-over effect changes with analyte composition (peptides have minimal carry-over compared to proteins), it can not be trivially corrected for analyte mixtures.

Since size-exclusion chromatography ideally has no bonding interactions between the column and analyte, it has nearly no carry-over. This permits very substantial amounts of analyte injection per chromatography. Unlike RP-LC, SEC does not use gradients, instead using isocratic elution with a constant mixture of water and organic solvent. One result of isocratic elution is that solvent composition at the ESI

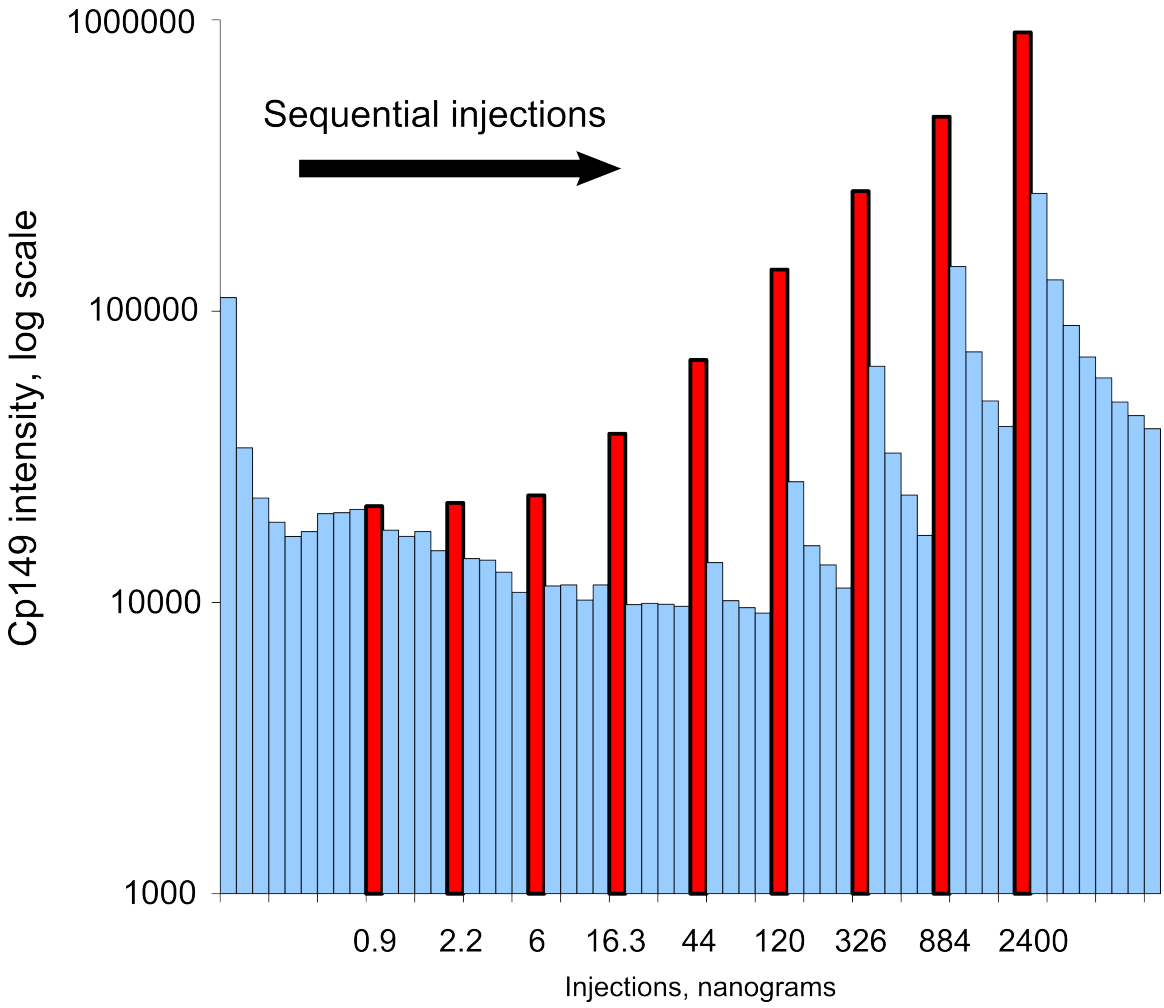


Figure 3.1: Cp149 carryover on RP column. Following every protein injection (white with heavy black outline), a series of blank runs were performed to check for carryover. Note that nearly 30 injections are necessary to reduce the background to the minimal possible levels: instrument cleanliness is a very substantial concern. An injection of 44 μg is very marginally acceptable, and every injection with more than that quantity causes very significant carryover that takes many blank runs to remove.

source is perfectly constant, so solvent-borne contaminants produce very consistent background signals and can be trivially removed. This background subtraction also allows compensation if carry-over is present: carry-over in SEC columns typically decreases linearly or exponentially with time, so the degree of carry-over from a previous run can be estimated by the analyte signal present just before the normal elution time. An example of this is shown in figure 3.2. For these reasons, SEC was selected as the method of choice for automated HPLC-MS analysis of kinetic hydrolysis reactions. The drawbacks of peptide signal complexity are negated by the usage as a routine analysis: peptide signals can be identified via more carefully directed analyses, and with prior knowledge of the expected signals, the elution complexity is not a problem.

Automated SEC-HPLC-MS

Hardware

One of the major factors contributing to the success of the HPLC-MS analysis method was the availability of high-quality HPLC autosamplers. Temperature-controlled Agilent 1200-series autosamplers were used as a starting point for analyte introduction into the HPLC. Since this series of autosampler is designed for high-precision quantitative chromatography, typically with UV-Vis detection rather than MS, accuracy and precision of the injection volumes is excellent. The actual specification on the G1377A μ-WPS autosampler is $< \pm 3\%$ for 0.2 to 1.0 μL injections, but our tests with reference peptide replicate injections indicated that real precision was substantially better than the specification: volumes of 0.01 μL could be drawn accurately and reproducibly from the sample vials. In practice, there is no reason to push the performance limitations in this manner, so volumes of 0.1 μL and larger were used for reaction sampling. The use of an autosampler also permitted reactions

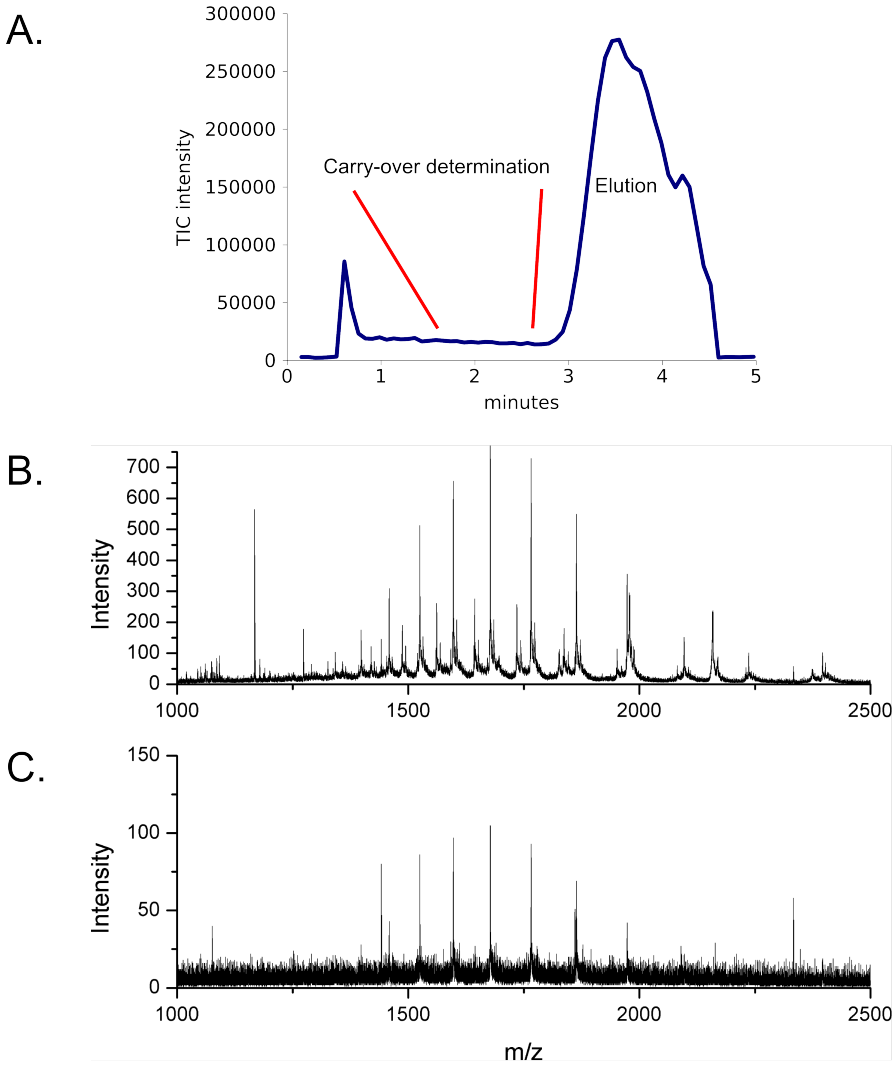


Figure 3.2: SEC provides minimal carryover and easy correction. A) Unlike RP, SEC produces carryover continuously, which allows easy correction within one run. The ratio of carryover (C) to the elution signal (B) is also much less in RP than SEC.

in progress to be left in the autosampler vials, with aliquots automatically removed at the appropriate timepoints.

Following the autosampler in the HPLC flowpath, the solvent was directed through a column selection valve and through the size-exclusion column. It is noteworthy that the injector loop valve and the column selection valve were the primary locations of flow obstruction for all HPLC runs of both RP and SEC methods. Since the HPLC-MS was not dedicated solely to Cp149 analysis, a very wide variety of analytes, columns, and solutions were used on the instrument in question, and many problems were traced to other experiments conducted on the same hardware. However, this provided a somewhat unbiased evaluation for the common problems associated with such HPLC systems. Although no detailed investigation was performed, the flow obstruction of the HPLC valves was traced to two primary sources. First, the use of PEEK tubing with fused-silica linings was highly correlated with flow path clogs. Specifically, the fused-silica tended to trap (or generate) particulate matter where the tubing connected to the valve port. This problem was far more common the more often the tubing was manipulated and was also extremely common when the fused-silica terminus had an imperfect cut or had been damaged. After all fused-silica had been removed from the HPLC flow path, the incidence of obstructions dropped substantially. Second, the infusion of certain types of capsid (CCMV, DPS, Norwalk, and others have been routinely analyzed on the HPLC-MS system) and capsid solution conditions into the HPLC flow path generated obstructions far more frequently than others. Considering the known issues associated with salt precipitation in high-organic solvent streams, and the tendency of proteins to aggregate and precipitate in the right conditions, these clogs can be attributed to sample-specific particulate matter, and care should be exercised to avoid such situations. However, the Cp149 system never displayed any tendency to form such precipitates, and very

few HPLC obstructions were observed that actually occurred during the course of kinetic hydrolysis experiments.

The column used for size-exclusion HPLC was a Phenomenex BioSep-SEC-S2000 in [DIMENSIONS] format, which has a theoretical effective range between 200 and 1 kDa. However, this assumes that analysis is conducted in solvent conditions with high concentrations of buffer, as would be performed with UV-Vis detection. The solvent selected for the kinetic hydrolysis experiments consisted of 50% water, 50% methanol, and 0.1% formic acid. This solvent provides efficient ionization at the ESI source, does not produce salt-based precipitates from the sample introduction, and helps to minimize backpressure. However, it does reduce the elution efficiency and resolution of the BioSep column, which was not a concern for this particular analysis.

The solvent flow rate through the HPLC was 100 μL/min, which produced very effective ESI at the source. The mass spectrometer was a Bruker micrOTOF equipped with the standard ESI source. This particular instrument uses an orthogonal TOF mass analyzer with a detector coupled to an ADC rather than the standard TDC. As a consequence, mass accuracy is excellent across a wide range of signal intensities, and detector saturation is much less of a concern compared to a typical TDC-equipped instrument. The source was always operated in the positive mode, which is standard for protein-based MS analyses. The combination of autosampler, SEC column, and flow rate described here allowed a turnaround time of less than 10 minutes per sample. Theoretically this value could be further reduced by making use of overlapped injections, but the tradeoff would be signal complexity and worsened carryover.

Protocols

Experiments for automated mass-spectrometry analysis were conducted with solvent conditions identical to the previous temperature-based study: 150 mM NaCl, 100

mM Tris pH 7.5, and with all the precautions previously described. However, protein concentration was reduced to 0.3 mg/mL. This was done for two reasons: first, in the presence of HAP compounds, 0.5 mg/mL may begin to assemble, and second, the protein stocks provided by collaborators were not sufficiently concentrated that it was possible to achieve 0.5 mg/mL in the final solution. Unlike the temperature-based study, these assays also contained 1% DMSO. Because the HAP compounds are not highly soluble in water, and are expected to degrade with time in aqueous conditions, all HAP stocks were provided in DMSO, and this caused the introduction of a small amount of DMSO into the final reaction. All control reactions also contained 1% DMSO to account for any minor effects it might have. Although intial tests used varying concentrations of HAP, the majority of experiments were conducted with 17.89 μM HAP in the final solution. This is a 1:1 stoichiometry of HAP:monomer of Cp149, and ensures that every dimer has the possibility of full occupancy. Reactions involving zinc also contained DMSO, but zinc concentration was held at 50 μM. That particular level was chosen based upon assembly experiments, which suggested that such a concentration of Zn^{2+} would be sufficient to cause changes in the dynamics, but would not be so much as to cause dramatic and immediate assembly.

Like the previous series of experiments, Cp149 was incubated in the presence of buffer an effector prior to initialization of the reaction. However, whereas previously it was important to initiate proteolysis as soon as possible after Cp149 had been added to the buffer mixture (due to dilution effects on the NaCl concentration), for HAP and Zn experiments there was little to no dilution effect, because trypsin concentrations (and thus volumes) were extremely low. Typical experiments were conducted with 1.5 μg/mL of trypsin, which is extremely dilute compared to previous concentrations [95]. As a consequence, the pre-trypsin mixture contained concentrations of salt and buffer that were insignificantly different from the final desired value. This is extremely

important because it permitted pre-incubation of Cp149 with the assembly effectors: all solutions were maintained for 10 to 15 minutes prior to addition of enzyme, to account for any slow binding interactions. However, this step may be superfluous considering the rapid induction of assembly seen upon adding effectors in the Zlotnick assembly experiments.

The elution times of Cp149, the peptide products, and the small-molecule contaminants were the constraints used to set the chromatography and data collection timing. Cp149 eluted from the column at approximately 3 minutes, with a slight tailing. Smaller peptides elute during the Cp149 tail, from approximately 3.5 minutes to 4.5 minutes. Beginning at approximately 4.5 minutes, a very large quantity of small molecule polymer contaminants begins to elute, combined with salts sometime slightly after. Because of the intensity of the polymer contaminants, which derived from the Cp149 protein stocks, the source/waste diversion valve must direct to waste beginning at 4.6 minutes. The polymer contaminants continue to elute until approximately 12.5 minutes post-injection with exponential decay, and with overloaded Cp149 injections, protein signal can be seen with exponential decay until approximately 13.5 minutes post-injection. This allows an overlapped injection scheme with 7 minute separation between injections: for the first 0.5 minutes following injection solvent is diverted to waste, since it contains carry-over polymer contaminants, salts, and other unknown small molecules. From 0.5 minutes to 3 minutes the signal is extremely clean, with minimal amount of Cp149 carry-over from the previous sample, which is used to correct for the quantitation of the current injection. Data collection between 3 minute and 4.5 minutes is representative of the injected sample, and is followed by diversion to waste again to avoid unnecessary contamination from the small molecules that elute at that time. All reactions were run with alternating control and experimental injections, with the second reaction of the series begun with a time offset equiva-

lent to the chromatography cycle time. This eliminates signal response variability as a concern, which could be a problem if control reactions were run one day and experimental the next.

Data Processing

As implied by the selection of size-exclusion chromatography and the data collection scheme, intact proteins and protein fragments are the primary signal of interest for this analysis. There are two reasons for this selection. The first involves the combinatorial problem of proteolysis and the possibility to detect quasi-equivalence, and is discussed in the next section with the shared kinetic model of proteolysis. The second reason pertains to equation fitting for mass spectrometry data. Unlike UV-Vis absorbance, which is inherently quantitative and reliable, mass spectrometry is much closer to fluorescence. Although the signal responses can be calculated and predicted for both, for complicated systems it is an imperfect prediction and subject to interfering effects. It is for this reason that the fluorescent assays of proteolysis (chapter 2, section 2) require that the assay be driven to completion, so that the observed intensity can be correlated with a physical concentration. In effect, both fluorescence and mass spectrometry involve a (usually unknown) signal response variable, which causes the equations for data fitting to have an additional degree of freedom. Although this parameter can be experimentally determined via sufficiently long reactions, it is imprecise and its inclusion reduces the precision of all other variables in the fitting equation. This problem of an unbounded upper end of signal response is primarily a concern with signals that increase with time: the appearance of fluorescence from a cleaved peptide or the corresponding peptide signal directly observed in a mass spectrometer. On the other hand, signals which start from a defined level at the beginning of a reaction, and decrease with some exponential trend, are inherently

bounded by zero at completion. This reduction in the degrees of freedom typically produces much higher quality fits for a given amount of noise, making these decaying signals the preferred ones to collect in such experiments. For proteolysis reactions, protein signals belong in this well-bounded decaying category, which is why they are the analyte of interest.

Protein signals are also the preferred category due to mass spectrometry signal response variation. Because the observed intensity in a MS spectrum is a function of the ionization, transfer, and detection efficiencies, the physical features of a molecule determine the final signal response. A natural consequence of this is that similar molecules tend to have similar signal responses, and the similarity can be roughly approximated by the relative change in molecule size (ignoring charge effects). For example, a 200 kDa protein would not be expected to show any substantial change in signal response upon loss of 5 random amino acids. On the other hand, if 5 random residues were removed from bradykinin, a 9-residue peptide, the change in signal response would be enormous. This signal response effect is not a concern for fitting kinetic data within one run, for one molecule, but it has larger implications. Day-to-day variation in mass spectrometry instrumentation and signal response is a concern, and in a shared-instrumentation environment, largely out of the individual's control. A very small peptide could be easily visible with high intensity one day but gone the next, which greatly complicates that data analysis. A robust signal response from a protein also provides an estimate for similar proteins: with a small terminal peptide removed from a protein, the remaining protein fragment should have roughly similar signal response to the original intact protein. This allows approximations regarding absolute quantitation without the need to directly calculate signal response factors.

Protein quantitation does involve one major drawback: because ESI mass spectrometry signals of large molecules tend to be highly charged, the observed signal

which arises from a molecule is scattered across the spectrum, with one peak for every charge state. The distribution of charges has some variability, but the primary problem is that S/N is reduced for every peak, and noise variability at every charge state introduces quantitation errors into the final calculation. A final relatively minor concern is that the data is not very accessible in this form: calculations must be performed in order to identify masses and low-intensity charge state envelopes can be difficult to detect. The solution to this problem is to use a process known as deconvolution. In the classic sense, deconvolution refers to the recovery of the original signal, given (or estimating) a convolution function which blurs the data. However, the mass spectrometry implementation of this term and process is slightly different. With regards to mass spectrometry, deconvolution does not imply a continuous or constant convolution function, but rather a convolution which is discontinuous and discrete according to the equations governing multiple charges in m/z. The most common current MS deconvolution algorithm is known as "maximum entropy deconvolution", and it can generate the theoretical spectrum of uncharged molecules most likely to produce the actually-observed convoluted spectrum. The exact implementation of entropy in this algorithm is unknown, and this is further discussed in the appendix, but the key point is that is does produce the *most probable* spectrum, after accounting for noise. The sensitivity to noise is a major concern for quantitative analysis: theoretically, a protein signal with high noise and high intensity will produce a lower final signal upon deconvolution, when compared to a low-intensity, low-noise signal. In practice, extreme swaps such as this example are very unlikely, but care must be taken to ensure that noise levels are very comparable between spectra. Variations in signal processing, such as averaging across slightly different extents of a peak, will cause intensity variations in the deconvoluted spectra.

The sensitivity of deconvoluted spectra to noise almost completely rules out manual data processing for high-throughput assays. Thankfully, the Bruker software provided for analysis of micrOTOF data, known as DataAnalysis, has a scripting engine based upon Visual Basic. This scripting back-end was used for all processing of mass spectrometry data, and example scripts can be found in appendix A. The general data handling scheme proceeded according to the physical characteristics of the SEC column: a background spectra was defined in the region between 1.5 and 2.5 minutes. Using this background spectra, a new TIC (total ion chromatogram) was defined. The background-subtracted TIC was averaged from 3 to 4.5 minutes, and the resulting spectrum was deconvoluted via the maximum entropy algorithm with a consistent definition of 3000 to 100000 Da at 1 Da intervals, 10000 resolution. This definition produces high-quality spectra for the Cp149 protein and proteolysis fragments with baseline background levels. Following deconvolution, the resulting spectra was exported to a flat text file format, and subsequent processing was performed with custom-written Python applications which performed quantitation on known mass ranges and performed data-fitting to the kinetic models.

Shared Parameter Kinetic Model of Proteolysis

Normally, interpretation of the MS results would proceed independently of the data analysis. However, proteolysis is a very computationally complex problem, and managing the data in a reasonable way requires that some assumptions be made. The fundamental problem is that for any protein with a reasonable number of cleavage sites, there are far too many possible protein fragments and peptides that can be produced. In the case of Cp149, there are nine possible cleavage sites, which gives rise to 55 potential protein fragments. That is a feasible number of signals to quantify,

but Cp149 is a dimer and also contains a potential disulfide. Regardless of whether the disulfide is oxidized or reduced, via mass spectrometry it is always seen as a dimer (noncovalent interactions are reasonably easy to maintain via ESI). Because the likely disulfide is at position 61, 30 of the 55 possible protein fragments contain the disulfide, and could remain bound to 30 possible protein fragments from the other monomer. The resulting 900 signals become much more troublesome to fully evaluate. To complicate this, possible modifications such as methionine, cystein, or tryptophan oxidation rapidly increase the number of covalently-bound masses to more than 100,000, and if noncovalent interactions keep partial fragments together the number increases further still. To make this problem tractable, two main criteria were applied. First, high-intensity signals were manually reviewed for sequence matches based upon high mass accuracy. Second, the peptides detected via SEC and RP methods were used to pinpoint known cleavage locations. By combining these approaches, two primary cleavages were identified for dimer and capsid forms of Cp149. The first cleavage location is between R82 and D83, and the second is between R127 and T128. These two cleavages together have the potential to produce 6 different proteins/peptides per monomer. Assuming the presence of the disulfide at C61, and not assuming symmetry between the two halves of the dimer, there are 15 protein fragments which could contain the disulfide, and 6 peptides without it. Because of the identical sequence between the two monomers, many fragments are isobaric (and identical if symmetry is assumed), so the 15 possible fragments are condensed to 6 observable masses, while the 6 peptides are reduced to 3 masses. These 9 masses were manually verified by examining the spectra, and are the basis for the quantitation of Cp149 kinetic hydrolysis reactions.

Structure to Model

A review of the capsid structure shows that there is no reason to enforce symmetry on a kinetic model of Cp149 proteolysis. Within the capsid, dimers span the gaps between 5-fold and (quasi-)6-fold centers of symmetry. These centers have noticeably different features, such as increased convex protrusion at the 5-fold compared to the flatter 6-fold, and the reduced symmetry of the 6-fold compared to the 5-fold. In addition, the crystal structure obtained with HAP-1 showed occupancy only at two of the three possible binding locations within the 6-fold center, which is further evidence that full symmetry is not maintained in the capsid form.

For this reason, the two cleavages (82/83 and 127/128) are not assumed to have identical kinetics for the two halves of the dimer. This gives the following possible fragments, which are denoted assuming the two monomers are connected at the N-terminus (since no cleavage site is N-terminal to C61), with the first number giving the terminal residue number on the first dimer, and the second number giving the terminal residue number on the second dimer.

1. 149-149
2. 149-127
3. 149-82
4. 127-149
5. 127-127
6. 127-82
7. 82-149

8. 82-127

9. 82-82

Note that items 2 and 4, and 3 and 7, are not equivalent in this scheme: they may have substantially different kinetics. The peptides use the same nomenclature but cannot be part of a disulfide-linked molecule, so both numbers refer to residues within one monomer.

1. 83-149

2. 83-127

3. 128-149

4. 83-149, dimer #2

5. 83-127, dimer #2

6. 128-149, dimer #2

These fragments and peptides are illustrated graphically in Figure 3.3.

Model Features

In order to realistically model unique but isobaric protein fragments, the kinetic model calculates each of the 12 possible masses individually, then sums the appropriate isobars. After summation, the isobar sets are transformed via the signal response equation to produce theoretical observations which are matched to the data. These isobars are proteins [2 and 4], proteins [3 and 7], and the peptides, each of which has a corresponding match from the other monomer. The kinetic trends for any particular mass (before summation of isobaric pairs) is calculated from a production/survival

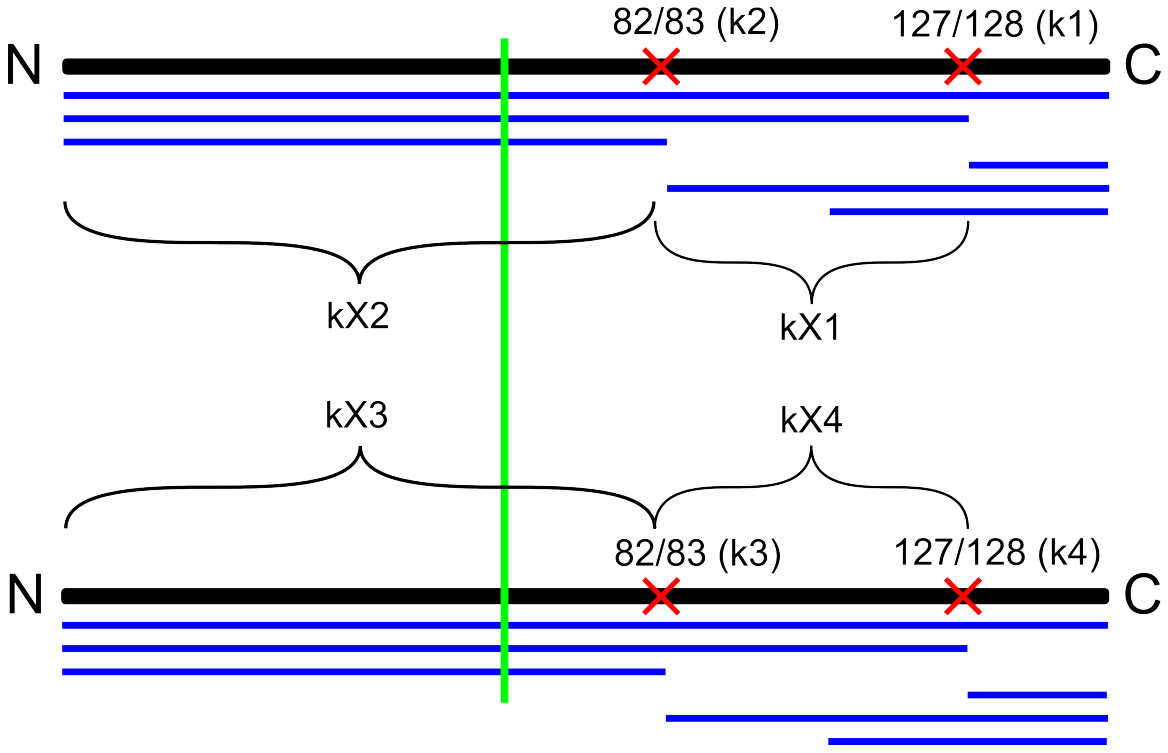


Figure 3.3: Schematic of the Cp149 proteolytic model. Each of the two monomers is represented as a black bar, with N-terminus at the left and C-terminus at the right. The two explicit cleavage locations are indicated with red X, and are labeled according to their designation. The possible cleavage products are shown in blue, and the possible disulfide is in green: any cleavage product from the first monomer could be connected to any cleavage product from the second, so long as it intersects the indicated disulfide. The peptides cannot participate in the disulfide bond. The three regions of implicit cleavages are also shown: k_{x1} cleavages occur between 83 and 127 of the first monomer, etc. These allow for model testing for unidentified cleavages.

exponential model, which is an adaptation of the proteolysis model of Vorob'ev [100]: the probability of observing a particular mass is the product of the probability of producing it and the probability that it has not yet been cleaved by any internal site. The probability of a mass being present is the product of producing a cleavage (or that cleavage being pre-existing) for both termini:

$$P(produced) = P(\mathrm{N}_{cleaved} \cap \mathrm{C}_{cleaved}) \tag{3.1}$$

$$P(\mathrm{N}_{cleaved}) = (1 - e^{-k_{N,cleave}}) + P(\mathrm{N}_{cleaved,t=0}) \tag{3.2}$$

$$P(survive) = e^{-t\sum_i k_i} \tag{3.3}$$

$$P(exists) = P(produced \cap survive) \tag{3.4}$$

In equation 3.2, the same form can be written for the C-terminus cleavage. The k_i constant in equation 3.3 is the cleavage constant for a cleavage site location within the given protein fragment. Terminal products (peptides) have no internal cleavage sites by definition, and this survival factor becomes 1. The only deviation between these equations and previously-published forms is that they take into account boundary conditions at the protein termini, which are non-trivial for smaller proteins. In such situations, or if the protein contains an already-nicked backbone, $P(cleaved) = 1$ for the appropriate terminus. For the Cp149 system, there are four explicit cleavages $82/83_1$, $82/83_2$, $127/128_1$, and $127/128_2$, (termed k_2, k_3, k_1, and k_4, respectively) but to account for undetected cleavages, four additional cleavage constants were defined: k_{x1}, k_{x2}, k_{x3}, and k_{x4}. These constants represent cleavages present within the terminal protein fragments of each monomer: residues 83-127 of monomer 1, 1-82 of monomer

1, 1-82 of monomer 2, and 83-127 of monomer 2. No constants are given for the terminal region 128-149 because it contains no trypsin sites.

Data Fitting

This kinetic model results in a very large number of parameters, so judicious care should be taken when using it to fit parameters to data. There are 4 normal kinetic constants, 4 additional constants to fit hidden cleavages, and 10 isobaric sets, each of which has an offset and slope factor: this produces 28 possible variables for data fitting! Overparameterization is assured unless some variables can be fixed at reasonable limits. For example, most of the offset factors can be set to 0 when using deconvoluted data, and the 4 hidden-cleavage constants can be left at 0 except for experimental simulations. These two changes alone will reduce the number of variables to 14. The slope factors also tend to be reliable: the intact protein signal response is bounded by the start of the reaction, and it can be calculated via standard curves. For species which appear then disappear later in the reactions, the slope factor can also be computed with relative certainty. However, even with only 8-12 real variables, the system may still be overparameterized, and data-fitting should be performed with many initial estimates to sample the full space of possible solutions.

One compounding problem for this data fitting process is the inability to pre-set certain values for the slope factor in the signal-response equations. In theory, these factors should be very reliable once calculated. However, the cleanliness and stability of the micrOTOF has been insufficient to make such assumptions in the past. Considering the substantial improvements in those metrics in the very recent past, future experiments should be able to assume consistent signal response on a day-to-day basis. If so, that would reduce the parameters from the total of 28 down to just

4-8, and with 10 masses observed in every time point, very accurate determinations of the kinetic constants should be possible.

Results and Discussion

Regardless of possible instability in the data fitting process, the raw data from the kinetic hydrolysis experiments shows a very clear general trends. Without HAP or zinc present, dimer is readily degraded over the course of several hours, producing 149-127 fragments and the corresponding 128-149 peptide (figure 3.4a,b). The 149-127 fragment appears relatively rapidly, then decreases with time as it is replaced with the 127-127 fragment, produced as the result of a second cleavage. However, there is no evidence for activity at the 82/83 cleavage site: neither 149-82, 82-82, or any other related product shows up in significant quantity. In contrast, capsid shows substantial production of the 149-82 fragment, which has the fastest initial kinetics. The 82-82 fragment also appears rapidly, and is near completion by 300 minutes. Since that particular terminal product requires multiple cleavages at the 82/83 site (one per monomer), it suggests that the capsid has a very high degree of exposure for the 82/83 cleavage location, which is situated near the top of the tower on the dimer structure. The capsid proteolysis trends are also interesting in that the intact protein decreases in intensity with time, but not with a typical exponential decay (Figure 3.4c,d). Instead, there is a slow linear decrease with time. There are two possible explanations for this. First, capsid proteolysis may be highly sequential, with an initial cleavage leading to a cascade of secondary cleavages. An alternative explanation is that the pool of intact protein is not homogeneous: if a small portion of capsid is responsible for nearly all of the observed proteolysis, then full production of products would be observed (as with the 82-82 fragment) with a relatively minimal decrease in the overall intact protein levels. Because of the signal response factor

problem of mass spectrometry, it is impossible to say if this is the case. However, both the sequential model and the population-partitioning model would fit well with overall knowledge of Cp149 capsid structure and assembly. The network of weak interactions which help to maintain capsid stability may fail to protect a subunit once it has been cleaved or adopts a highly unfolded conformation. A distribution of varying structures within the pool of capsids would also be easy to achieve. If some capsids have a post-translational modification such as oxidation or small molecule adducts, they could be protected from cleavage via steric clashes. The apex of the tower structure is known to be amenable to nearly any artificial insertion without disrupting particles, and any modification in that region could block the 82/82 site from proteolysis.

Influences of HAP12

In the presence of 17.89 μM HAP12, all kinetic hydrolysis trends are dramatically altered, especially for dimer. The most obvious change is an enormous decrease in the disappearance rate for the intact protein: without HAP12, approximately 50% of the intact protein is consumed within the first two hours of the reaction, and the 149-127 fragment has become the dominant species. However, stoichiometric levels of HAP12 cause the intact protein to persist for up to 700 minutes. At the same time, cleavage products are also substantially reduced, which rules out the possibility for dramatically accelerated secondary cleavages. The slow decline of the intact protein bears striking similarity to the linear decrease in the capsid protein under control conditions, which might suggest that HAP12 has caused the dimer to rapidly assemble. This possibility can be ruled out due to the pattern of the cleavages: all capsid reactions, including reactions in the presence of HAP12, have a strong signal which arises from the 82-82 terminal cleavage product. Despite the

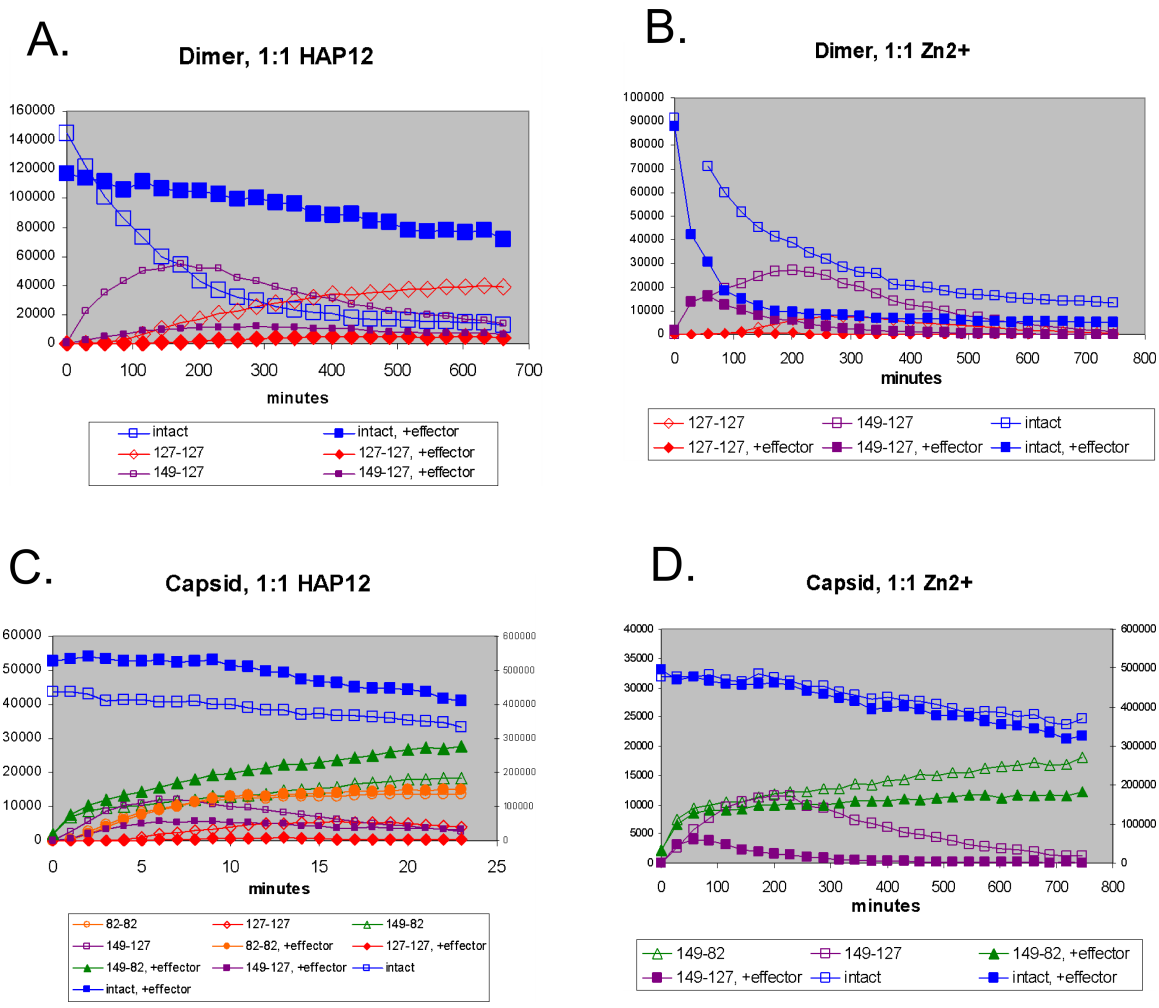


Figure 3.4: HAP12 and $Zn^{(2+)}$ have substantial effects of Cp149 dynamics. Intensities of the deconvoluted MS signals are shown for the primary fragments: low intensity fragments are not shown. The capsid plots show the intact protein on a separate (right) axis due to very high intensity. All other trends are shown on the left axes. Experiments with effector present are filled markers, while corresponding controls are with empty markers. In plot D, the 82-82 fragment is omitted for clarity, but shows no substantial change relative to the normal behavior, as in C.

slowed proteolysis, dimer does not display any significant production of this protein, which shows continued protection at the top of the towers. Likewise, the 149-82 fragment cannot be seen in dimer reactions, with or without HAP12. While this does not rule out the possibility that HAP12 has induced a substantial change in the dimer structure, and perhaps caused nucleation of trimers of dimers or other such possible intermediates, but it does rule out the possibility that HAP12 has caused the dimer solutions to assemble into native-type capsid structures.

As previously discussed, fitting this data with the kinetic model is not a highly robust process. However, with sufficient care some approximate rate constants can be extracted. In general, the proteolysis trends from dimer tend to be better fits to the model, which may be a reflection of cleavages occurring more independently in dimer solutions, as compared to sequentially (with allosteric-like trends) for capsid reactions. The data presented in figure 3.4 matched the kinetic model well, and gave $k_{observed}$ constants for the 127/128 site of 0.0036 min^{-1} for the dimer, and 0.0024 min^{-1} for the capsid (note that these values are in EX2 conditions, with 1.5 μg/mL of trypsin. Visually, these values may seem counterintuitive, but the apparent proteolysis rate constant for capsids is slowed via cleavage at the 82/83 site, which is internal for all 127/128 cleavage products. Upon addition of HAP12, the dimer's cleavage rate constant drops to 0.0008 min^{-1}, and the capsid likewise decreases to 0.0007 min^{-1}. Although it is premature to say that the similarity in these values implies a similar structure in the C-terminus, it does indicate that there may be a connection, as was seen with exposure equilibria as a result of temperature dependence. Also note that the 82/83 cleavage site is not substantially perturbed. The 82-82 terminal fragment observed in capsid reactions is nearly identical between experimental and control reactions. If HAP12 induced a sequential cleavage cascade, the 82-82 fragment would be expected to increase due to the availability of precursor material. The 149-82 frag-

ment is also minimally different in the presence of HAP12, and the slight increase in the experimental reactions can be attributed to a decrease in cleavage at the internal 127/128 cleavage site. Taken together, these observations suggest that HAP12 has very dramatic effect upon the C-terminus but minimal impact on the dimer tower. The kinetics of proteolysis are also very "balanced" in these reactions. Comparing control to HAP12 reactions, the different fragments have different cleavage constants but overall behavior is matched: for example in this situation the assembly effector does not dramatically increase the 127/128 cleavage while decreasing the 82/83 cleavage activity, which would induce production-peak shifts on the time axis.

Influence of Zinc

Both HAP12 and zinc are potent assembly activators for Cp149, but while HAP12 decreases the dynamics of motion, zinc seems to dramatically increase it. As shown in figure 3.4, zinc causes rapid cleavage of the intact protein, such that over 50% is consumed within the first hour. The same effect is not seen with capsid: the slow linear decline is approximately the same between experimental and control reactions, although the reactions in the presence of Zn^{2+} may have a very minor rate increase. However, as was the case with the control capsid reactions, the rate constants do not match between the intact capsid protein and the proteolysis fragments, which may be further support that proteolysis is taking place only on a sub-set of the total capsid protein. In figure 3.4d, the data for the 82-82 fragment is not shown, since it closely overlays the 149-82 data: it retains the behavior seen in HAP12 reactions, with very minimal change upon addition of zinc. The relatively subtle visual changes in the 149-127 trend, as a result of HAP12, are now dramatic and severe. The initial phase of 149-127 production appears to be the same in control and experimental, but an unknown secondary cleavage rapidly reduces the concentration. Since the 82-82 and

149-82 fragments display minimal change, this implies that unknown cleavages in the 83-127 region are now highly active. Unfortunately, no protein or peptide signals have been found to account for that hypothesis.

Very similar behavior is seen for dimer reactions in the presence of zinc. The 149-127 fragment disappears extremely rapidly from an unknown source, but dimer also displays corroborating evidence in the form of the 127-127 fragments. Compared to the normal trends (3.4), the 127-127 fragment is nearly nonexistent in all time points. This type of behavior would be expected for a highly active cleavage somewhere within the body of 127-127, but unlike the capsid reactions, there is no 82-82 or 149-82 to help narrow the possible locations of proteolysis. Although it is attractive to think it is found in the same locations as in capsid (likely between residues 83 and 127), there is no evidence to support a specific claim of that nature.

Conclusions

The results of proteolysis reactions in the presence of HAP12 and zinc are extremely exciting, considering the differing effects on dynamics for two compounds which both promote assembly. Although the effects and magnitudes vary, with zinc greatly increasing proteolysis and HAP compounds decreasing the cleavage rate, there can be no doubt that the compounds have very substantial and widespread effects upon the dynamics, which supports the hypothesis that HAP12 and zinc function as allosteric effectors. The cleavage rates, especially at the C-terminus for HAP12 and N-terminal of residue 127 for Zn^{2+}, undergo dramatic changes in the presence of effectors. The decrease in cleavage for HAP12 could be explained via steric hindrance, but the putative binding location of the HAP compounds is distant to the cleavage site: HAP binds between residues I105, F25, and Y118, while cleavage is on the op-

posite end of the C-terminal helix. At best, stabilization of the C-terminal loop is the most substantial effect which could be directly attributed to HAP binding, without invoking the possibility of much larger overall conformational changes in the body of the dimer. As for zinc, there is no model to date to explain the enormous increase in proteolysis. The inability to identify the cleavage locations is a huge impediment to a full understanding of zinc-induced structural changes, which is why efforts are ongoing to identify all products of the proteolysis reactions.

One promising approach for more thorough mass spectrometry signal identification involves detecting relevant signals based upon their kinetic trends, rather than mass-matching or intensity. To conduct this analysis, all time-series spectra for dimer and capsid, control and experimental, are collected into a sequentially-ordered dataset: this consists of approximately 120 spectra. Then each individual Da value (the corresponding mass to m/z but in deconvoluted spectra) is plotted versus time and normalized. This provides about 100,000 to 500,000 unique trends, each with 120 samples on the x-axis. The 500,000 trends are then clustered with an iterative algorithm into a full tree, and the tree is cropped to produce 50 to 100 clusters. Each cluster thus represents one form of kinetic trend for a given mass. Examples of the resulting plots are shown in figure 3.5.

From plots such as these it is easy to visually identify masses which have very different behavior between dimer and capsid, or between control and drug-treated Cp149. After selection of the relevant clusters, the clusters are mapped back onto the original spectra to show the source of the signals. In many cases, the original masses are present as very low intensity signals. Using this approach, many promising masses have been identified, and they are currently being screened against a library of possible adducts and cleavages to identify their source on the Cp149 protein.

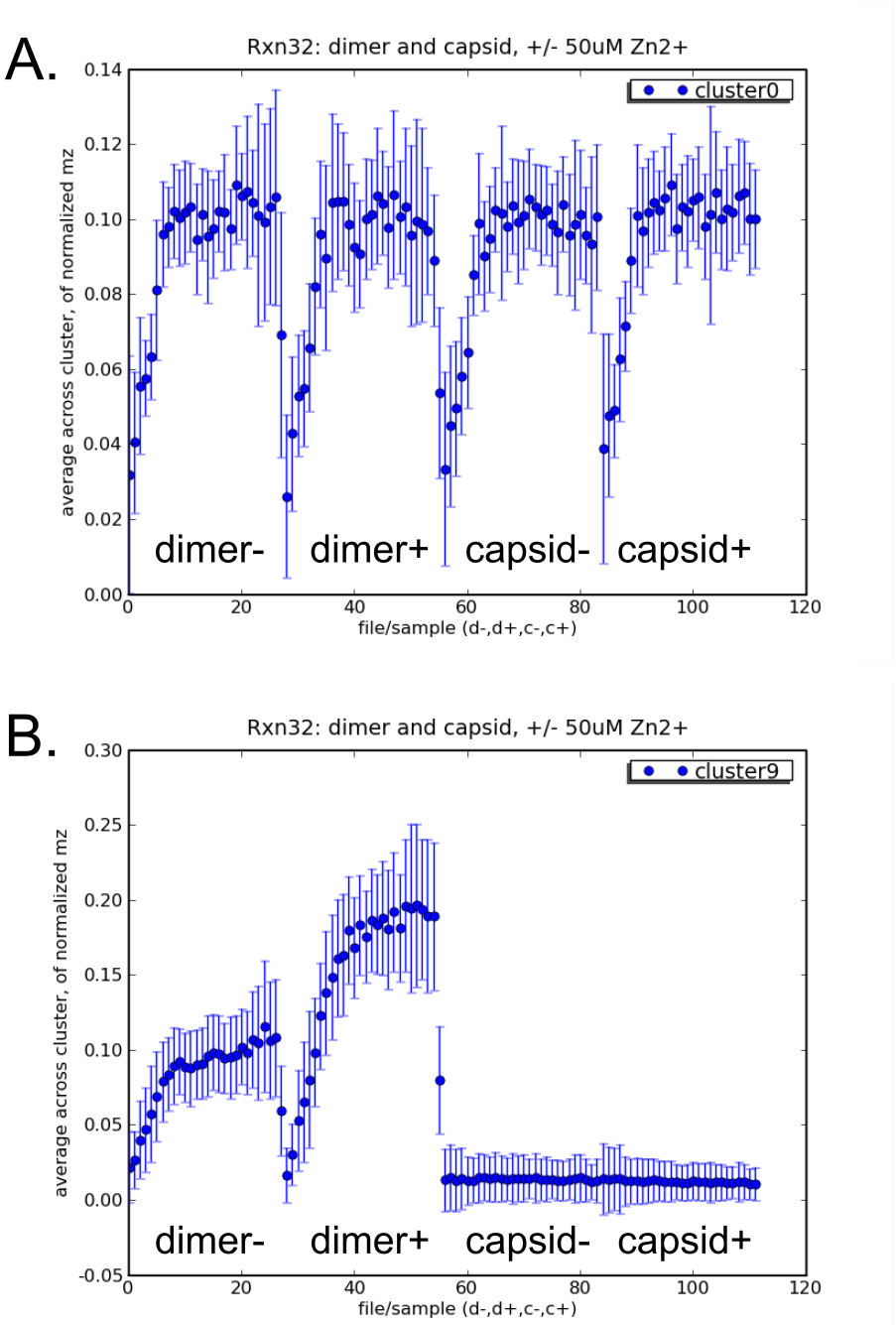


Figure 3.5: Clustering analysis of mass spectrometry signals reveals regulated but unknown masses. A) An unknown mass cluster with rapid kinetic behavior but lack of any response between dimer, capsid, experimental, and control. B) A regulated mass cluster shows no appearance in the capsid solutions, and sensitivity to zinc in dimer solutions.

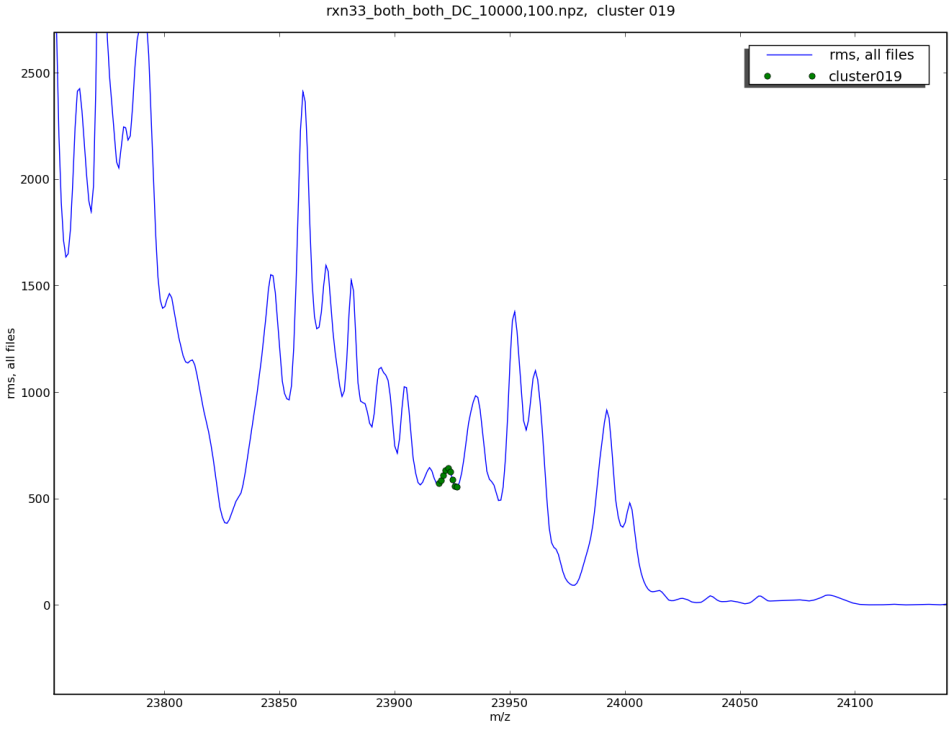


Figure 3.6: Identification of masses from clusters. After clusters have been manually evaluated for interesting behavior, they can be mapped back to the original spectra. The indicated series of points from one cluster (green circles) clearly comes from one low-intensity peak. This particular peak is only present in dimer, and is up-regulated in response to zinc. However, without the clustering analysis it would never be identified due to the low signal intensity and the close overlap with more-intense signals.

HX ASSAYS OF ASSEMBLY EFFECTORS: SMALL-SCALE DYNAMICS

Introduction and Background

The application of kinetic hydrolysis is invaluable for the ability to assign specific locations to dynamic motion, and because protease cleavage is an excellent probe of large-scale dynamic motions. However, proteolysis may be completely insensitive to smaller-scale fluctuations due to the large physical distance between the open and closed conformations: if a 12 Å motion is require for protease cleavage, then kinetic hydrolysis will be silent to 10 Å motions, despite the possible significance of such dynamics (chapter 1, section 1). Our results with kinetic hydrolysis of Cp149 in response to temperature and assembly effectors have shown that Cp149 is capable of very large scale dynamics, but it does not answer the question as to whether there are alterations in the more ubiquitous smaller-scale dynamics. To address this issue, we have used hydrogen-deuterium exchange to probe the entire protein backbone at a much smaller scale. The results from such experiments can be used to further develop the systematic model of Cp149 dynamics as they relate to complex assembly and drug or effector interactions.

Back-Exchange

A general background of the utility of HX can be found in chapter 1, but a more detailed treatment is necessary to actually carry out such experiments. The unavoidable problem with HX derives from the fact that deuteration of labile protons is a reversible process. Therefor, if a protein is placed into D_2O it will begin to assimilate deuterium, especially at the amide protons, but those deuterium labels will be lost if the protein is again exposed to water. For analysis this represents

a serious problem, since it results in a decrease in the mass shift, and the rates of deuterium loss (known as "back-exchange") can be difficult to predict or calculated.

Therefore, precautions must be taken to minimize back-exchange. The two most common ways to accomplish this involve dropping the pH to approximately 2.7, and keeping the protein chilled as cold as possible. Reduced temperatures slow the overall exchange rates, and low pH assists because HX is an acid- and base-catalyzed process: at pH 2.7 the two different pathways for exchange are minimized.

Regardless of the solution conditions, back-exchange will continue to take place at a significant rate. For this reason rapid analysis is critical, to minimize analysis-induced loss of deuteration signals. Accomplishing extremely rapid analyses necessitates careful optimization of the analysis protocol to achieve both speed and efficient analyte processing.

Theory

Assuming that deuterium labeling remains a first-order process over the course of the experiment, the probability of labeling a particular labile-proton site (or the percentage labeled for a population of sites) can be described with the following equations:

$$D_{forward} = D_s(1 - e^{-(kt)_f}) \tag{4.1}$$

$$D_f = (D_{LC} - D_{forward})(1 - e^{-(kt)_b}) + D_{forward} \tag{4.2}$$

$$D_f = (D_{LC} - D_s(1 - e^{-(kt)_f}))(1 - e^{-(kt)_b}) + D_s(1 - e^{-(kt)_f}) \tag{4.3}$$

$$D_f = D_{LC} - D_s(1 - e^{-(kt)_f}) - D_{LC}e^{-(kt)_r} + D_s e^{-(kt)_b}(1 - e^{-(kt)_f}) + D_s(1 - e^{-(kt)_f}) \tag{4.4}$$

$$D_f = D_{LC} - D_{LC}e^{-(kt)_b} + D_{solution}e^{-(kt)_b}(1 - e^{-(kt)_f}) \tag{4.5}$$

$$D_f = D_{LC}(1 - e^{-(kt)_b}) + D_s e^{-(kt)_b}(1 - e^{-(kt)_f}) \tag{4.6}$$

$$D_{f,final} = D_{LC}P(back) + D_{solution}(1 - P(back))P(forward) \tag{4.7}$$

In these equations, $D_{forward}$ is the probability of labeling a site with a deuterium during the forward exchange step of the experiment. D_f is the probability of the site remaining labeled by the end of the analysis and back-exchange. D_{LC} is the percentage of deuterium (relative to labile hydrogen) present in the analysis solutions, and D_s is the percentage of deuterium present in the forward-exchange solvent conditions. The simplified final equation confirms intuition: the probability of finding a deuterium, having started without one, is the sum of two different pathways. The first pathway is the probability of a new H/D swap during the back-exchange period. The second pathways is the probability of first obtaining a deuterium during the normal forward-exchange time, then not losing it to a reverse swap in the back-exchange step. The equation can also be rearranged to group exponents together:

$$D_f = D_{LC} + (D_s - D_{LC})e^{-(kt)_b} - D_s e^{-(kt)_f - (kt)_b} \tag{4.8}$$

From this equation it can be seen that the observed deuterium measurement depends on three factors: the easily-computed deuterium level in the analysis step, an exponential that involves the time spent in the analysis step, and a double exponential which involves both the initial deuterium incorporation stage and the analysis step. Unfortunately, we are generally only interested in the first deuterium incorporation step: the analysis step does not provide useful information, since it usually involves protein denaturation and irrelevant conditions such as low pH and cold temperatures.

If the system in question involved just a single hydrogen/deuterium binding site, the equations could be simplified. The incorporation of deuterium in the forward stage would be $(D_{final} - D_{LC}P(back))(D_{solution}(1 - P(back)))^{-1}$, and the $P(back)$ factor could be experimentally estimated. As a function of time. However, for proteins with many labile protons, the summation factors introduced by the inability to differentiate between protons further complicates the situation. For a protein:

$$\sum_n D_{f,n} = nD_{LC} + n(D_s - D_{LC})\sum_n e^{-(kt)_b} - nD_s\sum_n e^{-(kt)_f-(kt)_b} \tag{4.9}$$

Mathematically, the summation of the final term cannot be split, and it represents an ill-posed problem: without more information, there are very many equally valid solutions. From a biochemistry perspective, it can be explained more simply: it is difficult to know the correlation between a proton's environment when the protein is folded ($e^{-(kt)_f}$) and when it is unfolded (($e^{-(kt)_b}$)) unless the protein is already well-characterized. It is possible that a well-protected proton site remains that way in organic solvents, or it could become more exposed. Likewise, an exposed proton could become trapped in a molten globule state when placed into the same organic solvent. The sum total of those two proton examples could be difficult to distinguish from a second situation, in which two moderately-exposed protons remain that way in organic solvent.

However, there are some ways in which this system can still be usable. By making some assumptions, the ill-posed nature can be avoided to give meaningful solutions. The first possible assumption is that fast-exchanging protons in native conditions remain fast-exchanging in denaturing conditions, or at least that the rank (in order of exchange speed) is maintained between the two states. We refer to this as fulfilling Wilcoxon criteria, and it provides very useful results. If the fastest exchanging protons in the forward step remain the fastest exchangers in the back-exchange step, then

deuterium incorporation follows a first-in-first-out pattern, and back-exchange will always be maximized, since the fastest back-exchanging protons are also the first to be incorporated (back-exchange of a proton which has not yet swapped for a deuteron has no effect). An alternative assumption is that the ranks between forward and back-exchange are nearly randomized. In these conditions, back-exchange has a more mild effect, and the observed deuterium incorporation is closer to the deuterium incorporation from just the forward-exchange step of the reaction: the back-exchange is proportional to the amount of forward exchange. From a theoretical basis, it is difficult to justify the random-rank criteria, since protein denaturation is not a random process.

The third way to deal with the forward/back rank problem is to ignore it entirely. Although back-exchange is difficult to directly correlate to the forward exchange, using either the Wilcoxon assumption or random rank assumption, the net effects of forward and back exchange will be more heavily biased for forward exchange (there is a net gain of deuterium, starting with native protein). Therefore, uncorrected HX measurements will still be a probe for small-scale dynamics in the forward reaction, regardless of signal dilution and interferences. Preliminary experiments with Cp149 used this approach, although control experiments were also conducted to estimate the back-exchange.

Chromatography Methods

The size-exclusion chromatography approach for analysis of Cp149 proteolysis reactions is completely unsuitable for hydrogen-deuterium exchange measurements. The maximum flow rate possible for the BioSep column at room temperature is approximately 0.1 mL/min. With increased back pressure at lower temperatures, the flow rate must be dropped even further, which linearly scales the time required to elute

protein. Since at least 3 minutes are required for protein elution with the maximum possible flow rates, any additional increases in chromatography time are unacceptable. In order to achieve the fastest possible chromatography times, a macrotrap C4 column from Michrom Biosciences was used for reverse-phase chromatography. These columns have extremely low backpressure as a result of minimal bed size and wide bore, and can loaded and eluted manually via syringes if desired. However, they also work very well with HPLC, and tolerate 1 mL/min flow rates and higher. To minimize carry-over and to reduce the time required for protein elution, the column was equilibrated and loaded with high concentrations of organic. Previous carry-over tests had established a 50/50 mixture of acetonitrile and isopropanol as the optimal solvent for reverse phase chromatography, and tests determined that a gradient beginning at 55% organic solvent was the maximum possible for the Cp149 system. With these conditions, flow-through was minimized and no bleed-through following the wash phase was observed. Protein remained on the column so long as it was held at 55%, but eluted rapidly upon increasing the concentration of organic solvent in the gradient. For reasons that are not fully understood, carry-over is nearly nonexistent with the reverse-phase column was loaded and eluted at high flow rates, such as 0.8 mL/min. Low temperatures tend to worsen the carry-over effect, but even with the column maintained at 0 °C, no carry-over was detected. Due to the small size of the column and the high flow rates, column washing was performed from 15 seconds prior to switching to the source. The gradient percentages were as follows: 55% organic from 0 to 0.25 minutes, linear gradient to 95% organic by 1.0 minutes, hold at 95% organic until 2.0 minutes post-injection, immediate drop to 55% organic at 2.0 minutes, and held at 55% organic until 3.0 minutes. Using this scheme, protein eluted from the C4 column at approximately 45 seconds post-injection. For back-exchange tests the "B" channel solvent contained 10% H_2O/D_2O mixture, so the gradient was

increased to a maximum of 100% channel B by 1.0 minutes in order to maintain equivalent ratios of H_2O/organic mixtures with time.

Protocol

Before the start of every HX reaction, buffers were prepared containing all components except for Cp149. This was the opposite procedure used for proteolysis experiments, since the dilution effect on the NaCl concentration is again a concern. However, while proteases should not be incubated in advance at high-activity pH levels, there is no such concern for deuterium. Theoretically, it should be possible to prepare large stocks of Tris/NaCl/D_2O in advance, this was intentionally avoided to ensure that every prepared reaction was a full experimental replicate. The reaction buffers for HX were a match to proteolysis reaction: 100 mM Tris pH 7.5, 150 mM NaCl, 1% DMSO, 17.89 μM HAP or 50 μM Zn^{2+} for effector reactions, 0.3 mg/mL Cp149, and 50% D_2O. The concentration of D_2O was intentionally kept low for these reactions, for two reasons. Protein stocks for Cp149, especially the dimer form, are not available at high concentrations: losses during purification are too high, and assembly is a constant concern. The relatively low concentrations of protein stocks (1-2 mg/mL, typically) restict the maximum dilution that can be made, and thus the amount of D_2O that can be present in the reaction. By keeping the D_2O concentration low, all reactions, regardless of original protein stock concentration, can be performed in the same conditions.

The other reason for intentionally reduced D_2O percentages is the ability to probe back-exchange. Equation 4.7 shows that the final observed deuterium incorporation has a component that scales linearly with the D_2O percentage in the analysis solvent conditions. When the HPLC system contains more deuterium than the original reac-

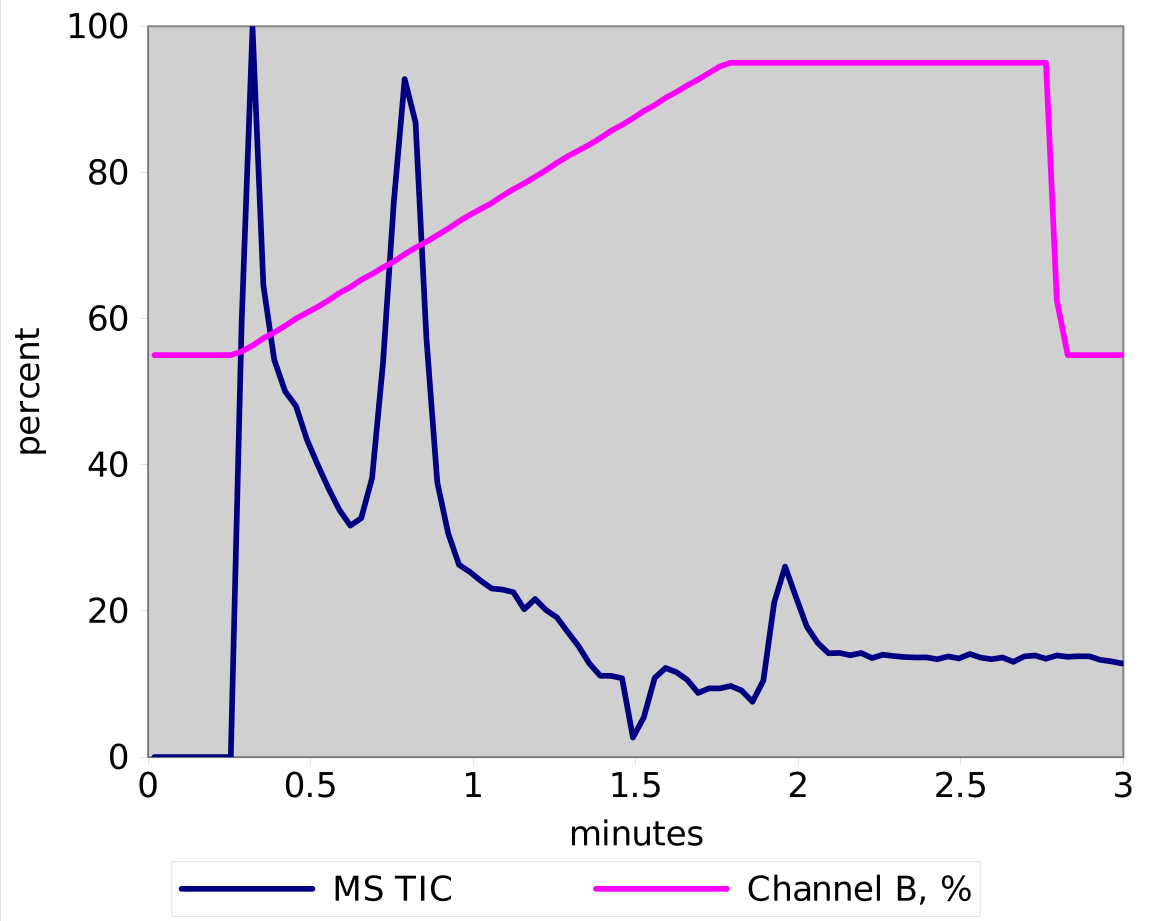


Figure 4.1: Rapid chromatography for HX. The total cycle time is 3 minutes, and protein is within the HPLC system for less than a minute. The elution peak at approximately 0.75 minutes corresponds to the intact protein signal. The peak which decays from waste valve switch at 0.25 minutes to 0.5 minutes is due to small molecules which elute more rapidly.

tion, the "back" exchange actually progresses forward, resulting in more deuterium incorporation. This sets an upper limit for the deuterium incorporated by the forward reaction: it can never be greater than the observed incorporation, so long as the HPLC D_2O levels are greater than the reaction levels. Likewise, lower limits can also be set with HPLC D_2O concentrations which are below the concentrations in the reaction. By keeping the reaction D_2O at 50%, a reasonable compromise is reached which provides sufficient signal quality via deuterium incorporation, while at the same time making it possible to achieve D_2O percentages in the HPLC system either above or below the 50% reaction level. If the protein reaction is carried out in 99% D_2O, it is nearly impossible to create back-exchange conditions which can incorporate additional deuterium, without using extremely pure chemicals, deuterated acids, etc.

Back-Exchange Tests

Based on the above equations, experiments were conducted with deuterated HPLC solvents to determine the extent of back-exchange. In place of the 100% H_2O, 0.1% formic acid (channel A) and 50% ACN, 50% 2-PrOH, 0.1% formic acid (channel B) used in normal experiments, the HPLC was used to deliver 80% H_2O, 20% D_2O, 0.1% formic acid (channel A), and 45% ACN, 45%2-PrOH, 8% H_2O, 2% D_2O, 0.1% formic acid (channel B). A minor correction factor (increasing the D_2O slightly from these exact values) was applied to account for formic acid as a proton donor. The resulting This composition provides a ratio of 4:1 protons to deuterons, regardless of the ratio of solvents A to B. These solutions were used to run several blank gradients to fully equilibrate the C4 column, then a normal injection of the standard Cp149 reaction mixture was performed. However, the reaction mixture contained 0% D_2O, and thus every available proton was available to be exchanged in the HPLC system. This experiment is therefor analogous to what would be seen if 100% deuterated

protein were analyzed in a H_2O-containing HPLC system: the observed deuteration levels represent the maximum possible back exchange that could ever occur, given the operating parameters of the HPLC.

Results

An example of the data collected from a typical HX experiment is show in figure 4.2. The very first sample measurement made following the start of the reaction already has a substantial amount of observed exchange, on the order of 65 to 75 Da. Considering that no back-exchange correction has been applied, this suggests that the true exchange as a result of just the forward reaction is very substantial. Interestingly, there are no visible subpopulations in the raw data (figure 4.2a), showing that the protein is behaving in a homogeneous fashion. In order to better determine if any subpopulations are present, the HX-shifted spectra were deconvoluted by using the reference spectra (collected under identical conditions but without D_2O) as the point-spread-function via Richardson-Lucy deconvolution. This is possible because a HX-shifted spectra is the result of convolving a reference spectra with the deuterium probability-density "blurring" function:

$$Observed = Reference \otimes PDF_{deuterium} \tag{4.10}$$

Since convolution is commutative, either the $Reference$ spectra or the PDF can be treated as the point-spread-function in the Richardson-Lucy algorithm. Although the deuterium PDF is the true point-spread-function in an HX reaction, it is the unknown factor computationally, and the reference spectra is easily obtained experimentally. Further details and an implementation of the Richardson-Lucy algorithm are provided in the appendix.

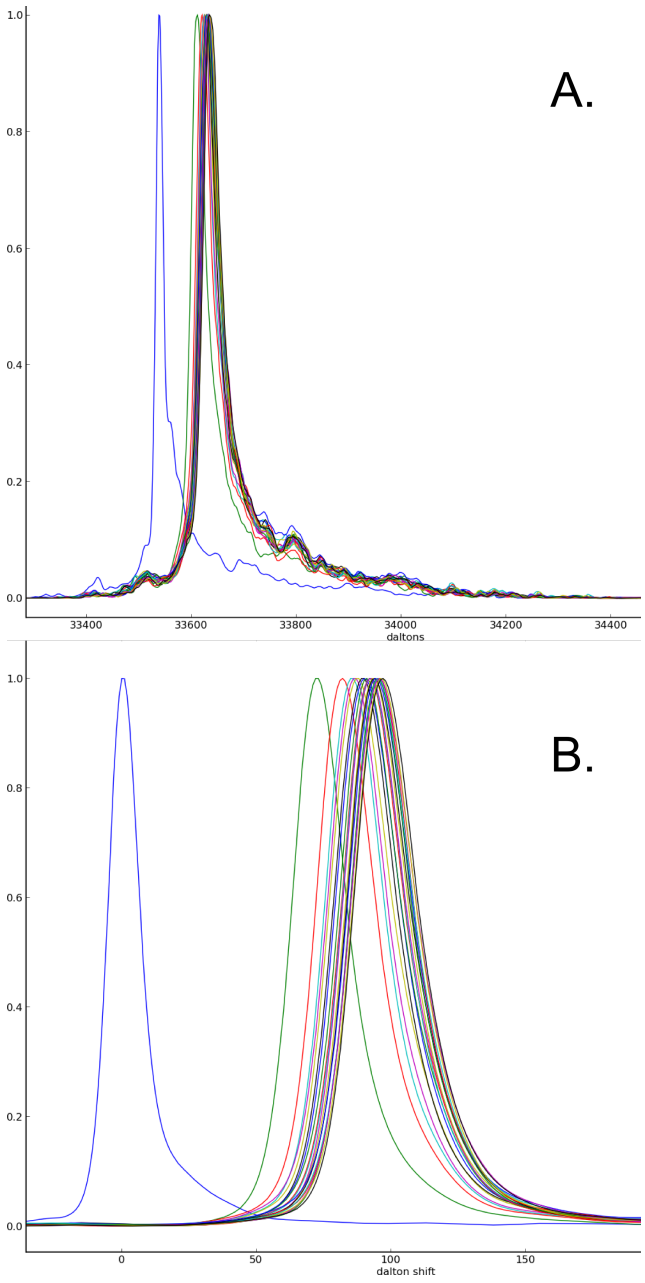


Figure 4.2: Intact protein mass additions as a result of HX. A) The general shape of the Cp149 intact protein signal (with adducts etc) is maintained as deuterium is added to the protein backbone. A very rapid mass shift is followed by much slower further mass addition. B) Traditional signal-processing deconvolution is applied to observe the HX pdf: the probability-density-functions for HX addition are shown for the spectra seen in A. The HX pdf spectra are the result of 10 iterations of Richardson-Lucy by using the reference spectra in place of the point-spread-function.

Signal deconvolution, which should not to be confused with mass spectrometry charge-deconvolution, was found to be an invaluable tool for the study of HX. The large numbers of spectra acquired in the course of testing both dimer and capsid in the presence of drugs could not be processed in an efficient manner without a high degree of automation. However, intrinsic signal noise complicates the ability to use explicit calculations for the determination of mass shifts. For example, determination of peak center is a complicated process and highly influenced by decisions regarding peak cutoff limits (chapter 6). With signals such as those seen in figure 4.2a, even slight signal noise near the peak cutoff limit will have a huge effect on the peak centroid calculation. Peak apex is similarly unsuitable: random noise throughout the peak makes any one single data point unreliable as a reference point. One solution to this problem is to use a smoothing function to remove noise from the signal peak, followed by an apex determination. This could be a valid approach, but protein mass spectra tend to be asymmetric, with adducts but not losses, and symmetric smoothing functions have the effect of shifting the smoothed peaks to a more symmetric shape. Protein spectra peaks which have been smoothed with a filter sufficiently large to remove noise also tend to be biased. For HX experiments, this causes the calculated rate of deuterium addition to be systematically high. In contrast, deconvolution is a nearly perfect solution: noise is explicitly accounted for, with the algorithm output being analogous to a best-case data fit, multiple overlapping HX distributions can be resolved, and it is amenable to high-throughput automated data processing. The use of automated deconvolution also helped to confirm the stability of the HX assay: both dimer and capsid proteins produced nearly identical trends, with no exceptional artifacts in either control or experimental.

The same deconvolution technique was applied to the back-exchange tests. One caveat of this approach is that it is impossible to have truly identical reference/ex-

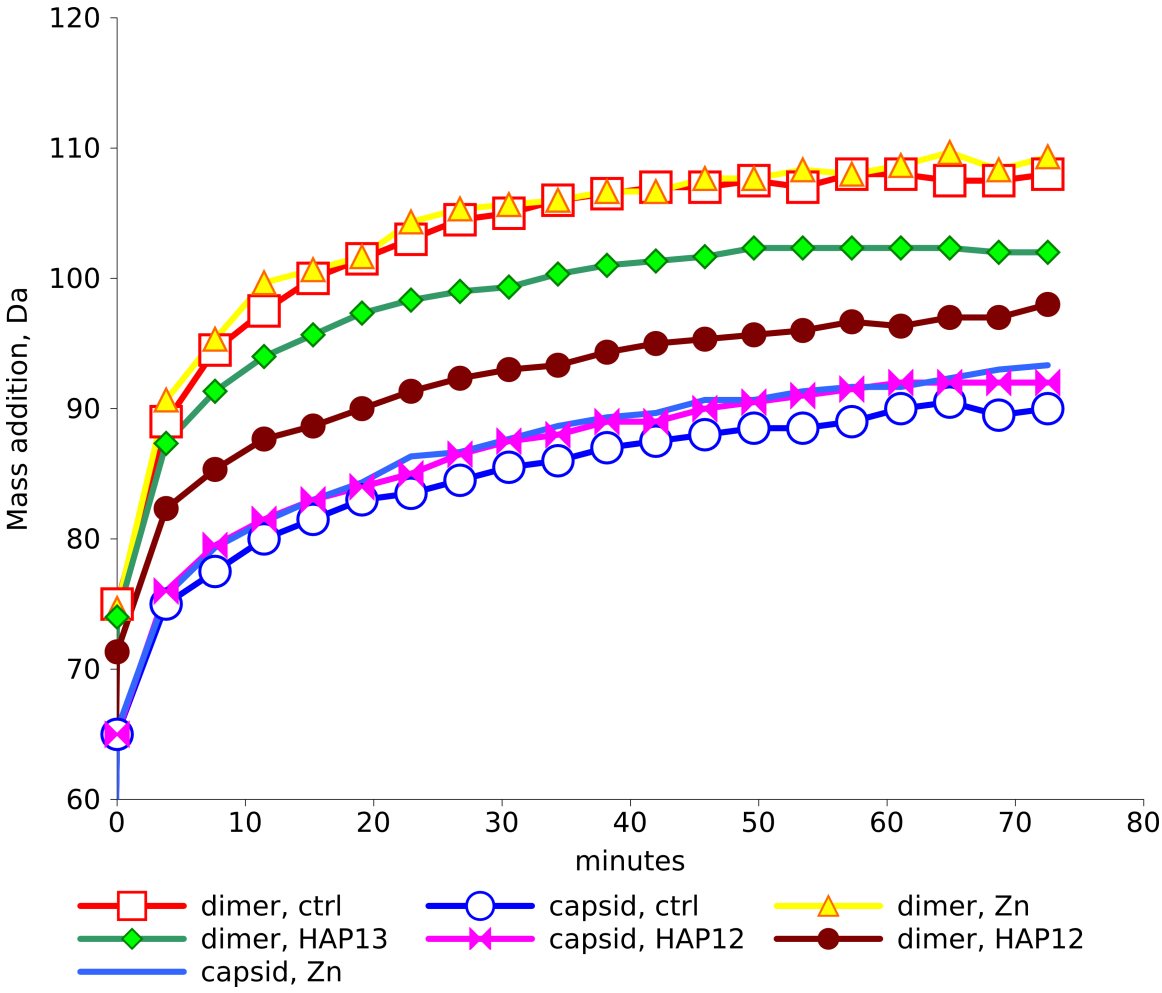


Figure 4.3: HX mass shifts of Cp149 in response to assembly effectors. All plots are the result of averaging multiple runs, and errors are typically ∼1.5 Da. Dimer shows more overall shift, and capsid the least. In the presence of HAP12 or Zn^{-2}, the mass shift for capsid does not change. Zinc also has no effect on dimer, but HAP13 and HAP12 cause a decrease in the exchange rate.

perimental spectra: performing HPLC analysis with deuterated solvents necessitates a separate instrument setup (different solvent bottles, etc) and different solvents, and as a result, the background signals and noise levels are slightly different between deuterated analyses and normal water-based HPLC analyses. For these reasons our observations err on the cautious side when interpreting the possibility of secondary HX distributions in the back-exchange tests, etc. That said, the back-exchange analyses look effectively identical to the normal reference spectra, but with the same homogeneous HX shift. For capsid, this amounts to a total of 198 Da of mass shift for 100% D_2O. The back-exchange mass shift for dimer was similar: 184 Da. Considering the overall size of the Cp149 dimer unit, these values indicate that overall back-exchange is very low. $Cp149_2$ has approximately 300 backbone amide protons capable of exchange, and the labile side-chain protons compose another large pool of protons of similar magnitude. One standard practice in HX is to assume that all side-chain deuterium incorporation is lost upon back-exchange because the side-chain exchange rates are much higher on average. Although we cannot confidently assign all of the observed back-exchange to side-chain HX, the overall numbers suggest that back-exchange in these conditions has a minimal effect upon the amide protons, since only about 25% of the theoretically-exchangeable protons show deuterium incorporation in the back-exchange experiments.

Discussion

The overall results of the HX assays, even without considering the effects of HAP or zinc, have interesting implications when compared to the proteolysis data. The basic trend of the HX results is one of systematic stability and straightforward trends: both capsid and dimer produced homogeneous shifts as a result of HX, with no

indication of sub-populations having different deuterium incorporation levels or rates. In some ways this is surprising considering the proteolysis evidence which may suggest heterogeneous populations of capsid. Also, every condition tested produced roughly the same results: a rapid initial deuterium incorporation phase was followed by a much reduced rate, and after 1 hour the mass shift was approximately 100 Da.

However, this should not be taken to imply that HX proves that Cp149 dimer and capsid are identical. Although the trend are generally the same, there are definite and significant differences in the magnitude of deuterium incorporation, both between dimer and capsid and as a result of HAP compounds. Dimer consistently produces a greater amount of mass shift, topping out at approximately 108 Da after one hour. Capsid has a mass increase of only about 90 Da by that time, and the residual rate is slightly higher than seen in the dimer trends (figure 4.3). These mass differences are the result of 50% D_2O incubation, so identical (theoretical) experiments carried out in 100% D_2O would show double the absolute mass difference. The lack of a back-exchange correction may also mask differences from assembly state or effectors, although the small magnitude of the back-exchange estimates, and the similarity between capsid and dimer, suggests that back-exchange has a heavy bias towards the extremely rapid exchangers. These rapid-HX protons are typically from the residue sidechains and would not be expected to be a good reporter of protein backbone dynamics, so the remaining signal (attributed heavily to backbone protons) most likely captures most of the key information from assembly state and drugs.

It is immediately apparent that Zn^{2+} has little to no effect on the backbone dynamics of either dimer or capsid. Dimer plus zinc is completely indistinguishable from control reactions (figure 4.3), and capsid plus zinc has an increase of only 2 Da. These results are in stark contrast to the proteolysis results, in which zinc substantially increased proteolysis cleavage through an unknown mechanism. Current

dogma regarding proteolysis and HX would therefor suggest that zinc induces large structural changes, but without increasing the overall solvent accessibility of the protein structure. Considering the highly helical and extended structure of the Cp149 dimer unit, such a hypothesis is not unthinkable, but the extreme similarities between control and experimental assays for both dimer and capsid are difficult to reconcile with the concept of a hugely dynamic protein that retains its small-scale fluctuations with great fidelity.

The proteolysis-slowing effects of the HAP compounds are mirrored in the HX results for the dimer. HAP13 induces a relatively minor decrease in dynamics, but the effect is definitely significant. HAP12 has approximately double the effect of HAP13, decreasing the exchangeable protons by approximately 10%. This effect is especially significant considering the hydrophobic pocket presumed to be occupied by HAP12 is relatively deficient of labile protons compared to more hydrophilic regions of the protein: localized effects are insufficient to account for the change in exchange observed here. In contrast to the dimer reactions, capsid in the presence of HAP12 displays only a very minor increase in exchangeable protons, similar to the magnitude of the effect seen with zinc. Considering the substantial decrease in proteolysis at the 127/128 site in response to HAP12, this increase is unusual. However, the overall proteolysis exposure observed in the initial temperature-dependence study (chapter 2) may reconcile these observations. At room temperature, the C-terminus of Cp149 is very exposed as probed with protease: nearly 1% of the sites populate the cleavage-accessible conformation. Since the capsid crystal structure places the C-terminus is located at the capsid surface, there are substantial steric hindrance effects for protease approach. To achieve the high degree of open conformation despite the close steric hindrance, the C-terminus may adopt a very solvent-exposed and dynamic conformation in solution. By nature of being so solvent-exposed, the C-terminus

could fall into the category of rapid-exchangers, with any deuterium incorporation lost in the back-exchange phase. This would leave the rest of the protein to account for the 2 Da increase in mass upon HAP12 treatment. Since the crystal structures with HAP demonstrate relatively rigid-body rotations of the Cp149 dimer at the capsid surface, it is very feasible that such transitions would be silent in HX assays.

Conclusions

Considering the large changes in dynamics observed via kinetic hydrolysis experiments (approximately 5-fold changes with temperature or assembly effectors), the small changes in hydrogen-deuterium exchange are somewhat surprising. The maximum difference seen between any two HX assays is approximately 20%, and corresponds to the difference between dimer and capsid: no effector tested substantially increased small-scale dynamics, and no change was more pronounced that the effect induced by capsid assembly. However, this should not be taken to imply that the dimer and capsid are static structures, insensitive to HAP or zinc. The results of the proteolysis experiments clearly show that large-scale conformational changes are present in the native structures of both the dimer and the capsid form, so the HX results must be interpreted with this knowledge in mind. This apparent discrepancy between kinetic hydrolysis and HX results can be resolved by considering the nature of each of the two measurements: kinetic hydrolysis has the potential to be sensitive to secondary structure, but only in the absence of other steric constraints. On the other hand, HX has nearly no steric constraints, as it is limited strictly be solvent accessibility and hydrogen bonding.

By placing the HX results into this context, the implications become clear: regardless of the large-scale dynamics of Cp149, the solvent accessibility and hydrogen

bonding are relatively unchanged. For a highly helical and non-globular protein like Cp149, it is easy to envision such a situation. The five alpha helices likely remain intact, but may shift dramatically in relation to one another. Since kinetic hydrolysis has displayed a heavy bias for the C-terminus, combined with the lack of structure in the crystal structure and the known dynamics of the C-terminal region in HBV biology (chapter 5), any gain or loss of secondary structure likely occurs at or beyond the C-terminal alpha-5 helix. This is further supported by the structure of assembled capsid: hydrophobic contacts between the C-terminal helices are the primary driving force of assembly, so an increase in hydrogen bonding or solvent exclusion can be likewise localized to this region. Since zinc has effectively no effect at all on the hydrogen-deuterium exchange measurements, it can be assumed that zinc does not perturb secondary structure at all, and likely binds or interactions with a core region which is silent to HX measurements carried out on this timescale.

Although it is somewhat speculative, it is interesting to note that if all 20% of the HX decrease, when comparing capsid to dimer, is located from residues 110 to 149, then the difference between dimer and capsid structures in that region must be substantial. This region, composed of the α-5 helix and terminal loop, comprise approximately 30% of the protein sequence. The vast majority of this section would need to be highly solvent-exposed in the dimer form, and completely protected in the capsid form, to account for the 20% decrease. The hydrophobic contacts in assembled capsid are relatively minimal, and are insufficient to account for this loss in solvent exposure. The hydrogen bonding in helix α-5 would be another possible source of labile hydrogens, but this would necessitate nearly complete unfolding in the dimer state. If we assume that the dimer does not adopt a completely unfolded C-terminal region, the only alternative is to assign a not-insignificant portion of the HX decrease to another region of the protein structure. However, since HAP binding

and capsid contacts are presumed to be highly localized to the inter-dimer contact region, any decrease in labile protons at more distant locations would necessitate allosteric structural changes. These observations are explained further in the context of structural models in chapter 5.

HBV STRUCTURE AND DYNAMICS

Structure of the HBV Core Protein Dimer

To date, there are just three atomic-level structures for the HBV core protein. The first two structure are both of protein which is assembled into the capsid form: one without any assembly effectors, and the second with HAP-1 bound in the inter-dimer contact region [76]. These structures have been available for several years now, but no similar atomic-level structures have been collected for free dimer, due to the problems with self assembly into capsids or aberrant structures at higher concentrations. Although the dimer structure visible within the capsid model is a good analog for the likely structure in solution, biophysical evidence has long supported a variety of differences between the dimer and capsid conformations (reviewed below). In order to fill this gap in our knowledge regarind HBV structure, the Zlotnick lab has recently collected x-ray density sets from crystallized dimer, which was made possible by using an assembly-defective mutant. Publication of their findings is imminent, but currently all comments regarding the dimer structure are via personal communication with A. Zlotnick. The particular mutant used, Y132A, removes hydrophobic contacts from the inter-dimer region of the capsid, but on isolated dimer the alanine is on the extreme periphery of the α-5 helix, and does not participate in contacts with the rest of the body of the dimer. For these reasons, it is reasonable to assume that the Y132A mutant does not introduce new structural effects.

Although the Y132A mutant is assembly-deficient, it is still capable of forming intact capsids in the presence of *wt* protein, and maintains relatively normal contacts in the hydrophobic inter-dimer region [101]. Perhaps for this reason, the dimer structure shows a trimer of dimers in the asymmetric unit of the crystal cell, with

the normal contacts present but slight distorted. The net result of these distorted contacts is that the normally pseudo-planar trimer is twisted into a helix, with two normal hydrophobic contact patches but with the third broken by separation of the dimer units at the termini of the trimer helix. These contacts produce 4 of the possible 8 contacts with some degree of hydrophobic contact, and this is expected to help induce the structural conformation created by normal capsid formation, if such a transition exists: in the hypothetical absence of these hydrophobic contacts, the free dimer structure should be less "capsid-like".

There are many interesting features to the dimer structure, but the most notable feature is the rotation of the helices around certain invariant regions. The central core of the dimer, termed the chassis, has extremely good alignment values when comparing the dimer structure with the capsid structure (figure 5.1b, blue). Each of the helices, although rotated in orientation, remains highly intact and can be easily overlaid with the capsid structure. However, the center of the towers contain a series of proline and glycine residues which disrupt the vertical helices and form a pivot point (figure 5.1b, green). From this pivot, one monomer in the dimer structure deviates outwards from the other monomer, breaking symmetry and opening the top of the tower (figure 5.1a). The α-5 helix is also rotated slightly outwards from its location in the capsid structure, and this region in particular should be expected to show mediation of the dimer-induced structural changes as a result of packing in the dimer crystal. These structural observations of the dimer, although subtle and still incomplete, provide dramatic evidence and background for both HBV biology, as well as our measurements of dynamics, and are discussed more in detail in the appropriate sections below.

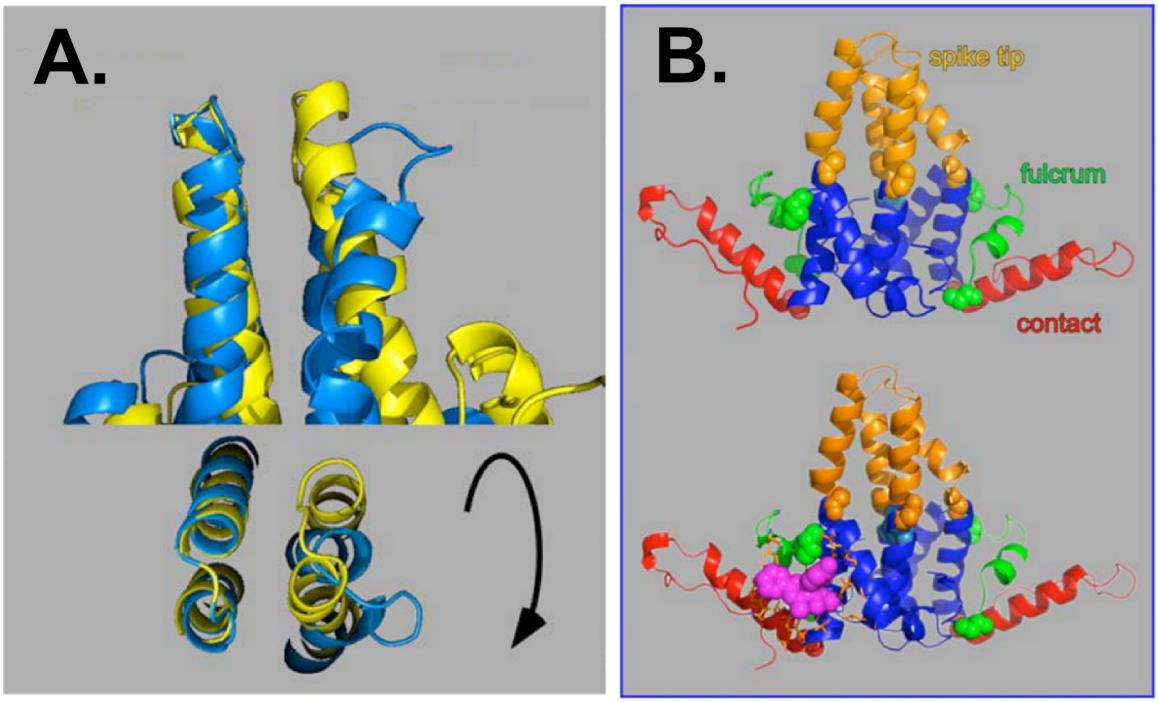


Figure 5.1: Crystal structure of the core protein dimer. Despite the presence of some of the possible inter-dimer hydrophobic contacts present in capsid, the dimer structure displays notable differences compared to capsid. A.) At the top of the tower, one monomer bends away from the other monomer, breaking symmetry with a substantial translocation/rotation. B.) The dimer structure is shown by itself (top) and in the presence of modelled HAP (bottom). The central chassis region (blue) remains very constant between dimer and capsid, but the helices extending away from it are capable of rotations, which are mediated by the fulcum regions. Figure courtesy of A. Zlotnick.

Implied Dynamics of HBV Biology

Our studies of Cp149 in solution have shown that this protein, in both the dimer and the capsid forms, is highly dynamic, existing as an ensemble of conformations. In some ways, our results be unexpected, such as the large equilibrium values observed for the opening transition at the 127/128 cleavage site: the K_{open} constant of approximately 0.01 is much larger than values observed in many proteins. However, in light of the overall knowledge of HBV structure and biology, conformational flexibility, could, or even should, be expected and teh studies presented in this document help bring together a large amount on in vivo and in vitro data present in the literature.

Viral Life Cycle

A full understanding of all the details of HBV replication is hampered by the inability to study the viral lifecycle in culture: it is not possible to stably infect cultured cell lines, and there is a very strict species specificity, so only chimpanzees can be infected with the human-specific hepatitis B virus. The utility of the chimpanzee animal model is further reduced due to the variability of disease progression in both humans and chimpanzees [102]. However, it is possible to infect human hepatoma cell lines (primarily HepG2 and Huh7) with HBV, and follow one round of viral replication. Studies using this approach have been used to achieve a general consensus of the HBV lifecycle. Virion entry to the cell is preceeded by binding to an unknown cellular receptor, possibly transferrin or liver endonexin [103]. The viral envelope then fuses with the cellular membrane at some point during the endocytosis process, and the interior capsid is released into the cytosol [104]. The capsid then binds to the nuclear pore complex, possibly with the assistance of importin beta and alpha, before passing through the pores into the nucleus [105]. The inability to detect assembled

capsids within the nucleus was originally interpreted to imply that capsids dissociated at the NPC, which would be consistent with the difficulty in importing such a large structure intact [103]. However, more recently it was found that the lack of capsid detection within the nucleus is due to the fact that they fully dissociate after passing through the NPC, forming a pool of available dimers which later reassemble [68]. This assembly and disassembly is of particular interest, since it is in apparent contradiction to the observed stability of both capsids and virions [89, 91]. Viral replication proceeds with capsid assembly in the cytoplasm, where it packages RNA, the polymerase, and a kinase (PKC) [105]. The exact effect of the RNA is ambiguous, since some studies have found that it promotes assembly and stabilizes capsid [68, 69], while other suggest that phosphorylation of the capsid protein produces a more pronounced increase in capsid stability, while simultaneously displacing the RNA [106]. The assembled capsid is then targeted to the endoplasmic reticulum via an unknown mechanism, where it associates with similarly-targeted surface antigen proteins (HBsAg) and acquires the outer octahedral protein/lipid layer at some point during cell exit. It is also possible for capsids to bypass cellular export prior to encapsulation with the lipid layer: they can be recycled to the nucleus, where they amplify viral DNA levels. Whether this recycling is targeted and regulated is unknown. Capsids which are exported are almost universally encapsulated by the protein/lipid HBsAg layer, which has the ability to form pleomorphic octahedral empty particles, but adopt an icosahedral conformation when surrounding the inner capsid [107]. Based on size measurements of icosahedral virions and the octahedral HBsAg particles, the inner capsid makes direct contact with the lipid layer via the tops of the capsid towers, which may provide the mechanism for influencing assembly conformation and for providing signal transduction between the layers.

Protein Phosphorylation

In addition to the life cycle signaling provided by capsid assembly and nucleic acid encapsidation, phosphorylation of the core protein provides another layer of regulatory control over key events in the viral life cycle. There are 10 serine residues within the structural body (up to residue 149) of the core protein, and several more within the arginine-rich 34 amino acids of the C-terminus (primarily at S157, 164, and 172) [108]. Both classes, structural and arginine-terminus, have been implicated in structural roles for the capsid and complete virion. The most well-known phosphorylation sites are those within the arginine-rich region. Capsid assembly occurs around RNA and packages with it a viral polymerase and two copies of protein kinase C (PKC) [109]. In native virions, the phosphorylated C-terminus is incapable of directly binding either RNA or DNA [106, 108], but has a strong affinity for RNA when unphosphorylated [68, 108]. This suggests that the initial RNA packaging occurs before the protein has been phosphorylated. Once fully assembled as capsids, there is a competition between the protein kinase and the RNA for the phosphorylation sites, and the eventual phosphorylation causes RNA to dissociate from the C-terminus, which permits reverse transcription to DNA [108]. Although all of the monomer subunits are capable of being phosphorylated, only 50% of them are: this matches the percentage of monomers which can be proteolyzed at the junction between the structural body of the protein and the arginine-rich terminus [110], and supports a model of allosteric motion with one dimer when just one monomer has been modified. It should be noted that these serine residues between 157 and 172 overlap the predicted nuclear localization signal (NLS), which is critical for capsid transport to the NPC [111]. These two roles for this region, phosphorylation via an encapsulated enzyme, and exposure to the cytoplasm, imply that the C-terminus can

adopt multiple conformations with respect to the surface of the capsid. Although such conformational motions are expected to be very substantial in magnitude, in order to achieve the requisite degree of exposure, our findings support the potential for such mobility at the C-terminus [95]. For capsids which are recycled back to the nucleus the NLS exposure and resulting nuclear import must be carefully timed to avoid premature release of the polymerase-DNA complex, to avoid terminating the synthesis of the second (+) DNA strand [112]. By tightly coupling DNA formation to exposure of the NLS, via phosphorylation at the C-terminus, this timing can be achieved, but dynamics of the C-terminal region is a secondary requirement to explain this coupling of function.

The second region of phosphorylation involves the main structural body of the core protein, specifically Ser87 and Ser106. These two residues are both located at critical regions of the core protein structure. Ser87 is present along the side of the helix-bundle tower, protruding away from the main body. Ser106 is near the base of the alpha-4 helix, prior to the turn into the C-terminal alpha-5 helix. The serine at position 87 is not involved in dimer-dimer contacts involved in capsid formation, nor is it part of the intra-dimer contacts. Regardless, phosphorylation at this location was observed to facilitate capsid formation by increasing the affinity between dimers [113]. There is currently no explanation for this effect, but naturally-occurring mutations in the body of the towers (specifically, F97L) have also been observed to increase assembly [72]. Considering the tower distortion and hinges seen in the center of the tower in the recent dimer structure (A. Zlotnick, in preparation), phosphorylation and mutations in this region likely stabilize assembly-competent structural conformations or dynamics. Very similar effects were seen for phosphorylation at Ser106. Although there is an inexplicable loss of alpha-helical character (as measured with CD) as a result of phosphorylation, the resulting dimeric structure is more resistant to urea

before losing secondary structure, and capsids assembled from this phosphorylated material will tolerate more urea before disassembly [114]. These results also have no explanation in the literature at the current time, but the dimer structure discussed below offers possible explanations for such effects.

Capsid Protein Conformations

Like many viruses, HBV makes efficient use of DNA by producing multiple proteins from a single genomic sequence. The "normal" capsid protein is produced as 183 amino acids, and contains an arginine-rich C-terminus which interacts with the packaged nucleic acids (chapter 2). The same sequence can also be produced from a second in-frame start codon upstream, which produces an additional 29 amino acids, and is known as the e-antigen (HBeAg) [115, 116]. The first 19 of these amino acids form a signal peptide which targets the protein to the endoplasmic reticulum, where the signal peptide is cleaved off. The remaining 10 residues are hydrophobic and contain a cysteine in the -7 position, relative to the start of the core protein. That cysteine may form a disulfide bond with Cys61 in the body of the protein, which then prevents formation of the normal homodimer structure seen in the core protein [117, 118]. However, there is some debate as to whether the disulfide is present in the native form of the protein and if the monomer is ever present in vivo [82]. Regardless of the structural fate of HBeAg, it is agreed that the C-terminus is removed by a basic endopeptidase, resulting in a protein that extends from the -10 to 149 position relative to the core protein. This construct is then secreted from the infected cell.

There are several interesting aspects to the HBeAg protein. First, it has never been fully characterized structurally. The entire existence of HBeAg is based upon serological characterization and the formation of antibodies specific to a presumed form of protein: no purified form has been isolated and there is no direct structural

data. Protein sequences designed to mimic HBeAg assemble into capsids in recombinant systems, which calls into question the strictly non-capsid assignment of HBeAg [119]. Studies regarding epitope presentation of HBeAg have also had mixed results. The primary epitope of the e-antigen (HBe1Ag) is linear and assigned to the region which is on the top of the tower in the capsid structure [91]. This same region is the dominant epitope for the core protein antigen (cAg), and is strictly associated with the capsid conformation, but in that form the epitope is structural and spans both monomers at the top of the tower (approximately residues 77-83) [91]. A second epitope on HBeAg has been identified localized to the C-terminus (HBe2Ag), but there is some conformational contribution to antigen recognition provided by the N-terminus [91]. Initial characterization of this epitope was conducted prior to determination of the crystal structure, so the HBe2Ag epitope was assigned to the region of residues 10 and 130 together, with an unknown contribution from the central region of the protein functioning to bring the two termini together. The crystal structure from the capsid shows how the N-terminus is spatially close to the C-terminus, but the C-terminal helix forms an obstruction between the two halves of HBe2Ag (figure 2.1). It is also difficult to reconcile the disulfide between C-7 and C61 with an extended conformation that bridges the C-terminal loop, beyond the alpha-5 helix. The sum of data on antibody binding to the hepatitis B core protein can best be described as mixed: in some conditions "dimer" can be recognized as capsid via anti-HBcAg, and capsids can also be disrupted to generate HBeAg signals [120]. Historically the HBe2Ag was used to reconcile the overlap between the HBcAg and HBe1Ag epitopes, but the apparent production of the e-antigen from disassembled capsids occurs despite the lack of a C-7/C61 bond, which is supposedly characteristic of the HBeAg form of the protein. Antibodies for HBe2Ag have also been observed to react with assembled capsids: this has been interpreted as being the result of broken capsids and a mix of

assembly states [121, 122], but our observations suggest that intact capsids transiently expose unfolded regions of both the towers and the C-terminus (chapter 3 and [95]), offering an alternative explanation for antibody binding.

The substantial cross-reactivity between HBeAg and HBcAg observed in many studies might tend to suggest that the HBeAg form of the protein does not exist as a unique structure at all, that it is an artifact of the experiments and is structurally uninteresting. This point of view can be immediately discounted due to the extremely strong selective pressure for the HBe antigen: with the exception of relatively few known mutants, which are selected against strongly, all HBV strains conserve the production of HBeAg. Although the exact function of HBeAg is currently unknown, it undoubtedly serves a crucial biological role in the viral life cycle. Anti-HBe production has a strong correlation with active and advanced hepatitis, and is typically viewed as an indicator of the carrier being infectious [123, 124, 116], but late-term response to HBeAg, especially the linear epitope, also is associated with virus clearance [121]. The exact role of HBeAg remains elusive, but depending on the situation (carrier vs active hepatitis, adult vs neonate, early vs late infection, etc) the same protein can have either immunogen or tolerogen characteristics [116].

Structure and Biology Conclusions

These observations (structural pleomorphism, capsid targeting and transport, assembly/disassembly, etc) regarding the hepatitis B virus are incompatible with the view of the core protein as just a static framework upon which the virion is assembled. Every step of the virus life cycle implies that structural dynamics are a critical element of regulating viral replication. Regulation of the two different start codons (HBeAg vs HBcAg) may determine the overall balance of non-capsid vs capsid protein production, but what is the ultimate fate of the HBeAg form of the protein?

Anti-HBc is one of the first immune responses generated by the host, but exposed capsids are never released from infected cells. At the same time, HBeAg is produced at substantial levels and is directly targeted for secretion. Assembly studies using model HBeAg have successfully produced capsid forms of the protein, and evidence for a monomeric form of the core protein is limited and questionable [82]. Together, it may be speculated that secreted HBeAg is designed to produce capsids as a decoy for the immune system [103]. But while the biology of infection could be a target for HBeAg, there is no question that it can play a role in regulating the kinetics and thermodynamics of capsid assembly. Capsids produced from mixed populations of mutant and native protein are well known for HBV: Y132A, which decreases capsid assembly affinity and a GFP-fusion, which allows tracking of dimer location [101, 125]. Both of these mutants can form complete but heterogeneous capsids in the presence of native protein. Given that HBeAg is expected to share the same overall structure as HBcAg, it would have the same behavior. In this manner, HBeAg could serve two roles: by forming weaker capsids, perhaps incapable of properly packaging nucleic acids, the pool of available HBcAg dimer would be reduced and capsid assembly could be intentionally stalled. These sequestered capsids could then be recycled into dimer and eventual infectious virions. Secondly, capsids formed with a fraction of HBeAg could be targeted for secretion without first acquiring the proteolipid outer coat of HBsAg. Although this is speculative, it would account for the early and strong production of anti-HBcAg observed in infected individuals [103]. Both of these possible explanations are dependent upon structural dynamics of the core protein. Our work has demonstrated that restricted motion of the C-terminus in HBcAg is associated with an assembly-active state, which would encourage interactions between HBeAg subunits, or between heterogeneous combinations of HBeAg and HBcAg (chapter 2 and [95]). Because the core protein must be proteolyzed to generate HBeAg, the accessibility

of the C-terminal regions to proteolysis via dynamic motion is also highly relevant: the generation of secretion-ready HBeAg can be stalled by preventing proteolysis at residue 149. Taken together, these observations suggest that substantial structural motion, or the cessation thereof, is required to form HBeAg, and the formation of this variant of core protein structure may be a key regulatory element in the HBV life cycle.

Once assembled, structural motion continues to play a role in the function of the HBV capsid. Phosphorylation is a common post-translational modification responsible for a wide variety of effects. In the case of HBV, it is used to coordinate the targeting of the immature capsid to the nucleus with the formation of double-stranded DNA from the initially-bound single-stranded DNA, which may be critical for early amplification of viral DNA in the nucleus. However, phosphorylation of the arginine-rich C-terminus region via an internalized kinase must be reconciled with the need for an external NLS in the same region. Previous studies have found that this region is susceptible to proteolysis, and our findings support such claims, with up to 1% of the C-termini exposed on the exterior of the capsid at any given time [95]. Production and secretion of mature capsids also implies dynamic motion of the buried C-termini. A lysine at position 97 was found to be critical for interaction with gamma-2-adaptin, a putative endosomal trafficking adapter and known binding partner of the viral envelope proteins [126]. This residue is located within the hinge region of the dimer tower, which straightens and becomes more symmetric upon capsid formation (A. Zlotnick, in preparation). The second half of this ubiquitination scheme is provided via the PPAY ubiquitin-binding motif, which is present at the C-terminal end of the alpha-5 helix. By blocking the ubiquitin ligase-core protein interaction with mutants abolishing the binding motif, HBV egress from the cell was reduced [126]. Although this site does have a small amount of solvent exposure on the assembled capsid, full

access to the PPAY sequence would require some protein unfolding. This particular sequence happens to be just one residue after the observed R127/T128 cleavage site in proteolysis experiments, which confirms substantial availability of this region of the structure.

Unified Model of Kinetic Hydrolysis and Hydrogen-Deuterium Exchange

Put into the context of the structures and function of the HBV capsid, the dynamics observed via kinetic hydrolysis and hydrogen-deuterium exchange provide excellent solution-phase confirmation of the structural motion implied by the crystal structures and the biological activities of the virus. The thermodynamic and kinetic trends collected from assembly assays have long suggested that allostery plays a very significant role in the regulation of capsid assembly, but these observations from kinetic hydrolysis and hydrogen-deuterium exchange are some of the first direct evidence for concerted structural changes. HAP compounds are known to act as a steric wedge, changing the angular orientation between subunits in assembled capsids, but we also find that these compounds have a direct allosteric effect upon the large-amplitude dynamics of distant regions of the dimer structure. It is very attractive to hypothesize that the primary role of HAP compounds is not to act as an inert steric wedge (though it clearly has that function), but rather to modulate dimer-dimer interactions by regulating the motion of large regions of the dimer. Such an explanation could help tie in discordant observations, such as the inexplicable effects of ionic strength, and the powerful activity of zinc, which has no known binding location.

Although highly speculative, the substantial quantity of circumstantial evidence points to concerted rigid-body motion within the HBV dimer, with simultaneous

regulation of a highly dynamic C-terminus. These motions can be summarized as involving a radial distortion of the tower apex, with the dimeric state tending to be highly asymmetric and more solvent exposed. This rotation is mediated by a slight outward translocation and bending at the hinge region in the center of the tower. The C-terminal helices are likewise radially distorted, with a hinge at the base of the dimer in the "chassis" region. Because the dimer structure was obtained from a set of three dimers in a distorted version of the normal trimer-of-dimers, the dimer structure may represent an assembly form with partial capsid character: in the absence of these hydrophobic contacts in the critical C-terminal region, the free dimer structure in solution may contain even more distortion than the current "dimer" structure. A cartoon schematic of the proposed dynamic model is shown in figure 5.2. Hinge regions, mediated by the chassis and the fulcrum regions, allow substantial rotational motion of helices. With the exception of the C-terminal helix, these helices remain highly intact in the course of core protein interactions with assembly effectors, contact with other dimers, and other structural events. In support of this theory, there is minimal evidence for any alteration of the secondary structure of the core protein: of non-denaturing treatments, only phosphorylation at S106 has been observed to cause a decrease in the helical content via CD [114].

On a grand scale, the motions observed in the dimer crystal structure are relatively small: the structure of the individual helices are maintained almost perfectly, but it would be a mistake to discount or marginalize the observed rotations, considering the body of biological evidence which support them. The hinge region contains capsid-stabilizing mutations and phosphorylation sites, and the distortion at the tower apex is the location of epitopes which distinguish between dimer and capsid: it is notable that the capsid-specific epitope is linear, which the dimer form is conformational. The chassis region of the structure is the location of another phosphorylation site, as

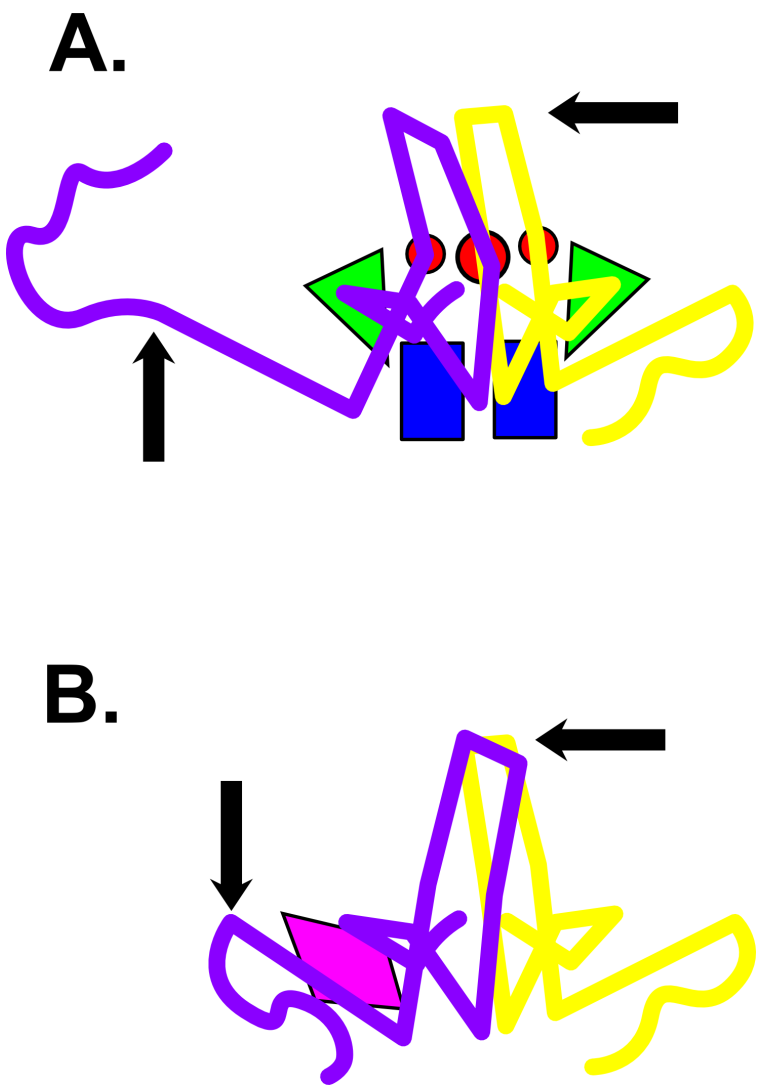


Figure 5.2: Model of HBV core protein dynamics. A cartoon of the core protein is shown, with the right half of both panels indicating the static conformation observed in the capsid crystal structure. A.) Structural distortions of the free dimer modulate kinetic hydrolysis measurements (known cleavages shown with arrows), but do not alter the secondary structure. The fulcrum region is in green, chassis in blue, and the helix-breaking pivots are in red. The highly extended C-terminus allows efficient HX with the solvent. B.) In the presence of HAP, the fulcrum is tightly bound and restricts the free rotation of the C-terminal helix. Cleavages (arrows) are dramatically reduced, but because the C-terminus remains highly solvent-exposed along the bottom and sides, it continues to display a large amount of hydrogen-deuterium exchange.

well as a putative zinc binding site, both of which promote capsid assembly. Moving away from the chassis to the C-terminus, the same region which unfolds to allow proteolysis access also serves as the inter-dimer contact region of the assembled capsid, the binding location for assembly-effecting HAP compounds, and a ubiquitin-ligase binding site. At the extreme arginine-rich C-terminus, one region of protein backbone much simultaneously participate in phosphorylation events on the capsid interior, while also serving as an external nuclear localization signal.

Further refinement of the dynamic-dimer hypothesis is provided by the dramatic differences in kinetic hydrolysis experiments and the subtle changes in HX signals. As a probe of large-scale motions, proteolysis reveals that both dimer and capsid are highly sensitive to temperature and assembly-effecting compounds. In particular, dimers are somehow protected from proteolysis at the apex of the towers, while capsids are cleaved relatively rapidly at the 82/83 location (figure 5.3). While the structural models do not yet provide a clear explanation for this effect, the distortions observed in the dimer structure may be a portent of similar motion possible in the assembled capsids. Since capsids easily tolerate substantial conformational stress imposed by mutants and bound antibodies at the tower apex, it is reasonable to assume that the crystal structure is representative of a static structural variant, in which motion has been quenched by crystal packing forces and temperature. Dimer and capsid also show differing responses to HAP and zinc. While both assembly states have decreased large-scale dynamics in the presence of HAP compounds, dimer is much more severely affected, with reduced cleavage rates, but only in the dynamic C-terminal region following residue 110. Zinc shows similar trends, but in the opposite direction: zinc substantially increases cleavages in the 83-127 region, which corresponds to the chassis and putative zinc binding site.

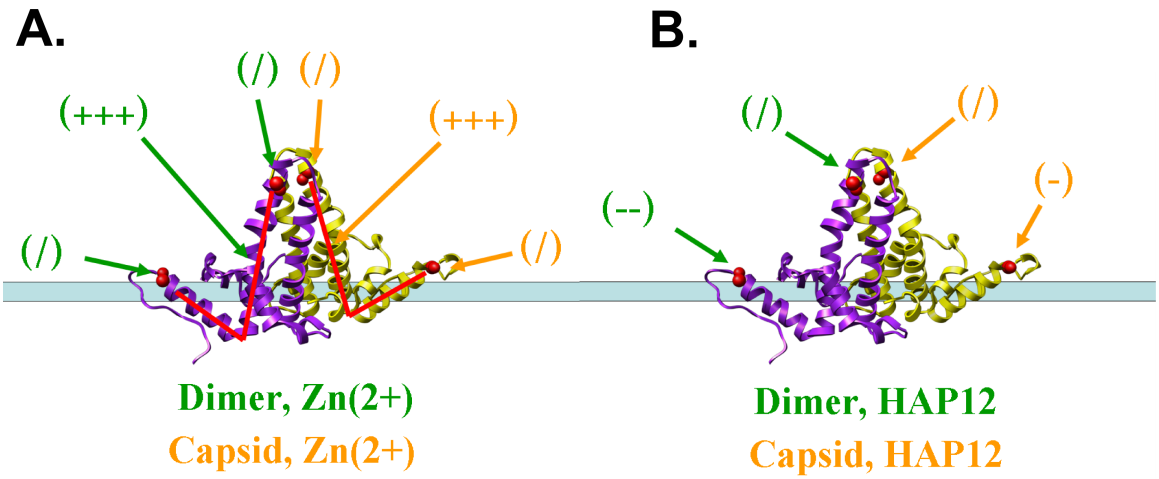


Figure 5.3: Summary of proteolysis measurements of dynamics. A.) In the influence of zinc, there are very small changes in the cleavage rates at 127/128 and 82/83 for both dimer and capsid. However, unidentified cleavages between 83 and 127 (red lines) dramatically accelerate overall proteolysis of the intact core protein. B.) HAP12 decreases proteolysis at the C-terminus, but has a very minor (though measurable) impact on the top of the towers. The cleavage rate is decreased much more in dimer compared to capsid.

By using D_2O as a molecular probe of backbone motion, we observe fairly small changes between dimer and capsid. Likewise, the addition of HAP compounds has minimal effect on the overall dynamic motion, with no real mass shift observed for capsid despite high concentrations of HAP. Dimer is slightly more sensitive, with HX-probed dynamics decreasing about 10%. Although speculative, the similarity of dimer/HAP trends to those observed with capsid or capsid/HAP does suggest that HAP is inducing a conformation very similar to the structure of the intact capsid. But while both HAP and zinc are strong assembly-inducing compounds, zinc is effectively silent. Since zinc binds into part of the "core" of the dimer structure, and involves proton-rich histidine residues, the lack of HX effect with zinc is especially surprising. However, all these observations reconcile well with the rigid rotational model: assuming that the Cp149 dimer subunit is a highly modular structure, with rigid and

stable helices connected by flexible linker regions, secondary structure should remain mostly intact despite large-scale conformational changes. At the C-terminus, the extent of protein unfolding dynamics is highly influenced by the environment of the alpha-5 terminal helix: hydrophobic contacts can stabilize an extended conformation as implied by the HBe2 epitope, reduce cleavage at the ubiquitin-ligase binding site, and mediate interactions between phosphorylation and the presentation of the NLS. We find evidence for the burial and sequestration of labile hydrogens in the process of forming these hydrophobic contacts: HX experiments show a decrease in the exchangeable protons of the protein core. However, changes in secondary structure pale in comparison to the influences acting on rotations of the helices, as measured with proteolysis in the presence of HAP or zinc.

These results represent substantial progress to determining the relationship between structural dynamics and function for the HBV Cp149 protein. By combining techniques which each report on a different scale of dynamics, we have provided some of the first direct experimental evidence for structural motion in the hepatitis B core protein. Kinetic hydrolysis works extremely well for extended reactions and provides very detailed information about backbone exposure throughout the body of a protein. However, it is insensitive to small structural changes and can easily generate data which is beyond our ability to fully model. For more rapid assays, or to evaluate smaller structural motions and hydrogen bonding, HX is far superior: a single effector can be assayed in a hour or less, the data is very manageable, and theoretically every labile hydrogen is probed via deuteration. The combination of these approaches supports a chassis/arm rotational model for helix motion in the HBV core protein, which helps to explain many previously-observed characteristics of the hepatitis system. Future work using this approach will further characterize the details of the structure/function relationship, such as the effects of ionic strength,

the quantitative effects of phosphorylation, and the binding location and mechanism of zinc. Since kinetic hydrolysis and HX assays are both very amenable to high-throughput processing via automated mass spectrometry, incremental improvements in the assays techniques will provide exponential benefits in experimental results. Together with more traditional biophysical tools, these assays provide a very detailed portrait of solution-phase behavior, and they have a promising future as mainstays for the study of supramolecular dynamics.

PHYSICAL SIGNAL MODULATION

Introduction and Background

Many mass spectrometry-based approaches to the study of viral capsid dynamics demand extremely high mass accuracy, although the exact reasons for this vary depending on the nature of the experiment. The capsids themselves can be a very high-complexity sample, especially when dealing with capsid composed of multiple different proteins. Capsids which have post-translational modifications, either natural or engineered, also can produce a huge possible pool of protein and peptide fragments. As one example, the Cp149 protein for the HBV capsid contains one disulfide bond. Assuming that the disulfide could be either present or reduced as a consequence of sample handling, when using a relatively non-specific protease, there are millions of possible protein fragments which could be observed. Even with the use of MS/MS fragmentation data, which necessitates some drawbacks such as reduced scan time, the search space for MS data is too large without the use of high-precision measurements. Similarly, high precision is an absolute requirement when working with protein sequences of unknown composition, when PTMs are not known in advance, or samples contain trace quantities of small molecule contaminants.

The need for high precision and accuracy was highlighted by the inability to positively identify proteolysis fragments in the presence of zinc. Although zinc adducts were part of the search to find matches for observed weights, no mass-based confirmations were made. Besides the extremely large search space when searching with such variable modifications, the other complicating fact was the inability to achieve the extremely high mass accuracy needed for positive mass-based sequence confirmation.

Methods capable of producing high mass accuracy would therefor be invaluable as part of both routine and targeted mass spectrometry analyses.

However, high precision and accuracy are not trivial to achieve even with modern instrumentation. Many publications in the field of mass spectrometry use accuracy and precision terminology and protocols in a conflicting manner: the lack of consensus in the field is a reflection of the difficulty in systematic statistical evaluation, as well as an indictment of community usage and standards for high-precision measurements. To ensure that the results obtained for mass spectrometry analysis of viral dynamics experiments is rigorous and justified, we have conducted a study of the primary factors influencing mass accuracy for time-of-flight (TOF) mass analyzers operating under typical conditions.

Abbreviations and Conventions

MMA = mass measurement accuracy Δx = the spacing between data points on the x-axis: this extends to Δt, Δ m/z, etc. W = the width of a peak, which may have varying definitions: FW20M = the peak width at a height equal to 20% of peak maximum. FW50M = the peak width at a height equal to 50% of peak maximum. FW80M = the peak width at a height equal to 80% of peak maximum. σx = the error of x. Unless otherwise noted, error values are for one standard deviation. K = absolute number of data points per peak in a given spectrum: a function of W and Δx only. R = resolution: x/W, where W is typically FW50M

Effects of Data Sampling Interval on Mass Accuracy and Precision

Mass spectrometry instrumentation has become sufficiently advanced in the last 10 years that reasonable mass accuracy and precision are now considered routine. Although there are no formal definitions, community standards typically refer to a

mass accuracy error of approximately 5 ppm as being "high precision", although it would be more correct to refer to that metric as a measure of accuracy. Precision is rarely addressed directly, although replicate measurements published within one study are usually exceptionally precise: one the order of 1 ppm or less. Especially in the light of this study, we find such claims highly doubtful, and more likely to represent perfect replication of imperfect measurements, as compared to a true stastical evaluation of a variable with a Gaussian or similar distribution of noise. In our experience, an accuracy level of 5 ppm can always be achieved, even with older generation mass spectrometry hardware. However, this result is not always trivial, i.e. instrument adjustment may be necessary to achieve a 5 ppm accuracy level. The question then becomes, is it possible to be certain of 5 ppm MMA for any unknown experimental measurement, regardless of signal-to-noise or other factors? Furthermore, what are the factors affecting the precision of a particular measurement?

To address these issues, we must first clarify some terminology. Mass error can refer to several forms of error: it can be the propagated error of calculations, such as those used to assign peak position, the inter-spectrum variation in peak positions (effective mass precision), or the difference between the true m/z and the observed value (mass accuracy). Because the last form of error, mass accuracy, defines the limits for mass-based search criteria, it is the primary metric of interest. The common procedure is to assume that mass accuracy is free from systematic errors: with a perfectly accurate instrument, effective mass accuracy would be determined by the error components of peak determination and scan-to-scan variation (the first two error components above). For this reason we follow the convention of using error and accuracy interchangeably, since the latter is a function of the former.

The true accuracy error of an experimental measurement is difficult to determine. Because there is currently no universally-accepted protocol, several different

approaches are used to estimate mass error. The most straightforward approach is to empirically make numerous replicate measurements, with the goal of obtaining an unbiased distribution of samples. Such analyses are time-consuming and measurement and calculation bias is often difficult to detect: we will present one major source of bias here. A second commonly-used technique is to use the residuals of the calibration spectra and function as an estimate for any m/z measured within the calibrated range. Using residuals in this manner to estimate the error is based numerous presumptions, including the choice of calibration function, the weights applied to the calibration function residuals, the mass range used for calibration, etc [127, 128]. The most rigorous approach for error estimation is to measure many known compounds, ideally spanning a significant period of time, to generate a large sample of observed error measurements. As with the first approach, these proxy measurements are time-consuming, and may not be an accurate estimate for the true error of any one particular experimental signal.

As an alternative to estimating error via empirical measurements as described above, attempts have been made to estimate the error directly from the experimental signal. This approach is very attractive for two reasons: error can be estimated without additional experiments, and the error estimate will be specific for the particular signal, rather than extrapolated from controls which may differ in intensity, m/z, background noise, etc. However, error estimates which arise directly from experimental signals tend to be measurements of peak calculation error (precision), rather than true mass accuracy. The original derivation of peak calculation error was proposed by Campbell and Halliday, who modeled the peak shape as a triangle [129]. For such a triangle, counting noise in intensity causes errors between the apparent center of

mass (or centroid, x_c) and the true center of mass, for a discrete set of points:

$$x_c = \frac{\sum xy}{\sum y} \tag{6.1}$$

The error in centroid position is given by:

$$\sigma x_{c,ppm} = 10^6 \frac{W}{x_c \sqrt{24 N_p}} \tag{6.2}$$

In this equation x_c is the centroid for the particular discrete set of points representing a peak, W is the width of the peak, N_p is the number of ions within W, and $\sigma x_{c,ppm}$ is the error of the centroid calculation. In this form, $\sigma x_{c,ppm}$ is a precision factor for determining the centroid location, but only for one particular set of points. This equation has been used in recent years to evaluate precision as a function of data collection time [130, 131]. The most common variant of this equation is

$$\sigma x_{c,ppm} = \frac{10^6}{\sqrt{N_s}} \tag{6.3}$$

, where N_s is the number of spectra (and proportional to N_p). Error values as calculated with this equation have been used as an estimate of error for exact mass measurements, which is inappropriate on modern instrumentation with high signal-to-noise: the Campbell-Halliday formula provides reasonable estimates of accuracy error only when instrument noise is the dominant source of measurement error (i.e. precision is a far greater concern than accuracy). As instrument noise and error decreases, centroid calculation of x_c becomes excessively precise: on a typical modern TOF sub-ppm levels for σx_c are achieved with less than 30 seconds of averaging (precision for a particular centroid is extremely high), while true peak accuracy error remains much larger and dominated by other hidden factors, such as instrument and signal artifacts.

Figure 6.1 illustrates the fundamental problem with the use of centroids. The peak shape is ideal, noise is completely nonexistent, and there are five data points

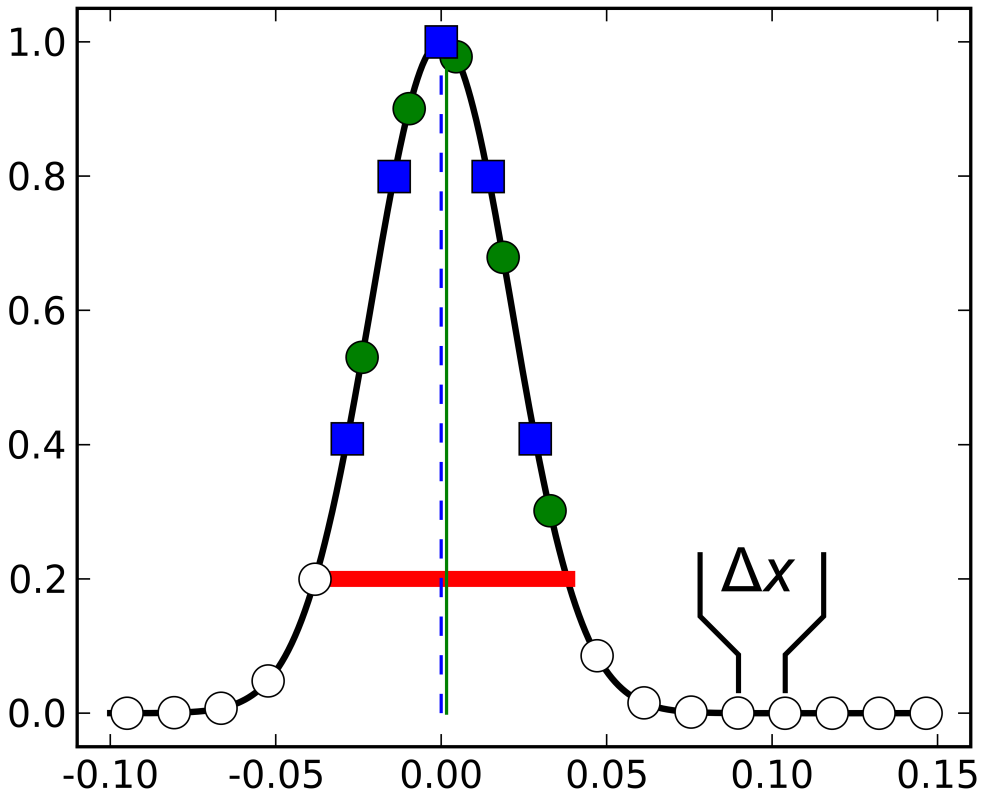


Figure 6.1: Data sampling of peaks gives discrete data sets with random centroiding errors. A representative peak is shown, with 10,000 resolution at 500 m/z. The x-axis is zeroed at the peak center. Two possible sampling schemes are shown: both have identical spacing (Δx) between points. Data points above the standard FW20M cutoff (red line) are shown filled, and are included in centroid calculations. Data points below that threshold (shown empty, squares omitted for clarity) have no impact on the centroid calculation. The blue squares data samples are perfectly symmetric, and accurately capture the true center of the peak. The green circles have a slight offset which results in substantial accuracy errors: the centroid calculation (vertical green line) is +3.3 ppm relative to the true center (dashed vertical blue line). The smaller Δx is, the smaller the maximum possible offset and resulting errors can be.

within the centroid cutoff (FW20M in this example): the Campbell-Halliday formula correctly predicts zero precision error in the centroid calculation. However, the centroid for the second data set (green circles) has +3.3 ppm accuracy error. By using a discrete calculation, the set of points used for a centroid (a sample) will not be able to recover the center of mass from the whole continuous peak (the population) unless the sample is very large. With centroiding algorithms, peak-assignment is performed via direct calculation with no explicit partitioning of error (we refer to any such method as peak-calculation). Alternatively, peak-assignment may be performed by data processing which deals with error explicitly, such as fitting the discrete data points with a known peak-shape function (often Gaussian): we term this peak-fitting. The error associated with peak-fitting has been analyzed in detail by Lee and Marshall [132], who derived the following:

$$\sigma x_{f,ppm} = 10^6 \frac{\sqrt{W\Delta x}}{S/N C(x) x_f} \tag{6.4}$$

Δx is the data sampling interval along the horizontal axis, S/N is the signal-to-noise ratio for the peak as calculated, $c(x)$ is a constant specific to the peak shape (0.708 for Gaussian, 0.627 for Lorentzian, etc), x_f is the assigned location of the peak on the x-axis, , and $x_{f,ppm}$ is the error in estimating the peak location via peak-fitting. Because the peak-fitting approach treats error explicitly, it is relatively insensitive to the subset of possible data used for fitting, and with correct models for peak and error it can reduce or eliminate systematic error. The second equation relates a peak width W and the data point spacing to the number of samples per peak, K. If the common relationship applies (data dispersion is finite), the Lee-Marshall formula takes the same general characteristics as the Campbell-Halliday formula: both predict zero precision error as the noise decreases to zero. However, the meaning of the error value changes with the two equations. With peak-fitting, the error refers to the

estimate of the true peak center, ideally free of any bias. In peak-calculation, the equation describes the error of a discrete mathematical construct: the center of mass for a particular set of data points. Because peak-calculation is an explicit calculation, x_c is far more sensitive to noise and artifacts than x_f. Deviations between centroid calculation and true peak center have been observed in practice as a result of too few data samples and asymmetric sets of points across a peak [130]. Therefore, we propose that the Campbell-Halliday formula must be used with the original context in mind: the centroid precision error σx_c may approach zero, but this factor is for the error of a particular calculation, and it cannot be used as an estimate of the absolute accuracy error between true and calculated m/z. Accuracy error estimates cannot be achieved via peak-calculation (with typical sampling schemes) if sampling only one representative distribution of points across a peak.

To our knowledge there are no equations for prediction of true mass accuracy resulting from peak-calculation, despite the fact that peak-calculation remains the default method for peak assignment in most software packages. Some publications cite 4-5 data points per peak as sufficient to accurately determine the peak location, but none contain derivations of this rule or experimental data to verify such a claim [1] We have addressed this gap in understanding by simulating millions of peaks with known parameters, then performing peak-calculation with the standard methods. The results of these simulations have confirmed that centroid asymmetry is the largest factor responsible for systematic accuracy errors in peak-assignment when using the centroid peak-calculation method: this error arises from the discrete nature of the calculation procedure itself, and exists even in ideal data with no noise (i.e. highly precise centroid

[1] A 1977 work claims 4-5 points per peak as sufficient, but it is unavailable for review, and as noted here, the improvements in S/N for modern instrumentation likely invalidate the justification for few data points per peak [133].

with low σx_c). The most straightforward way to reduce this systematic error is to sample more data points per peak, by way of reducing the data point spacing, Δx.

Data point spacing is also a critical component in the mass accuracy and precision of peak-fitting analyses. The Lee-Marshall formula indicates that arbitrarily high precision can be obtained with sufficient signal-to-noise, but this is under the assumption that peak intensity is substantially above baseline and that the peak is well-isolated and not overlapping other peaks. Such conditions are not trivial to achieve, and many papers address the problems of calculating correct mass for incompletely-resolved signals [134, 135, 136]. Furthermore, with few samples per peak, it is impossible to determine the presence of unresolved signals without prior knowledge. For these reasons we conclude that the instrument parameter Δx is just as important as the highly-regarded resolution factor in determining the performance of a mass analyzer: a "high quality" spectrum should have high resolution, high signal-to-noise, and high sampling frequency. For example, by using Equation 4, we can compute the peak-fitting precision of a TOFMS given the resolution, flight length, kinetic energy, and digitizer frequency: if the resolution of a 500 m/z peak is 10,000 (as measured at the FW50M), with a 2 meter flight path and with ions accelerated to 25keV, the flight time is approximately $2 \times 10^{-}5$s, and peak width is $2 \times 10^{-}9$s. Using a 2 GHz digitizer, the peak will be sampled with approximately 4 points. Entering these parameters into the Lee-Marshall equation, we can calculate that the S/N needs to be over 100 in order to achieve 5 ppm mass precision under ideal conditions. If the resolution were to be doubled without changing the digitizer, there would not be enough data points to perform a curve-fitting procedure. On the other hand, if the Δx sampling was changed such that 8 data points were collected across a peak, precision would be improved by 40% without any change in resolution. Reducing Δx can be an important strategy for achieving high mass accuracy.

Reduction of the Data Interval Δx

Dependent on the type of mass analyzer, Δx can be adjustable, such as with quadrapoles, or fixed such as with TOFs. The fixed flight path of a TOF means that data point spacing is defined by the speed with which ions can be recorded at the detector. This limitation produces some undesirable characteristics for estimating error and achieving high-quality signals. The fundamental operation of time-of-flight mass analyzers is based upon the velocity imparted to a charged molecule via an accelerating potential:

$$c = t^2 \tag{6.5}$$

In this equation c is a constant which accounts for flight path distance, acceleration potential, etc. All instruments will have some deviation from this ideal behavior, but it serves as the best first-order approximation. Because m/z is proportional to the square of the flight time, a constant time-frequency sampling rate gives rise to a m/z scale with varying $\Delta m/z$ increments: $\Delta m/z \propto m/z$, and the peak width on the time scale is also proportional to the time-of-flight: $W_t = \frac{t}{2R}$. This causes more data points per peak for higher m/z ions as the peak width increases. For a time-of flight analyzer, these physical characteristics give varying signal quality across a single spectrum: for TOF data in the m/z domain, both the Campbell-Halliday and the Lee-Marshall equations for precision/accuracy (Equations 6.2 and 6.4, respectively) must be used on a peak-by-peak basis. One study has addressed the issue of irregular data spacing by intentionally undersampling higher m/z values: they collected fewer points per peak, but peaks had improved signal-to-noise. That approach provided consistent data sets with efficient reduction in data size requirements, but introduced varying S/N, required direct control over the digitizer, and can not be extended for over-sampling instead of undersampling [137].

Physical Signal Modulation Increases Data Sampling and Reduces Noise

Our solution to the problem of reducing $\Delta m/z$ on TOF mass analyzers is to intentionally distort instrument operating parameters to induce peak shifts across the fixed sampling increments on the m/z axis: this is known as physical signal modulation (PSM). Several instrument parameters have been found to produce this effect, but we have experimentally determined TOF reflector voltage to be the most ideal for our particular instrument. This voltage is varied in small increments, which gives the effect of changing the flight path length and inducing peak shifts across the m/z axis. Following this intentional distortion, the spectra are re-aligned by a parameter-free pattern matching technique. This approach allows a given peak to be sampled by predetermined and arbitrary data point spacing, regardless of digitizer hardware limitations. The resulting highly-sampled peaks allow more precise determination of peak location by both peak calculation and peak-fitting methods, reduce systematic error from centroid asymmetry, and allow determination of peak position for unresolved peaks.

Physical signal modulation also revealed that mass spectra can contain substantial coherent noise, which is very distinct from normal random noise, and much more problematic This coherent noise is in phase with the digitizer cycle, and will not average out regardless of the number of scans acquired. However, because the coherent noise is unaffected by changes in the reflector voltage, physical signal modulation dephases the noise and makes it possible to remove the noise from the real ion signals. The resulting spectra have substantially improved signal-to-noise, and when combined with the improvements in peak sampling, permit accurate detection of peaks in regions that have S/N ratios at or below 1.0 in a non-PSM spectrum.

Experimental

Materials

HPLC-grade methanol and water were purchased from Sigma-Aldrich. Trifluoroacetic acid was purchased from Sigma-Aldrich, 99% purity. Sodium trifluoroacetate calibration solutions were prepared according to Moini and Jiang, with a final composition of 50% MeOH, 50% H2O. A second calibration mixture was prepared from Agilent electrospray calibration mixture (G2421A) by diluting the stock solution 50-fold into MeOH.

Instrumentation

All measurements were performed on a Bruker micrOTOF mass spectrometer equipped with an ESI source and an Acquiris AP200 ADC card. The control software version was micrOTOFcontrol 1.1 and some analysis was performed with the Bruker DataAnalysis software package, versions 3.2 and 4.0. Mass spectra were collected in the 50-1700 m/z range in the positive mode. The capillary voltage was set to 4200 V and the end plate offset was -400 V. The needle nebulizer gas was adjusted to maintain 0.2 Bar, drying gas was held at 3.0 L/min, and the heater was 200C. The reflector voltage was kept at 1080 V except when intentionally altered for the PSM method. For data collection 50,000 hardware spectra were summed by the digitizer hardware for every recorded spectrum with no rolling averaging or other processing. Instrument calibration was not performed, as all spectra were reference spectra containing known internal peaks. Spectra collected to evaluate centroid errors were collected with nearly identical conditions except with a summation of 5,000 (the Bruker-recommended default) rather than 50,000. These spectra were collected for an

extended length of time (approximately 20 minutes and 10,000,000 hardware spectra) to examine spectrum-to-spectrum variations.

Spectra Simulations

Evaluation of peak-calculation error was performed with a series of programs written in Python, using the SciPy and NumPy libraries. Versions were: Python 2.5.2, SciPy 0.7.1, and NumPy 1.1.0. Peaks were simulated based upon a Gaussian shape with varying height, width, and a per-point noise in the vertical axis. When used, the noise was modeled with a relative Gaussian component, which is a reasonable approximation for counting noise when the counts are greater than 15 [132, 131]. For each peak shape variant, a matrix of variables was evaluated which included 1000 values for Δx between 10 and 0.005 W units on a logarithmic scale, and 1000 values for an x-axis sampling offset x_{offset}, between 0 and Δx on a linear scale. The combination of these two values defined the data points sampled along the x-axis: the continuous peak function was sampled at the appropriate locations for later processing. After data point selection, relative Gaussian noise was added for some analyses. No knowledge of the original peak function was retained for peak-calculations: all centroid operations were performed blind using only the quantized data.

Spectra Calculation

Peak calculation was performed with the standard FW20M centroid definition, with some simplifications due to the known presence of just a single peak (no S/N cutoffs or overlapping peak corrections required). The data point with maximum intensity was found, and the intensity of that point was used to define a cutoff for data inclusion in the peak calculation: any point with intensity greater than 20% the

maximum intensity was retained, with remaining points discarded. This cutoff threshold was also varied to test the effects of different peak-calculation definitions (FW50M and FW80M). Centroid calculations were then performed according to Equation 6.1. The full matrix of possible variables to evaluate (W, Δx, x_{offset}, peak height, and noise) included over 20,000,000 simulated peaks.

For each instance of peak height, W, and noise, resulting centroid peak calculations were clustered according to Δx. Because x_{offset} was observed to have a linear relationship with the error of the peak calculation, and x_{offset} has a maximum value of Δx, the extreme spread and standard deviations of error were recorded for every value of Δx, as representative of the error associated with peak-calculation. For additional details see Section 6.

The same peak-calculation routine used for simulated data was used to evaluate experimental data. Experimental spectra were also processed with a peak-fitting routine, with an explicit triple-Gaussian equation [138]. To derive the explicit peak description function, all spectra were combined into one master spectrum without averaging: each peak consisted of about 20000 data points. The peaks of interest (m/z 622 and 922) were extracted from the master spectrum, and each peak was fit to a function defined by the sum of three Gaussian curves. The location, amplitude, and width of each Gaussian component were allowed to vary to achieve the best fit with the experimental data. The two triple-Gaussian functions computed with this method were slightly different when comparing the 622 and 922 peaks: this is to be expected (due to different peak widths), and as we are not trying to find a universal peak shape, it is not a concern when evaluating scan-to-scan variation for a particular peak. To fit the 622 and 922 peaks for each spectrum in the set of $\sim$1900 scans, the appropriate triple-Gaussian function was used, but only the amplitude and an m/z-location offset were allowed to vary: the relative parameters for each of the three

Gaussian subcomponents was fixed. Although the determination of the master triple Gaussian functions was performed with data points that extended to baseline on both sides of the peak, for peak-fitting of each individual spectra only the points within the appropriate FW20M, FW50M, or FW80M thresholds were used. This was done to ensure that the input data was identical between peak-calculation and peak-fitting, and the only differences were due to the peak determination routines.

Physical Signal Modulation

The optimized reflector voltage for normal operation of the Bruker micrOTOF was 1080 V. For the physical signal modulation procedure, this value was altered in 4 V steps between 1060 and 1100 V. The signal was allowed to stabilize, and sufficient scans were collected to ensure that the representative averaged spectrum would have very high signal-to-noise ratios, typically achieved with more than 500,000 hardware scans. All spectra variants were collected within one data collection run (over less than 10 minutes), and separated with later analysis in the DataAnalysis software package.

Data Processing

Data analysis was conducted primarily with custom-written programs in Python. Bruker DataAnalysis was used to evaluate standard peak calculation routines, and it was used to generate averaged spectra and export data into a text-based format for subsequent processing with our Python package. CompassXport, also from Bruker, was used to generate mzXML files. The majority of the data processing was carried out in a series of Python routines to handle spectral alignment and processing. These routines have been implemented as a Python library called PySpecTools and are available at http://chemistry.montana.edu and http://pyspectools.sourceforge.net. An

overview of select PySpecTools processing routines is outlined in the Supplementary materials.

Results and Discussion

Centroid Calculations of Simulated Spectra

It has long been known that asymmetry and few data samples per peak can lead to errors in peak-calculation, however a systematic study of these effects has not been completed. It has been reported that approximately 5 data points per peak may be sufficient for an accurate mass determination [137]. To test this, we simulated millions of peaks with the full range of possible discrete sampling schemes: the resulting peaks are representative of what could be observed for any given combination of m/z, instrument resolution, and digitizer sampling rate. These were then centroided using three different cutoffs based on peak height. The resulting centroid calculations give a peak assignment (xc) which has substantial accuracy error compared to the true peak center. The distribution of the accuracy errors is approximately Gaussian, and the scale of the accuracy error is proportional to the peak width. Unlike a true Gaussian, there is a finite limit to the accuracy error imposed by the magnitude of Δx: a 100% confidence interval (extreme spread) of error is +/- 2.02 times the standard deviation. Plots showing centroid accuracy errors as a function of Δx are shown in Figure 6.2. When using the top 80% of a peak for calculation of the centroid position (FW20M, Figure 6.2 red), and 5 data points are available, the apparent center of the peak (x_c) has an accuracy error of approximately +/- 0.016 times the peak width W. A typical 500 m/z peak with a 10,000 resolution will have a peak width of 0.05 m/z at the FW50M level, and 0.076 m/z at FW20M. Because the mass error is proportional to the width used for the centroid calculation, the width value W used

for these calculations is 0.076 m/z. Using that value of W, the accuracy error of centroiding is is equivalent to +/-2.4 ppm (one standard deviation) for a typical 500 m/z peak. These errors are due entirely to the centroid calculation procedure, and are present even in an ideal signal with zero noise (see Figure 6.1). The underlying cause of this error is asymmetry in the observed distribution of data points relative to the true peak location: a perfectly symmetric set of data points recovers the exact peak location.

The different peak-width cutoff definitions have a measurable impact on the precision of calculations, for data both with and without noise. All three width definitions produced the same general trend, but with FW20M giving much more accurate calculations than FW50M or FW80M, which had increasing error, respectively. The differences in precision were exaggerated when simulated noise was added to the data. When using the typical FW20M centroid calculation (this is the cutoff most often reported in the literature), the addition of 10% relative normal noise to the vertical axis increased the error in centroiding by approximately 26%. When the FW50M definition of the peak was used, the addition of 10% noise increased the error by 27%. This effect was further magnified in calculations which used the FW80M definition of the peak. After the addition of 10% noise, the error increased by 240% when the FW80M definition was used.

It should be noted that the error of the centroid calculation (sxc) is relative to peak width, and the width changes with the choice of peak definition; for a Gaussian shape, a FW20M width is 2.7 times wider than the FW80M for the same peak. This makes it more difficult to achieve a large number of data points within a peak when using higher centroid cutoffs, but the precision of the centroid calculation σx_c is improved due to the smaller value of W (Figure 6.2). Our results show that the typical peak-calculation cutoff of the top 80% of the peak (FW20M) gives the least

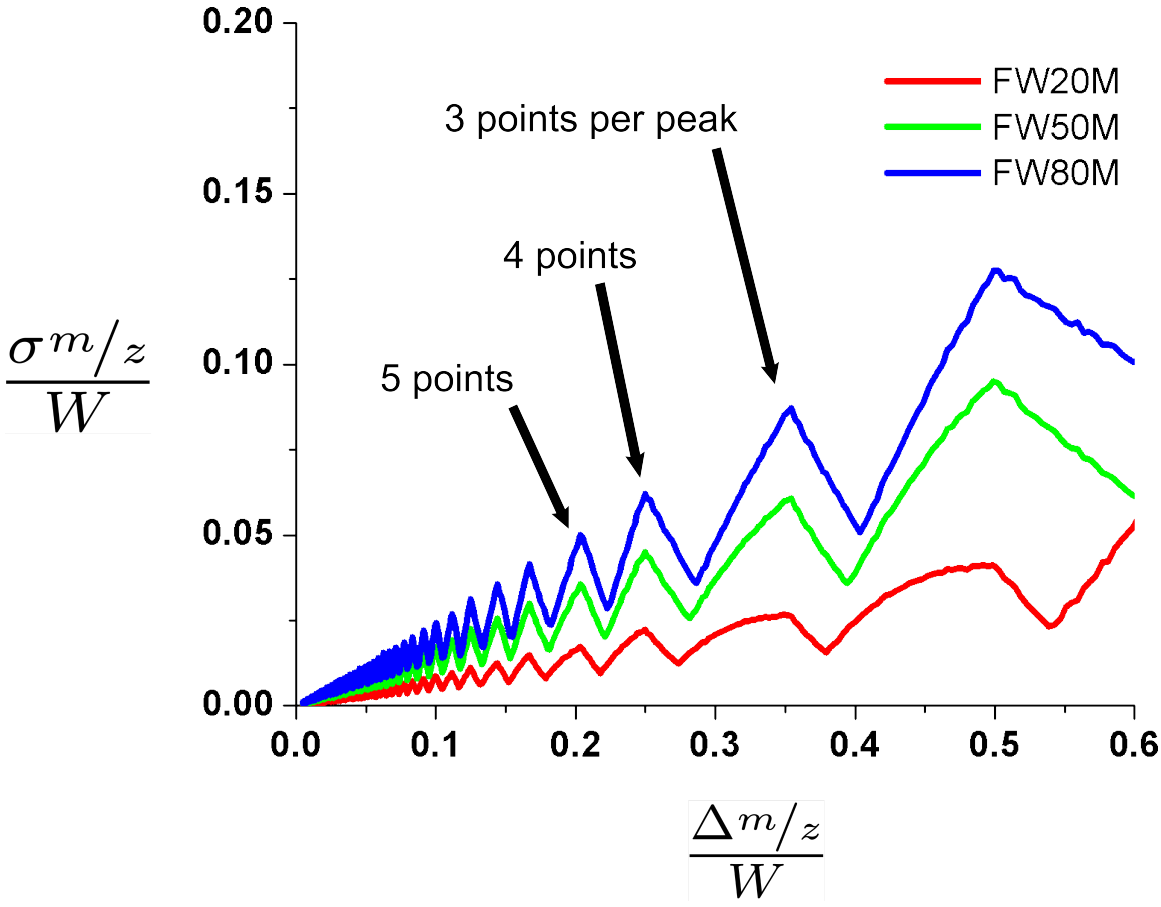


Figure 6.2: Accuracy error in centroid calculation as a function of $\Delta m/z$. The accuracy error of randomized centroid calculations (one standard deviation, in units of peak width), is shown for three different peak definitions: FW20M (red), FW50M (green), and FW80M (blue). Data is compiled from simulated peaks with no noise, and represents best-case scenarios. Lower peak height cutoffs produce more accurate calculations, and are less susceptible to noise (Supplementary Materials). For the example peak used in the text (500 m/z, R=10,000, 5 points), $W = 1.2x10^{-3}m/z$, and the error is 2.28 ppm, one standard deviation. The asymmetry and sawtooth pattern in the trends result from physical constraints of the centroid calculation. These features tend to broaden and disappear with the inclusion of noise.

error of the variants simulated for a given number of data points (K). However, our simulations have not investigated the effects of external noise such as faint overlapping signals, intrinsic baseline noise, etc [134, 136]. If peak-calculation must be used, these additional variables should be investigated for a thorough analysis of error associated with peak-calculation.

Several conclusions can be derived from our simulations. First, because peak-calculation continues to have systematic accuracy error even in the absence of noise, high-quality signals composed of many averages may be insufficient for accurate mass determinations if the number of data points per peak is too small. Highly stable mass spectrometry instrumentation actually exacerbates this problem: if a peak cannot be observed in multiple shifted locations relative to the data sampling grid, replicate measurements will faithfully reproduce the same systematic error. To obtain true replicates, it is necessary to collect spectra separated by a time interval sufficient to allow instrument drift, and replicate measurements conducted immediately and sequentially will likely not show adequate drift. Though rarely carried out in routine practice, analyses carried out on multiple days, such as that performed in the Bristow-Webb study, may be the only way to eliminate peak-calculation biases (Bristow and Webb 2003). As mass spectrometry instrumentation becomes more reliable and stable, even analyses carried out days apart may share systematic errors.

Centroid Calculations of Experimental Data

Because our findings regarding systematic accuracy errors of centroids are derived from simulations and theory, we also examined a large dataset of real spectra to look for examples of error induced by peak-calculation. Our model predicts an error for every peak as a result using centroid calculations, and to check this prediction we evaluated two high-intensity, well-resolved peaks in a large continuous dataset. The

two selected peaks are the highest intensity signals in the Agilent calibration solution (G2421A), and are composed of $C_{12}H_{19}O_6N_3P_3F_{12}^+$ and $C_{18}H_{19}O_6N_3P_3F_{24}^+$. They produce m/z signals at nominal m/z of 622 and 922, respectively. Despite good ion intensity, the 622 peak contains on average only 4.9 data points per peak, and the 922 contains 6.0 (Figure 6.3). The non-integers values reflect the varying number of data points used for each centroid calculation. The signal-to-noise ratio for a single scan is approximately 12,300 for the 622 peak and 75,000 for the 922 peak. Based on our simulations, we would therefore expect an accuracy error of +/- 3.7 ppm and +/- 4.6 ppm (95% C.I., $W_{622} = 0.078$ m/z, $W_{922} = 0.12$ m/z), for the two peaks and a centroid error transition of 7.6 ppm and 9.5 ppm for the 622 and 922 m/z peaks, respectively.

For both inter- and intra-scan comparisons, the errors associated with centroid calculation are not randomly distributed. With a stable mass spectrometer, the true positions of the 622 peak and the 922 peak should be highly correlated within a single scan: instrumental variation such as physical changes in flight tube length and orthogonal acceleration energy will have identical effects upon the experimental and control ions within one transmitted ion packet. Assuming a fixed m/z difference between these two ions, and the fixed m/z sampling grid, there is a correlation between the data sampling patterns for both peaks: as the 622 centroid data set becomes asymmetric and biased to low m/z values, the same should happen for the 922 peak. This pattern should be maintained until one of the two peaks crosses the centroid error threshold: the degree of asymmetry (or x_{offset}) cannot exceed Δx. Due to the periodic nature of the sampling, when the true peak center moves by more than Δx, the error swaps signs from negative to positive (or vice versa) and the calculated peak position shifts by the extreme-spread error value. Based on this characterization, when comparing two peaks calculated via the centroid method, we

expect to see multiple different populations of $(m/z_{\text{peak1}})/(m/z_{\text{peak2}})$ ratios, each with a high degree of correlation. The number of populations is related to the probability of the two peaks drifting across a centroid error threshold, and is more or less random depending on peak location relative to the m/z sampling grid. A second mechanism for generating quantized centroid shifts arises out of the necessity to define a discrete set of points for peak calculation: when a single data point is near the intensity cutoff limit, scan-to-scan noise may vary its inclusion in the calculation, which introduces a quantized shift in apparent mass.

To look for quantized mass shifts arising from centroid calculations, we processed over 1900 individual spectra from a continuously-infused calibration solution, collected over approximately 15 minutes. The raw data file was converted to mzXML format via Bruker CompassXport, and the resulting file was analyzed with a Python script which calculated the centroids of the 4 major peaks in the spectrum using both the standard centroid method and a triple-Gaussian peak-fitting routine, a variant of the method of Strittmatter and Rodriguez. The resulting peak locations for the 622 and 922 ions were plotted against each other, with each spectrum representing an individual data point. For the peak-calculation data, several discrete populations can clearly be seen, with identical slopes but with offsets between the different populations (Figure 6.3d). The distribution of centroid values for the 622 peak also shows two discrete populations even without comparison to the 922 peak. By coincidence the 622 peak happens to fall with two data points very close to the FW20M cutoff limit, and the binary noise from their inclusion or exclusion makes centroid calculation of the 622 peak nearly a worst-case scenario. In contrast, the same data shows no quantized errors when using peak-fitting routines (Figure 6.3c). To make the results between peak-calculation and peak-fitting directly comparable, the peak-fitting routine was restricted to using only data points within the FW20M threshold: in practice peak-

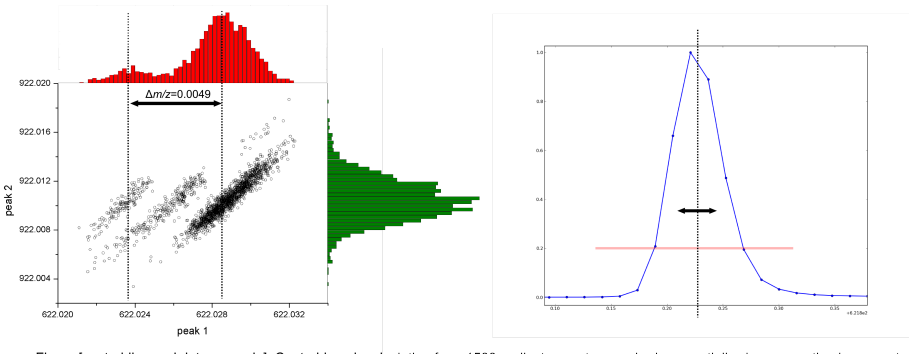


Figure 6.3: Centroid peak-calculations show artificial peak shifts. A) The 622 m/z peak is poorly-sampled and has two data points which fall close to the FW20M cutoff level (red line): scan-to-scan noise causes these to be randomly included or excluded from the centroid calculation. The accuracy error for this peak calculation is predicted to be +/- 3.7 ppm (95% CI), and the extreme-spread quantization transition is 7.6 ppm. B) The 922 peak has more data samples and data points are farther from the 20% cutoff limit. Accuracy predictions for this peak are +/- 4.7 ppm (95% CI) and 9.5 ppm transition. C) Peak-fitting with a multiple-Gaussian model produces good correlation between the 622 and 922 peaks, and the scan-to-scan variation remains low. For each of 1500 spectra, the 922 peak position (as fit with multiple-Gaussian model, see text) is shown on the y-axis, and the 622 peak on the x-axis. D) Centroid calculations produce multiple populations, especially for the 622 peak. At least three distinct populations of 922/622 ratio are visible. The axes for D have histograms displayed opposite the labels to indicate data point density. The 622 peak-calculations are bimodal, with a separation of approximately 7.9 ppm. The separation agrees with the predicted shift of 7.6 ppm, and is close to a worst-case scenario for predicted behavior. Multiple populations are not clearly present for the 922 peak, due to improved peak sampling and data point locations.

fitting can easily use a lower threshold, which would improve the precision of the fits. By using centroid-type cutoff thresholds, the input data for the peak-fitting routine contains the same binary noise of peak inclusion or exclusion as the peak-calculation method. While the centroid routine visibly suffers from this effect, peak-fitting correctly recovers the true peak locations with only a single population of peak ratios.

Centroid Errors and Lock-Mass Corrections

The presence of quantized shifts in calculated centroids has substantial implications to the use of intra-spectra mass corrections, commonly called lock mass. Error propagation should be expected to introduce error into the process of correcting mass shifts via lock mass, which could decrease overall accuracy and precision. Current application of the technique is based on the assumption that the scan-to-scan drifts, if uncorrected, are larger than the error introduced by the correction, resulting in a net improvement. Our results suggest that this may not be the case: individual centroid calculations in raw spectra have large accuracy errors which may easily be larger than the scan-to-scan variation, as seen in Figure 6.3d. Because lock mass has been shown to improve mass accuracy in practice, we offer an alternative explanation for its efficacy despite the usage of peak-calculation via centroids.

For simplicity, assume the parameter of interest is the mass correction calculated via difference between calculated peak locations for a control peak and an experimental peak, $\Delta M = M_{\text{control}} - M_{\text{experimental}}$. The process of calculating centroids for both M_{control} and $M_{\text{experimental}}$ introduces error, and with sufficient mass spectrometer stability and a limited timeframe of data collection, the calculated positions will be different from their true values. In such a situation, the measurement of ΔM will be incorrect, but it will be more accurate (though less precise) than the measurements of either M_{control} or $M_{\text{experimental}}$. Lock mass thus improves accuracy, but this is a statis-

tical generalization: the accuracy of ΔM will usually be improved, but it will also be worsened a small percentage of the time. To be sure that ΔM is very accurate, a large number of different $M_{\text{control}} - M_{\text{experimental}}$ calculations must be performed. This is the second way lock mass improves accuracy in practice: a value of ΔM is computed for every spectrum, which improves the sample set from which the average value of ΔM is calculated. The alternative approach (not part of normal lock mass procedure) would be to average spectra before performing a calculation of ΔM. By averaging first, the sample set of M_{control} and M_{unknown} is condensed to a single instance of each, and both peaks retain systematic error due to the peak-calculation errors inherent in computing the location of a single peak. The resulting value of ΔM is statistically certain to be off by some quantity, regardless of how many spectra were initially averaged.

The application of correction factors to improve mass accuracy and precision is not a trivial procedure, and we do not imply that this cursory analysis of lock mass is comprehensive. For example, Wu and McAllister have noted differences in mass errors when using a combine-then-center approach vs center-then-average when dealing with multiple spectra without lock mass, presumably due to intensity biases for spectral averaging [139]. There is also a substantial body of research regarding corrections for time-digital-converters and the dead time effect, calibration drifts with time, and selective signal averaging [140, 138, 141, 142, 143, 144]. The applicability of each such correction schemes should be evaluated on a case-by-case basis, but in most data processing schemes, centroid calculations play a pivotal role in the determination of peak location. Due to the ubiquity of the centroid calculation in mass spectral analysis, our results have negative implications for a large variety of data processing schemes for the determination of high mass accuracy. Based on our observations, peak-calculation introduces both random and systematic artifacts with accuracy er-

rors that can be much larger than the scan-to-scan variation commonly used as a metric for mass measurement precision. With modern computing power and efficient peak-fitting routines as a viable alternative, centroid calculations should not be used for determining peak location when high mass accuracy is required.

Physical Signal Modulation

Historically, improved accuracy on TOF mass analyzers has been addressed by improving resolution and increasing digitizer speed. These two factors work together and mass accuracy improvements of current high-performance TOFs must be preceded by the development of faster digitizers, so that the number of data points across a peak ($\Delta x/W$) can be increased, or at least maintained. An orthogonal approach to the problem of narrower peaks (decreasing W) is to shift arrival times of ions slightly and then combine spectra, effectively decreasing Δx. After completing testing on a number of instrument parameters, it was determined that the TOF reflector voltage could be used for this. Adjustment of the reflector voltage has a direct effect on the location of peaks, with lower voltages causing ions to have delayed arrival times (Figure 6.4a). Over the range of voltages sampled (approximately +/- 2%), the peak shifts in the time-domain were well-described (R2= 0.9999) by a linear relationship between the parameter c in Equation 6.5 and the reflector voltage. This relationship is the basis for a demodulation function. The linearity and stability of this response suggests that the PSM technique can be applied by collecting only a few reference spectra to calculate the function, then using the demodulation function to correct each modulated spectra without the need for matching internal reference points. Because this work is a proof of principle, every spectrum was experimentally aligned and manually verified. In practice, the simplex calibration method of Christian and

Reilly, which directly models physical instrument operating characteristics, would allow efficient and precise spectral alignment [128].

The plot of modulated spectra also shows a substantial component of coherent noise, which is in-phase with the t (and thus m/z) sampling intervals (Figure 6.4). Because the noise is coherent, it will not average out when combining a number of scans. The exact source of this noise is unknown at the present, but it is reasonable to assign its origin to the electronics of the detector and ADC processing pathway. We also measured this electronic noise by collecting data with the ESI source inactive. The general patterns of noise were the same as in the PSM experiments, but collecting noise trends with control runs in such a manner adds to the number of data collection runs that must be performed and requires additional assumptions, e.g. that the electronic noise will be the same when the detector is recording ion arrival as compared to just "dark" noise. Alternatively, it is trivial to calculate the noise from experimental spectra using the PSM approach. In the example spectra used in this study, each m/z location has been sampled 11 times, and even with the small reflector modulation increments used, several spectra can be found with no real ion peak overlapping any given point of interest. Because detector noise scales in magnitude with detector intensity, the low-intensity background signals have the smallest standard deviations as a set. To automatically choose the correct spectra to define noise at every point, the intensity values for a particular m/z value, from all 11 PSM spectra, were treated as a set. Using that set, all possible combinations of 3 values were calculated, and the combination subset with the smallest standard deviation was used to compute the coherent noise spectra for that m/z value. A plot of the coherent noise calculated with this method is shown in Figure 6.4b. The low-variance noise calculation method was also tested with combinations of size 5, but in regions of high peak density it

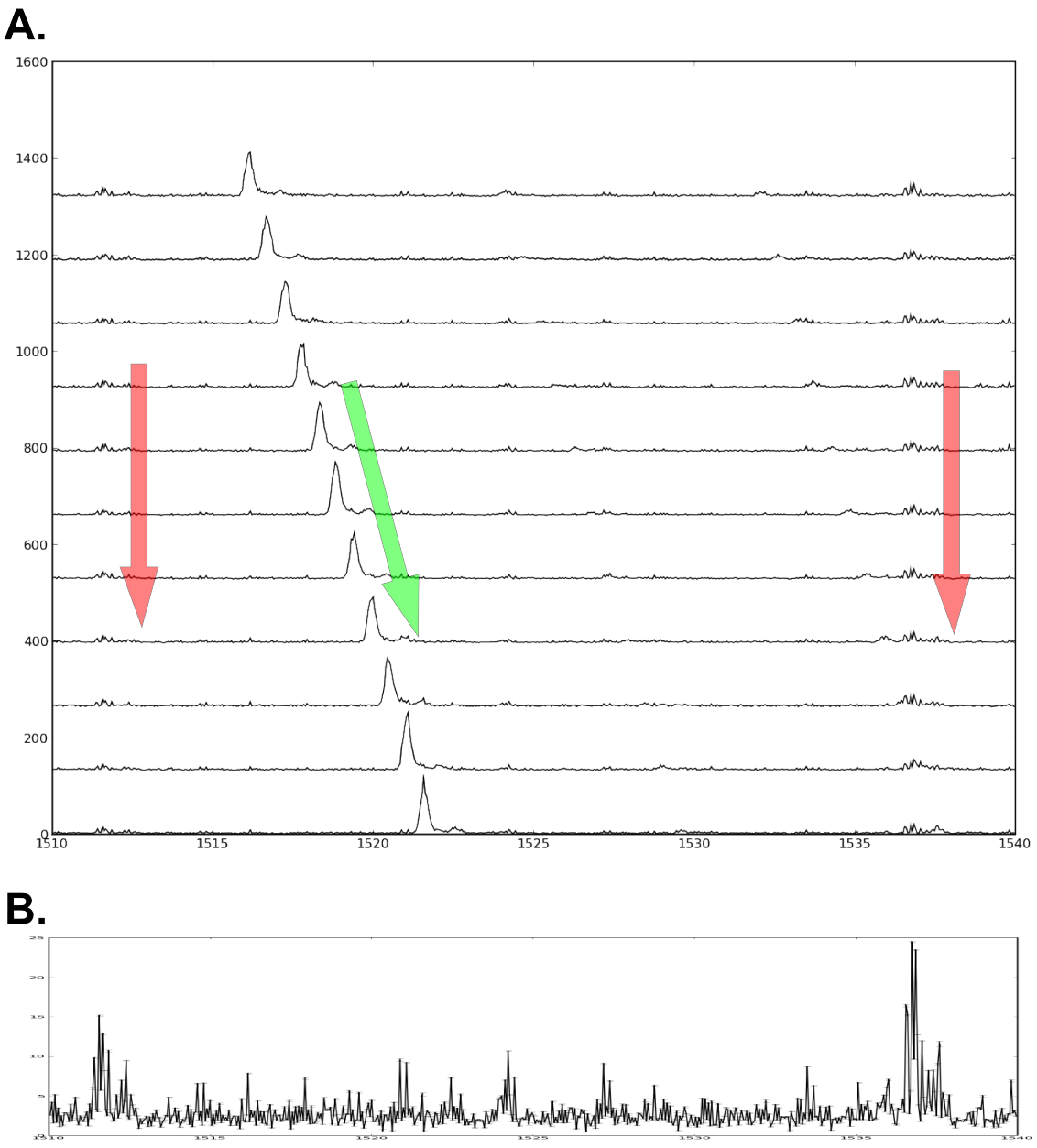


Figure 6.4: Physical signal modulation prior to spectral alignment shows coherent noise and highly predictable behavior. A.) True ion peaks shift across the m/z axis as the reflector voltage is altered (trend indicated in green). Coherent detector noise remains constant in all spectra (to left of red arrows). B.) For each m/z sampling location, removal of outliers from the set of spectra produces the average coherent noise spectra. The y-axis is enlarged relative to A. The cluster size for this spectrum is three samples. Error bars representing one standard deviation are shown on all peaks, but are extremely small. The high precision of background noise determination aids in accurately removing noise from low-intensity peaks.

was not possible to exclude all real peaks from the noise set, and the addition of two more noise spectra to the set did not improve precision of the noise estimates.

There are several strategies for dealing with the coherent noise once it has been calculated. One option is to immediately subtract it from each spectrum prior to spectral alignment. Alternatively, because averaging the demodulated spectra tends to average out noise, coherent noise subtraction can be omitted prior to spectral alignment, but this comes at the cost of signal quality for low-intensity peaks. Figure 6.5a shows a closeup of a highly-averaged spectra before any PSM processing. This spectra is the result of over a minute of data collection, and has no noise as calculated by the Blom metric (additional averaging will not improve it). However, true signal-to-noise in this region is extremely low, and coherent noise completely masks any real peaks that might be present. In comparison, Figure 6.5c shows the result of PSM and denoising: the background level is reduced to less than 1 unit of intensity, and peaks with intensity less than 10 can be clearly seen: due to the extremely high data density of PSM-processed spectra, outliers are inevitable (and visible in the figure), but they are inconsequential due to the large numbers of low-deviation data samples

Improvements in Δx

Besides the noise-reduction effects of PSM, this technique also improves signal quality by reducing the spacing between data points. As discussed in the Introduction and Background, there is a statistical uncertainly in mass accuracy and precision arising from the sampling interval: for a 500 m/z peak with a resolution of 10,000 and measured with a 2GHz digitizer, peak-calculation has a +/- 2.4 ppm error if the spectrum is perfectly noiseless, and peak-fitting has a 5 ppm error if the S/N ratio is 100 or less (95% CI level). By applying PSM to collect spectra with varying m/z sampling intervals, the Δx parameter can be arbitrarily reduced to achieve the desired

degree of precision. Figure 6.5b shows a low-m/z peak from a typical raw spectrum. As in Figure 5a, it is highly averaged, although coherent noise has not been removed. The same peak is much more highly sampled after PSM (Figure 6.5d), permitting greater precision in both peak-calculation and peak-fitting.

Figure 6.5b also illustrates the problem of calculating peak positions from peaks which are not fully resolved. There are only three data points within the FW20M cutoff threshold, which is barely enough to fit a single peak shape, much less the possibility of two overlapping peaks. In our study, we have improved peak sampling by approximately a factor of 10, allowing direct visualization of peak shapes even at low m/z values. This permits both qualitative detection of overlapping peaks, as well as appropriate input data to fit such a system with equations describing overlapping peaks [134].

Conclusions

The complexity of fully evaluating mass accuracy and precision has long been an impediment to cohesive reporting of mass spectrometry results. Even within the subfield of proteomics, Zubarev and Mann have identified three different usages of mass accuracy alone [145]. Our investigations of Δx and its relation to high mass accuracy measurements show that even informal assumptions regarding mass measurement precision may be invalid for current instrumentation. Specifically, the precision of mass measurements on a typical TOF mass analyzer far surpasses the true statistical accuracy, and systematic errors need to be addressed in order to achieve routine mass accuracy at or below 5 ppm. A major component of this systematic error is the use of centroids to calculate peak location, and even in ideal synthetic spectra, we find that centroids create mass errors due to peak asymmetry and the binary noise of data point

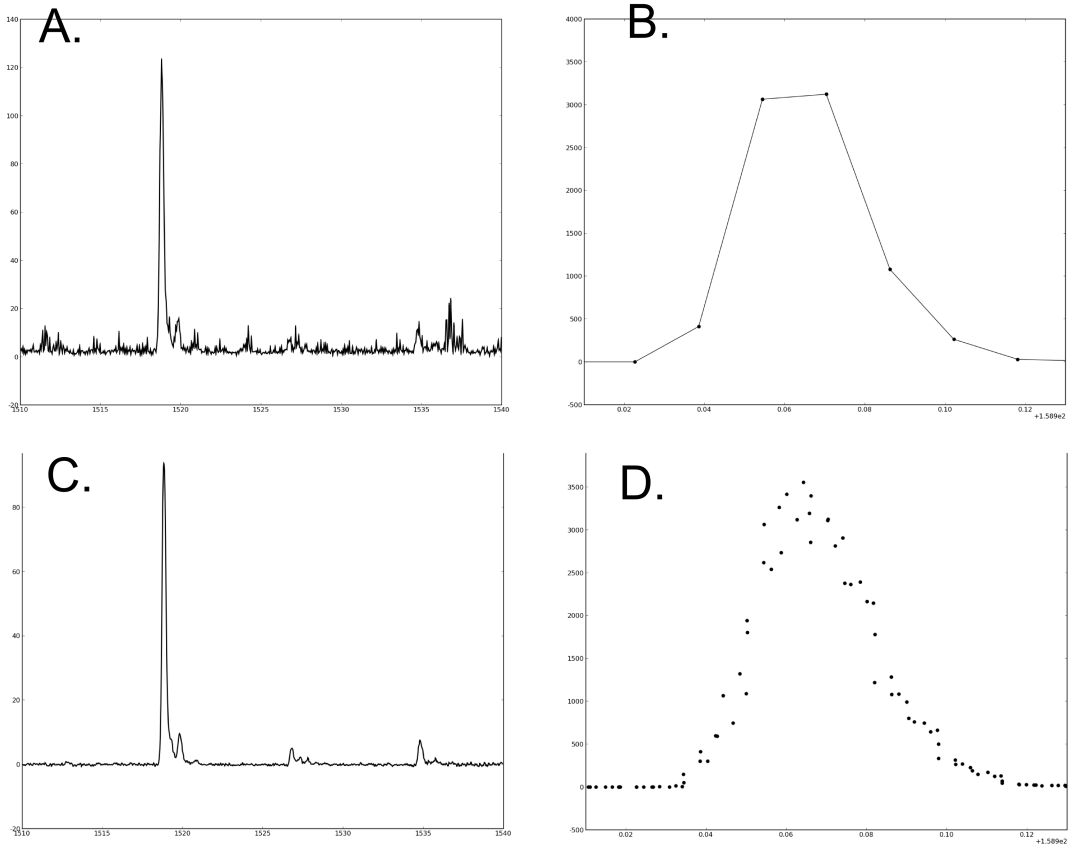


Figure 6.5: Comparisons between raw spectra and PSM-processed spectra. A) Closeup of peaks with very low S/N, from a highly averaged spectrum. High-intensity coherent noise is visible at 1537 m/z. Neither the real peak at 1535 nor the noise at 1537 improves with additional averaging (scan-to-scan intensity errors are too small to visualize). C.) Same region as in (A) but after PSM processing. A nearly perfect baseline is recovered, and coherent noise is removed. Data sampling is extremely dense (∼900 points shown) and is unfiltered to show unaltered data: for discussion see text. B.) Low m/z peak suffering from lack of sufficient sample points on the m/z axis. This peak has very large S/N (∼600) and intensity, but has a centroiding extreme-spread error of 7.7 ppm. D.) Same region as in (B) after processing. The experimentally-determined triple-Gaussian model is shown fit to the data.

inclusion/exclusion. Furthermore, the errors of centroid calculation are not randomly distributed and produce unusual artifacts when used for lock mass corrections. Due to the prevalence of the lock mass technique in many variants, this is a troubling discovery and merits further investigation.

Peak fitting is an alternative to using centroids (peak-calculation) for peak assignment. Peak-fitting has been predicted to provide more precise results [132] and our findings support this theory. In addition, the use of peak-fitting removes substantial systematic errors in accuracy. The accuracy and precision of both peak-calculation and peak-fitting depend highly upon the data point spacing Δx, and modern TOF instrumentation reflects a balance between improving resolution, which increases effective Δx, and digitizer speed, which decreases it.

For TOF mass analyzers, physical signal modulation provides an easy approach to reduce Δx increments, and has the additional benefit of removing coherent spectral noise. This technique can be carried out on any TOF mass analyzer without the need for hardware upgrades, and effectively extends the physical limits of the detector/digitizer system. Because all downstream processing can be fully automated via our PySpecTools software package, the PSM technique should be highly amenable to high-throughput analysis of high mass measurement accuracy samples. Efforts are currently ongoing to adapt the PSM approach to mass spectrometry instrument control software, removing the need for manual adjustment of TOF parameters, permitting automated data collection, noise reduction, and spectral alignment as part of the normal data collection process.

REFERENCES CITED

[1] A. C. Steven, J. B. Heymann, N. Cheng, B. L. Trus, J. F. Conway. Virus maturation: dynamics and mechanism of a stabilizing structural transition that leads to infectivity. *Curr Opin Struct Biol*, 15(2):227–36, 2005.

[2] P.E. Prevelige. *The Bacteriophages*. Oxford University Press, New York, 2005.

[3] J. F. Conway, W. R. Wikoff, N. Cheng, R. L. Duda, R. W. Hendrix, J. E. Johnson, A. C. Steven. Virus maturation involving large subunit rotations and local refolding. *Science*, 292(5517):744–8, 2001.

[4] M. A. Canady, H. Tsuruta, J. E. Johnson. Analysis of rapid, large-scale protein quaternary structural changes: time-resolved x-ray solution scattering of nudaurelia capensis omega virus (nomegav) maturation. *J Mol Biol*, 311(4):803–14, 2001.

[5] C. E. Fricks, J. M. Hogle. Cell-induced conformational change in poliovirus: externalization of the amino terminus of vp1 is responsible for liposome binding. *J Virol*, 64(5):1934–45, 1990.

[6] W. R. Wikoff, L. Liljas, R. L. Duda, H. Tsuruta, R. W. Hendrix, J. E. Johnson. Topologically linked protein rings in the bacteriophage hk97 capsid. *Science*, 289(5487):2129–33, 2000.

[7] R. L. Duda. Protein chainmail: catenated protein in viral capsids. *Cell*, 94(1):55–60, 1998.

[8] R. Lata, J. F. Conway, N. Cheng, R. L. Duda, R. W. Hendrix, W. R. Wikoff, J. E. Johnson, H. Tsuruta, A. C. Steven. Maturation dynamics of a viral capsid: visualization of transitional intermediate states. *Cell*, 100(2):253–63, 2000.

[9] J. B. Heymann, N. Cheng, W. W. Newcomb, B. L. Trus, J. C. Brown, A. C. Steven. Dynamics of herpes simplex virus capsid maturation visualized by time-lapse cryo-electron microscopy. *Nat Struct Biol*, 10(5):334–41, 2003.

[10] A. Zlotnick, V. S. Reddy, R. Dasgupta, A. Schneemann, Jr. Ray, W. J., R. R. Rueckert, J. E. Johnson. Capsid assembly in a family of animal viruses primes an autoproteolytic maturation that depends on a single aspartic acid residue. *J Biol Chem*, 269(18):13680–4, 1994.

[11] A. Schneemann, R. Dasgupta, J. E. Johnson, R. R. Rueckert. Use of recombinant baculoviruses in synthesis of morphologically distinct viruslike particles of flock house virus, a nodavirus. *J Virol*, 67(5):2756–63, 1993.

[12] A. Janshoff, D. T. Bong, C. Steinem, J. E. Johnson, M. R. Ghadiri. An animal virus-derived peptide switches membrane morphology: possible relevance to nodaviral transfection processes. *Biochemistry*, 38(17):5328–36, 1999.

[13] M. Chow, R. Yabrov, J. Bittle, J. Hogle, D. Baltimore. Synthetic peptides from four separate regions of the poliovirus type 1 capsid protein vp1 induce neutralizing antibodies. *Proc Natl Acad Sci U S A*, 82(3):910–4, 1985.

[14] C. Helgstrand, S. Munshi, J. E. Johnson, L. Liljas. The refined structure of nudaurelia capensis omega virus reveals control elements for a t = 4 capsid maturation. *Virology*, 318(1):192–203, 2004.

[15] S. Munshi, L. Liljas, J. Cavarelli, W. Bomu, B. McKinney, V. Reddy, J. E. Johnson. The 2.8 a structure of a t = 4 animal virus and its implications for membrane translocation of rna. *J Mol Biol*, 261(1):1–10, 1996.

[16] M. A. Canady, M. Tihova, T. N. Hanzlik, J. E. Johnson, M. Yeager. Large conformational changes in the maturation of a simple rna virus, nudaurelia capensis omega virus (nomegav). *J Mol Biol*, 299(3):573–84, 2000.

[17] M. Tomasicchio, P. A. Venter, K. H. Gordon, T. N. Hanzlik, R. A. Dorrington. Induction of apoptosis in saccharomyces cerevisiae results in the spontaneous maturation of tetravirus procapsids in vivo. *J Gen Virol*, 88(Pt 5):1576–82, 2007.

[18] D. J. Taylor, N. K. Krishna, M. A. Canady, A. Schneemann, J. E. Johnson. Large-scale, ph-dependent, quaternary structure changes in an rna virus capsid are reversible in the absence of subunit autoproteolysis. *J Virol*, 76(19):9972–80, 2002.

[19] K. K. Lee, J. Tang, D. Taylor, B. Bothner, J. E. Johnson. Small compounds targeted to subunit interfaces arrest maturation in a nonenveloped, icosahedral animal virus. *J Virol*, 78(13):7208–16, 2004.

[20] D. J. Taylor, Q. Wang, B. Bothner, P. Natarajan, M. G. Finn, J. E. Johnson. Correlation of chemical reactivity of nudaurelia capensis omega virus with a ph-induced conformational change. *Chem Commun (Camb)*, (22):2770–1, 2003.

[21] B. Bothner, D. Taylor, B. Jun, K. K. Lee, G. Siuzdak, C. P. Schultz, J. E. Johnson. Maturation of a tetravirus capsid alters the dynamic properties and creates a metastable complex. *Virology*, 334(1):17–27, 2005.

[22] J. M. Hogle. Poliovirus cell entry: common structural themes in viral cell entry pathways. *Annu Rev Microbiol*, 56:677–702, 2002.

[23] G. A. Farr, L. G. Zhang, P. Tattersall. Parvoviral virions deploy a capsid-tethered lipolytic enzyme to breach the endosomal membrane during cell entry. *Proc Natl Acad Sci U S A*, 102(47):17148–53, 2005.

[24] M. Agbandje, R. McKenna, M. G. Rossmann, M. L. Strassheim, C. R. Parrish. Structure determination of feline panleukopenia virus empty particles. *Proteins*, 16(2):155–71, 1993.

[25] S. F. Cotmore, Jr. D'Abramo A, M., C. M. Ticknor, P. Tattersall. Controlled conformational transitions in the mvm virion expose the vp1 n-terminus and viral genome without particle disassembly. *Virology*, 254(1):169–81, 1999.

[26] F. Sonntag, S. Bleker, B. Leuchs, R. Fischer, J. A. Kleinschmidt. Adeno-associated virus type 2 capsids with externalized vp1/vp2 trafficking domains are generated prior to passage through the cytoplasm and are maintained until uncoating occurs in the nucleus. *J Virol*, 80(22):11040–54, 2006.

[27] J. K. Lewis, B. Bothner, T. J. Smith, G. Siuzdak. Antiviral agent blocks breathing of the common cold virus. *Proc Natl Acad Sci U S A*, 95(12):6774–8, 1998.

[28] G. Vriend, M. A. Hemminga, B. J. M. Verduin, J. L. Dewit, T. J. Schaafsma. Segmental mobility involved in protein-rna interaction in cowpea chlorotic mottle virus. *Febs Letters*, 134(2):167–171, 1981.

[29] M. Roivainen, A. Narvanen, M. Korkolainen, M. L. Huhtala, T. Hovi. Antigenic regions of poliovirus type 3/sabin capsid proteins recognized by human sera in the peptide scanning technique. *Virology*, 180(1):99–107, 1991.

[30] M. Roivainen, L. Piirainen, T. Rysa, A. Narvanen, T. Hovi. An immunodominant n-terminal region of vp1 protein of poliovirion that is buried in crystal structure can be exposed in solution. *Virology*, 195(2):762–5, 1993.

[31] Q. Li, A. G. Yafal, Y. M. Lee, J. Hogle, M. Chow. Poliovirus neutralization by antibodies to internal epitopes of vp4 and vp1 results from reversible exposure of these sequences at physiological temperature. *J Virol*, 68(6):3965–70, 1994.

[32] M. P. Fox, M. J. Otto, M. A. McKinlay. Prevention of rhinovirus and poliovirus uncoating by win-51711, a new antiviral drug. *Antimicrobial Agents And Chemotherapy*, 30(1):110–116, 1986.

[33] D. K. Phelps, C. B. Post. A novel basis of capsid stabilization by antiviral compounds. *J Mol Biol*, 254(4):544–51, 1995.

[34] D. K. Phelps, C. B. Post. Molecular dynamics investigation of the effect of an antiviral compound on human rhinovirus. *Protein Sci*, 8(11):2281–9, 1999.

[35] B. Bothner, X. F. Dong, L. Bibbs, J. E. Johnson, G. Siuzdak. Evidence of viral capsid dynamics using limited proteolysis and mass spectrometry. *J Biol Chem*, 273(2):673–6, 1998.

[36] B. R. Szymczyna, L. Gan, J. E. Johnson, J. R. Williamson. Solution nmr studies of the maturation intermediates of a 13 mda viral capsid. *J Am Chem Soc*, 129(25):7867–76, 2007.

[37] S. J. Opella, A. C. Zeri, S. H. Park. Structure, dynamics, and assembly of filamentous bacteriophages by nuclear magnetic resonance spectroscopy. *Annu Rev Phys Chem*, 59:635–57, 2008.

[38] B. Bothner, A. Schneemann, D. Marshall, V. Reddy, J. E. Johnson, G. Siuzdak. Crystallographically identical virus capsids display different properties in solution. *Nat Struct Biol*, 6(2):114–6, 1999.

[39] A. T. Da Poian, J. E. Johnson, J. L. Silva. Protein-rna interactions and virus stability as probed by the dynamics of tryptophan side chains. *J Biol Chem*, 277(49):47596–602, 2002.

[40] R. Tuma, J. K. Bamford, D. H. Bamford, Jr. Thomas, G. J. Assembly dynamics of the nucleocapsid shell subunit (p8) of bacteriophage phi6. *Biochemistry*, 38(45):15025–33, 1999.

[41] R. Tuma, J. H. Bamford, D. H. Bamford, M. P. Russell, Jr. Thomas, G. J. Structure, interactions and dynamics of prd1 virus i. coupling of subunit folding and capsid assembly. *J Mol Biol*, 257(1):87–101, 1996.

[42] Jr. Prevelige, P. E., D. Thomas, K. L. Aubrey, S. A. Towse, Jr. Thomas, G. J. Subunit conformational changes accompanying bacteriophage p22 capsid maturation. *Biochemistry*, 32(2):537–43, 1993.

[43] D. A. Kuzmanovic, I. Elashvili, C. Wick, C. O'Connell, S. Krueger. Bacteriophage ms2: molecular weight and spatial distribution of the protein and rna components by small-angle neutron scattering and virus counting. *Structure*, 11(11):1339–48, 2003.

[44] D. A. Kuzmanovic, I. Elashvili, C. Wick, C. O'Connell, S. Krueger. The ms2 coat protein shell is likely assembled under tension: a novel role for the ms2 bacteriophage a protein as revealed by small-angle neutron scattering. *J Mol Biol*, 355(5):1095–111, 2006.

[45] P. L. Freddolino, A. S. Arkhipov, S. B. Larson, A. McPherson, K. Schulten. Molecular dynamics simulations of the complete satellite tobacco mosaic virus. *Structure*, 14(3):437–49, 2006.

[46] K. Y. Sanbonmatsu, C. S. Tung. High performance computing in biology: multimillion atom simulations of nanoscale systems. *J Struct Biol*, 157(3):470–80, 2007.

[47] M. K. Kim, R. L. Jernigan, G. S. Chirikjian. An elastic network model of hk97 capsid maturation. *J Struct Biol*, 143(2):107–17, 2003.

[48] F. Tama, 3rd Brooks, C. L. The mechanism and pathway of ph induced swelling in cowpea chlorotic mottle virus. *J Mol Biol*, 318(3):733–47, 2002.

[49] F. Tama, O. Miyashita, 3rd Brooks, C. L. Normal mode based flexible fitting of high-resolution structure into low-resolution experimental data from cryo-em. *J Struct Biol*, 147(3):315–26, 2004.

[50] E. C. Dykeman, O. F. Sankey, K. T. Tsen. Raman intensity and spectra predictions for cylindrical viruses. *Phys Rev E Stat Nonlin Soft Matter Phys*, 76(1 Pt 1):011906, 2007.

[51] K. M. ElSawy, A. Taormina, R. Twarock, L. Vaughan. Dynamical implications of viral tiling theory. *J Theor Biol*, 252(2):357–69, 2008.

[52] K. T. Tsen, E. C. Dykeman, O. F. Sankey, S. W. Tsen, N. T. Lin, J. G. Kiang. Probing the low-frequency vibrational modes of viruses with raman scattering–bacteriophage m13 in water. *J Biomed Opt*, 12(2):024009, 2007.

[53] F. Englert, K. Peeters, A. Taormina. Twenty-four near-instabilities of casparklug viruses. *Phys Rev E Stat Nonlin Soft Matter Phys*, 78(3 Pt 1):031908, 2008.

[54] J. Lanman, T. T. Lam, S. Barnes, M. Sakalian, M. R. Emmett, A. G. Marshall, Jr. Prevelige, P. E. Identification of novel interactions in hiv-1 capsid protein assembly by high-resolution mass spectrometry. *J Mol Biol*, 325(4):759–72, 2003.

[55] S. Kang, A. M. Hawkridge, K. L. Johnson, D. C. Muddiman, Jr. Prevelige, P. E. Identification of subunit-subunit interactions in bacteriophage p22 procapsids by chemical cross-linking and mass spectrometry. *J Proteome Res*, 5(2):370–7, 2006.

[56] R. Tuma, H. Tsuruta, J. M. Benevides, Jr. Prevelige, P. E., Jr. Thomas, G. J. Characterization of subunit structural changes accompanying assembly of the bacteriophage p22 procapsid. *Biochemistry*, 40(3):665–74, 2001.

[57] T. E. Wales, J. R. Engen. Hydrogen exchange mass spectrometry for the analysis of protein dynamics. *Mass Spectrom Rev*, 25(1):158–70, 2006.

[58] J. Lanman, T. T. Lam, M. R. Emmett, A. G. Marshall, M. Sakalian, Jr. Prevelige, P. E. Key interactions in hiv-1 maturation identified by hydrogen-deuterium exchange. *Nat Struct Mol Biol*, 11(7):676–7, 2004.

[59] B. Suchanova, R. Tuma. Folding and assembly of large macromolecular complexes monitored by hydrogen-deuterium exchange and mass spectrometry. *Microb Cell Fact*, 7:12, 2008.

[60] L. Wang, D. L. Smith. Capsid structure and dynamics of a human rhinovirus probed by hydrogen exchange mass spectrometry. *Protein Sci*, 14(6):1661–72, 2005.

[61] C. Y. Fu, Jr. Prevelige, P. E. Dynamic motions of free and bound o29 scaffolding protein identified by hydrogen deuterium exchange mass spectrometry. *Protein Sci*, 15(4):731–43, 2006.

[62] J. K. Hilmer, A. Zlotnick, B. Bothner. Conformational equilibria and rates of localized motion within hepatitis b virus capsids. *J Mol Biol*, 375(2):581–94, 2008.

[63] C. R. Bourne, M. Finn, A. Zlotnick. Global structural changes in hepatitis b virus capsids induced by the assembly effector hap1. *J Virol*, 2006.

[64] L. M. Stannard, M. Hodgkiss. Morphological irregularities in dane particle cores. *J Gen Virol*, 45(2):509–14, 1979.

[65] S. M. Freund, C. M. Johnson, A. M. Jaulent, N. Ferguson. Moving towards high-resolution descriptions of the molecular interactions and structural rearrangements of the human hepatitis b core protein. *J Mol Biol*, 384(5):1301–13, 2008.

[66] K. A. Dryden, S. F. Wieland, C. Whitten-Bauer, J. L. Gerin, F. V. Chisari, M. Yeager. Native hepatitis b virions and capsids visualized by electron cryomicroscopy. *Mol Cell*, 22(6):843–50, 2006.

[67] S. Zhou, D. N. Standring. Hepatitis b virus capsid particles are assembled from core-protein dimer precursors. *Proc Natl Acad Sci U S A*, 89(21):10046–50, 1992.

[68] B. Rabe, M. Delaleau, A. Bischof, M. Foss, I. Sominskaya, P. Pumpens, C. Cazenave, M. Castroviejo, M. Kann. Nuclear entry of hepatitis b virus capsids involves disintegration to protein dimers followed by nuclear reassociation to capsids. *PLoS Pathog*, 5(8):e1000563, 2009.

[69] M. Seifer, D. N. Standring. Assembly and antigenicity of hepatitis b virus core particles. *Intervirology*, 38(1-2):47–62, 1995.

[70] P. Ceres, A. Zlotnick. Weak protein-protein interactions are sufficient to drive assembly of hepatitis b virus capsids. *Biochemistry*, 41(39):11525–31, 2002.

[71] S. J. Stray, P. Ceres, A. Zlotnick. Zinc ions trigger conformational change and oligomerization of hepatitis b virus capsid protein. *Biochemistry*, 43(31):9989–98, 2004.

[72] P. Ceres, S. J. Stray, A. Zlotnick. Hepatitis b virus capsid assembly is enhanced by naturally occurring mutation f97l. *J Virol*, 78(17):9538–43, 2004.

[73] S. J. Stray, A. Zlotnick. Bay 41-4109 has multiple effects on hepatitis b virus capsid assembly. *J Mol Recognit*, 2006.

[74] C. Bourne, S. Lee, B. Venkataiah, A. Lee, B. Korba, M. G. Finn, A. Zlotnick. Small-molecule effectors of hepatitis b virus capsid assembly give insight into virus life cycle. *J Virol*, 82(20):10262–70, 2008.

[75] A. Zlotnick, P. Ceres, S. Singh, J. M. Johnson. A small molecule inhibits and misdirects assembly of hepatitis b virus capsids. *J Virol*, 76(10):4848–54, 2002.

[76] C. R. Bourne, M. Finn, A. Zlotnick. Global structural changes in hepatitis b virus capsids induced by the assembly effector hap1. *J Virol*, 2006.

[77] B. Bottcher, M. Vogel, M. Ploss, M. Nassal. High plasticity of the hepatitis b virus capsid revealed by conformational stress. *J Mol Biol*, 356(3):812–22, 2006.

[78] A. Harris, D. M. Belnap, N. R. Watts, J. F. Conway, N. Cheng, S. J. Stahl, J. G. Vethanayagam, P. T. Wingfield, A. C. Steven. Epitope diversity of hepatitis b virus capsids: quasi-equivalent variations in spike epitopes and binding of different antibodies to the same epitope. *J Mol Biol*, 355(3):562–76, 2006.

[79] D. M. Belnap, N. R. Watts, J. F. Conway, N. Cheng, S. J. Stahl, P. T. Wingfield, A. C. Steven. Diversity of core antigen epitopes of hepatitis b virus. *Proc Natl Acad Sci U S A*, 100(19):10884–9, 2003.

[80] J. F. Conway, N. R. Watts, D. M. Belnap, N. Cheng, S. J. Stahl, P. T. Wingfield, A. C. Steven. Characterization of a conformational epitope on hepatitis b virus core antigen and quasiequivalent variations in antibody binding. *J Virol*, 77(11):6466–73, 2003.

[81] P. Pumpens, E. Grens. Hepatitis b core particles as a universal display model: a structure-function basis for development. *FEBS Lett*, 442(1):1–6, 1999.

[82] J. F. Conway, N. Cheng, A. Zlotnick, S. J. Stahl, P. T. Wingfield, D. M. Belnap, U. Kanngiesser, M. Noah, A. C. Steven. Hepatitis b virus capsid: localization of the putative immunodominant loop (residues 78 to 83) on the capsid surface, and implications for the distinction between c and e-antigens. *J Mol Biol*, 279(5):1111–21, 1998.

[83] J. F. Conway, N. Cheng, A. Zlotnick, P. T. Wingfield, S. J. Stahl, A. C. Steven. Visualization of a 4-helix bundle in the hepatitis b virus capsid by cryo-electron microscopy. *Nature*, 386(6620):91–4, 1997.

[84] B. Bottcher, S. A. Wynne, R. A. Crowther. Determination of the fold of the core protein of hepatitis b virus by electron cryomicroscopy. *Nature*, 386(6620):88–91, 1997.

[85] R. A. Crowther, N. A. Kiselev, B. Bottcher, J. A. Berriman, G. P. Borisova, V. Ose, P. Pumpens. Three-dimensional structure of hepatitis b virus core particles determined by electron cryomicroscopy. *Cell*, 77(6):943–50, 1994.

[86] S. A. Wynne, R. A. Crowther, A. G. Leslie. The crystal structure of the human hepatitis b virus capsid. *Mol Cell*, 3(6):771–80, 1999.

[87] A. Zlotnick, N. Cheng, S. J. Stahl, J. F. Conway, A. C. Steven, P. T. Wingfield. Localization of the c terminus of the assembly domain of hepatitis b virus capsid protein: implications for morphogenesis and organization of encapsidated rna. *Proc Natl Acad Sci U S A*, 94(18):9556–61, 1997.

[88] A. Zlotnick, J. M. Johnson, P. W. Wingfield, S. J. Stahl, D. Endres. A theoretical model successfully identifies features of hepatitis b virus capsid assembly. *Biochemistry*, 38(44):14644–52, 1999.

[89] S. Singh, A. Zlotnick. Observed hysteresis of virus capsid disassembly is implicit in kinetic models of assembly. *J Biol Chem*, 278(20):18249–55, 2003.

[90] O. Weber, K. H. Schlemmer, E. Hartmann, I. Hagelschuer, A. Paessens, E. Graef, K. Deres, S. Goldmann, U. Niewoehner, J. Stoltefuss, D. Haebich, H. Ruebsamen-Waigmann, S. Wohlfeil. Inhibition of human hepatitis b virus (hbv) by a novel non-nucleosidic compound in a transgenic mouse model. *Antiviral Res*, 54(2):69–78, 2002.

[91] J. Salfeld, E. Pfaff, M. Noah, H. Schaller. Antigenic determinants and functional domains in core antigen and e antigen from hepatitis b virus. *J Virol*, 63(2):798–808, 1989.

[92] N. D. Rawlings, A. J. Barrett, A. Bateman. Merops: the peptidase database. *Nucleic Acids Res*, 2009.

[93] C. Park, S. Marqusee. Probing the high energy states in proteins by proteolysis. *J Mol Biol*, 343(5):1467–76, 2004.

[94] C. Park, S. Marqusee. Pulse proteolysis: a simple method for quantitative determination of protein stability and ligand binding. *Nat Methods*, 2(3):207–12, 2005.

[95] J. K. Hilmer, A. Zlotnick, B. Bothner. Conformational equilibria and rates of localized motion within hepatitis b virus capsids. *J Mol Biol*, 375(2):581–94, 2008.

[96] C. Park, B. R. Kelemen, T. A. Klink, R. Y. Sweeney, M. A. Behlke, S. R. Eubanks, R. T. Raines. Fast, facile, hypersensitive assays for ribonucleolytic activity. *Methods Enzymol*, 341:81–94, 2001.

[97] D. Xie, L. Suvorov, J. W. Erickson, A. S. Gulnik. Real-time measurements of dark substrate catalysis. *Protein Sci*, 8(11):2460–4, 1999.

[98] C. Park, S. Zhou, J. Gilmore, S. Marqusee, Q. F. Ma, J. Hu, W. H. Wu, H. D. Liu, J. T. Du, Y. Fu, Y. W. Wu, P. Lei, Y. F. Zhao, Y. M. Li, Y. Ye, H. W. Lee, W. Yang, S. J. Shealy, A. L. Wilkins, Z. R. Liu, I. Torshin, R. Harrison, R. Wohlhueter, J. J. Yang, T. Kowalik-Jankowska, M. Ruta-Dolejsz, K. Wisniewska, L. Lankiewicz, M. Yu, J. Summers. Energetics-based protein profiling on a proteomic scale: identification of proteins resistant to proteolysis characterization of copper binding to the peptide amyloid-beta(1-16) associated with alzheimer's disease metal binding affinity and structural properties of an isolated ef-loop in a scaffold protein cu(ii) interaction with n-terminal fragments of human and mouse beta-amyloid peptide multiple functions of capsid protein phosphorylation in duck hepatitis b virus replication. *J Mol Biol*, 368(5):1426–37, 2007.

[99] T. Komiyama, H. Suda, T. Aoyagi, T. Takeuchi, H. Umezawa. Studies on inhibitory effect of phosphoramidon and its analogs on thermolysin. *Arch Biochem Biophys*, 171(2):727–31, 1975.

[100] M. M. Vorob'ev, M. Dalgalarrondo, J. M. Chobert, T. Haertle. Kinetics of beta-casein hydrolysis by wild-type and engineered trypsin. *Biopolymers*, 54(5):355–64, 2000.

[101] C. R. Bourne, S. P. Katen, M. R. Fulz, C. Packianathan, A. Zlotnick. A mutant hepatitis b virus core protein mimics inhibitors of icosahedral capsid self-assembly (dagger). *Biochemistry*, 2009.

[102] N. Ogata, R. H. Miller, K. G. Ishak, R. H. Purcell. The complete nucleotide sequence of a pre-core mutant of hepatitis b virus implicated in fulminant hepatitis and its biological characterization in chimpanzees. *Virology*, 194(1):263–76, 1993.

[103] D. Ganem, R.J. Schneider. *Hepadnaviridae and their replication.* Fields Virology. Lippincott-Raven, Philadelphia, 2001.

[104] M. Qiao, T. B. Macnaughton, E. J. Gowans. Adsorption and penetration of hepatitis b virus in a nonpermissive cell line. *Virology*, 201(2):356–63, 1994.

[105] M. Kann, B. Sodeik, A. Vlachou, W. H. Gerlich, A. Helenius. Phosphorylation-dependent binding of hepatitis b virus core particles to the nuclear pore complex. *J Cell Biol*, 145(1):45–55, 1999.

[106] A. Machida, H. Ohnuma, F. Tsuda, A. Yoshikawa, Y. Hoshi, T. Tanaka, S. Kishimoto, Y. Akahane, Y. Miyakawa, M. Mayumi. Phosphorylation in the carboxyl-terminal domain of the capsid protein of hepatitis b virus: evaluation with a monoclonal antibody. *J Virol*, 65(11):6024–30, 1991.

[107] R. J. Gilbert, L. Beales, D. Blond, M. N. Simon, B. Y. Lin, F. V. Chisari, D. I. Stuart, D. J. Rowlands. Hepatitis b small surface antigen particles are octahedral. *Proc Natl Acad Sci U S A*, 102(41):14783–8, 2005.

[108] Y. T. Lan, J. Li, W. Liao, J. Ou. Roles of the three major phosphorylation sites of hepatitis b virus core protein in viral replication. *Virology*, 259(2):342–8, 1999.

[109] M. Kann, W. H. Gerlich. Effect of core protein phosphorylation by protein kinase c on encapsidation of rna within core particles of hepatitis b virus. *J Virol*, 68(12):7993–8000, 1994.

[110] M. Seifer, S. Zhou, D. N. Standring. A micromolar pool of antigenically distinct precursors is required to initiate cooperative assembly of hepatitis b virus capsids in xenopus oocytes. *J Virol*, 67(1):249–57, 1993.

[111] C. T. Yeh, Y. F. Liaw, J. H. Ou. The arginine-rich domain of hepatitis b virus precore and core proteins contains a signal for nuclear transport. *J Virol*, 64(12):6141–7, 1990.

[112] M. Nassal. The arginine-rich domain of the hepatitis b virus core protein is required for pregenome encapsidation and productive viral positive-strand dna synthesis but not for virus assembly. *J Virol*, 66(7):4107–16, 1992.

[113] H. Y. Kang, S. Lee, S. G. Park, J. Yu, Y. Kim, G. Jung. Phosphorylation of hepatitis b virus cp at ser87 facilitates core assembly. *Biochem J*, 398(2):311–7, 2006.

[114] H. Kang, J. Yu, G. Jung. Phosphorylation of hepatitis b virus core c-terminally truncated protein (cp149) by pkc increases capsid assembly and stability. *Biochem J*, 416(1):47–54, 2008.

[115] K. Takahashi, A. Machida, G. Funatsu, M. Nomura, S. Usuda, S. Aoyagi, K. Tachibana, H. Miyamoto, M. Imai, T. Nakamura, Y. Miyakawa, M. Mayumi. Immunochemical structure of hepatitis b e antigen in the serum. *J Immunol*, 130(6):2903–7, 1983.

[116] D. Milich, T. J. Liang. Exploring the biological basis of hepatitis b e antigen in hepatitis b virus infection. *Hepatology*, 38(5):1075–86, 2003.

[117] M. Nassal, A. Rieger. An intramolecular disulfide bridge between cys-7 and cys61 determines the structure of the secretory core gene product (e antigen) of hepatitis b virus. *J Virol*, 67(7):4307–15, 1993.

[118] G. Wasenauer, J. Kock, H. J. Schlicht. Relevance of cysteine residues for biosynthesis and antigenicity of human hepatitis b virus e protein. *J Virol*, 67(3):1315–21, 1993.

[119] A. Zlotnick, N. Cheng, J. F. Conway, F. P. Booy, A. C. Steven, S. J. Stahl, P. T. Wingfield. Dimorphism of hepatitis b virus capsids is strongly influenced by the c-terminus of the capsid protein. *Biochemistry*, 35(23):7412–21, 1996.

[120] H. Ohori, M. Yamaki, S. Onodera, E. Yamada, N. Ishida. Antigenic conversion from hbcag to hbeag by degradation of hepatitis b core particles. *Intervirology*, 13(2):74–82, 1980.

[121] M. A. Baumeister, A. Medina-Selby, D. Coit, S. Nguyen, C. George-Nascimento, A. Gyenes, P. Valenzuela, G. Kuo, D. Y. Chien. Hepatitis b virus e antigen specific epitopes and limitations of commercial anti-hbe immunoassays. *J Med Virol*, 60(3):256–63, 2000.

[122] D. R. Milich, J. E. Jones, J. L. Hughes, T. Maruyama, J. Price, I. Melhado, F. Jirik. Extrathymic expression of the intracellular hepatitis b core antigen results in t cell tolerance in transgenic mice. *J Immunol*, 152(2):455–66, 1994.

[123] Y. H. Shi, C. H. Shi. Molecular characteristics and stages of chronic hepatitis b virus infection. *World J Gastroenterol*, 15(25):3099–105, 2009.

[124] S. Tong, K. H. Kim, C. Chante, J. Wands, J. Li. Hepatitis b virus e antigen variants. *Int J Med Sci*, 2(1):2–7, 2005.

[125] K. Weigand, A. Knaust, H. Schaller. Assembly and export determine the intracellular distribution of hepatitis b virus core protein subunits. *J Gen Virol*, 2009.

[126] M. Rost, S. Mann, C. Lambert, T. Doring, N. Thome, R. Prange. Gamma-adaptin, a novel ubiquitin-interacting adaptor, and nedd4 ubiquitin ligase control hepatitis b virus maturation. *J Biol Chem*, 281(39):29297–308, 2006.

[127] H. J. Motulsky, L. A. Ransnas. Fitting curves to data using nonlinear regression: a practical and nonmathematical review. *Faseb J*, 1(5):365–74, 1987. 0892-6638 (Print) Comparative Study Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S. Review.

[128] N. P. Christian, R. J. Arnold, J. P. Reilly. Improved calibration of time-of-flight mass spectra by simplex optimization of electrostatic ion calculations. *Anal Chem*, 72(14):3327–37, 2000. 0003-2700 (Print) Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S.

[129] A.J. Campbell, J.S. Halliday. Mass measurement from low intensity ion beams. *ASMS Conference on Mass Spectrometry and Allied Topics, #13*, pages 200–203, St. Louis, MO, 1965.

[130] A. N. Tyler, E. Clayton, B. N. Green. Exact mass measurement of polar organic molecules at low resolution using electrospray ionization and a quadrupole mass spectrometer. *Anal Chem*, 68(20):3561–9, 1996. 0003-2700 (Print) Journal Article Research Support, U.S. Gov't, P.H.S.

[131] K. F. Blom. Estimating the precision of exact mass measurements on an orthogonal time-of-flight mass spectrometer. *Anal Chem*, 73(3):715–9, 2001. 0003-2700 (Print) Journal Article.

[132] H. N. Lee, A. G. Marshall. Theoretical maximal precision for mass-to-charge ratio, amplitude, and width measurements in ion-counting mass analyzers. *Anal Chem*, 72(10):2256–60, 2000. 1520-6882 (Electronic) Journal article.

[133] V. V. Raznikov, A. F. Dodonov, E. V. Lanin. Data acquisition and processing in high-resolution mass-spectrometry using ion counting. *International Journal Of Mass Spectrometry And Ion Processes*, 25(3):295–313, 1977. ISI Document Delivery No.: EA181 Times Cited: 9 Cited Reference Count: 12 English Article 0168-1176.

[134] K. F. Blom. Utility of peak shape analyses in determining unresolved interferences in exact mass measurements at low resolution. *Journal Of The American Society For Mass Spectrometry*, 9(8):789–798, 1998. ISI Document Delivery No.: 104VC Times Cited: 11 Cited Reference Count: 15 English Article 1044-0305.

[135] L. Sleno, D. A. Volmer, A. G. Marshall. Assigning product ions from complex ms/ms spectra: the importance of mass uncertainty and resolving power. *J Am Soc Mass Spectrom*, 16(2):183–98, 2005. 1044-0305 (Print) Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S.

[136] Eva Lange, Clemens Gropl, Knut Reinert, Oliver Kohlbacher, Andreas Hildebrandt. High-accuracy peak picking of proteomics data using wavelet techniques. *Pac Symp Biocomput*, pages 243–54, 2006. Journal Article; Research Support, Non-U.S. Gov't English 1793-5091.

[137] V. V. Raznikov, A. R. Pikhtelev, A. F. Dodonov, M. O. Raznikova. A new approach to data reduction and evaluation in high-resolution time-of-flight mass

spectrometry using a time-to-digital convertor data-recording system. *Rapid Commun Mass Spectrom*, 15(8):570–8, 2001. 0951-4198 (Print) Journal Article.

[138] E. F. Strittmatter, N. Rodriguez, R. D. Smith. High mass measurement accuracy determination for proteomics using multivariate regression fitting: application to electrospray ionization time-of-flight mass spectrometry. *Anal Chem*, 75(3):460–8, 2003. 0003-2700 (Print) Journal Article Research Support, U.S. Gov't, Non-P.H.S. Research Support, U.S. Gov't, P.H.S.

[139] J. Wu, H. McAllister. Exact mass measurement on an electrospray ionization time-of-flight mass spectrometer: error distribution and selective averaging. *J Mass Spectrom*, 38(10):1043–53, 2003. 1076-5174 (Print) Journal Article.

[140] L. Charles. Influence of internal standard charge state on the accuracy of mass measurements in orthogonal acceleration time-of-flight mass spectrometers. *Rapid Commun Mass Spectrom*, 22(2):151–5, 2008. 0951-4198 (Print) Journal Article.

[141] M. Colombo, F. R. Sirtori, V. Rizzo. A fully automated method for accurate mass determination using high-performance liquid chromatography with a quadrupole/orthogonal acceleration time-of-flight mass spectrometer. *Rapid Commun Mass Spectrom*, 18(4):511–7, 2004. 0951-4198 (Print) Journal Article.

[142] A. M. Laures, J. C. Wolff, C. Eckers, P. J. Borman, M. J. Chatfield. Investigation into the factors affecting accuracy of mass measurements on a time-of-flight mass spectrometer using design of experiment. *Rapid Commun Mass Spectrom*, 21(4):529–35, 2007. 0951-4198 (Print) Journal Article.

[143] V. V. Mihaleva, O. Vorst, C. Maliepaard, H. A. Verhoeven, R. C. H. de Vos, R. D. Hall, Rchj van Ham. Accurate mass error correction in liquid chromatography time-of-flight mass spectrometry based metabolomics. *Metabolomics*, 4(2):171–182, 2008. ISI Document Delivery No.: 298IA Times Cited: 1 Cited Reference Count: 32 English Article 1573-3882.

[144] J. Cox, M. Mann. Computational principles of determining and improving mass precision and accuracy for proteome measurements in an orbitrap. *J Am Soc Mass Spectrom*, 20(8):1477–85, 2009. 1879-1123 (Electronic) Journal Article Research Support, Non-U.S. Gov't.

[145] R. Zubarev, M. Mann. On the proper use of mass accuracy in proteomics. *Mol Cell Proteomics*, 6(3):377–81, 2007. 1535-9476 (Print) Journal Article.

APPENDICES

APPENDIX A

SUPPLEMENTARY BIOPHYSICAL RESEARCH TOOLS

In many ways, biochemistry as a field has been insulated from the need to stay current with the latest technological advances. For example, very old techniques such as SDS-PAGE remain mainstays of analysis, and their continuing utility demonstrates that in many situations, the simplest solution is the best. However, technical advances in techniques and instrumentation also open up new opportunities for experiments that would not be possible with any other method, and with these advances comes the need for user progress in the associated "housekeeping" skills: generation of large cryo-EM datasets necessitates the ability to manage and interconvert the resulting 3D files, systems biology studies require the manipulation of large numbers of unique genomic or proteomic squences, and advanced kinetic modeling requires a familiarity with numeric methods.

Most of the time, existing software packages and protocols are available to meet these needs, and whenever possible, they should be fully utilized to maximize efficiency and minimize the posibility for serious errors. But for some specialized situations, no pre-packaged solution is available, or it may be more efficient to perform a trivial data manipulation with a custom routine as compared to searching for the correct tool. Here, we present an introduction to data handling for scientific analysis, and provide explanations and code for the software used in our studies of viral capsid dynamics.

Mass spectrometry data

data export

The standard Bruker DataAnalysis software is unsuited for detailed quantitation, largely due to the ability to deal with overlapping signals, flexible definitions, etc. In order to process our data with more care, all spectra needed to be exported into a

common format. The format chosen was flat-file text, due to the support within DataAnalysis to write such files, the ability to review them manually, and the simplicity and widespread support for text-based files in other software packages.

DataAnalysis does have the very useful feature of possessing a Visual Basic-based backend which can access most of the manually-available features. However, there is some lack of overlap in both directions: the user cannot manually specify multiple regions-of-interest within the software (which is possible via scripting), and the scripting engine cannot adjust parameters such as smoothing, which must be set in the software method files (which are binary blobs and cannot be parsed or written). Example code for spectral processing is shown below.

```
Set subtracted_TIC = CreateObject("DataAnalysis.TICChromatogramDefinition")
subtracted_TIC.MSFilter.Type       =  daMSFilterMS
subtracted_TIC.ScanMode            =  daScanModeFullScan
subtracted_TIC.Polarity            =  daPositive
subtracted_TIC.BackgroundType      = daBgrdTypeConstant

Analysis.Chromatograms.AddChromatogram(subtracted_TIC)
Analysis.Chromatograms(Analysis.Chromatograms.Count).AddRangeSelection 0.61, 1.0, 0,
    0
Analysis.Chromatograms(Analysis.Chromatograms.Count).AverageMassSpectrum false, true
'Analysis.Spectra(Analysis.Spectra.Count).AddRangeSelection 1000, 2500, 0, 0
Analysis.Spectra(Analysis.Spectra.Count).Smooth
Analysis.Spectra(Analysis.Spectra.Count).Deconvolute
Analysis.Save

specCount = Analysis.Spectra.count
childCount = Analysis.Spectra.Item(specCount).ChildSpectra.Count
outFileName = cstr( Analysis.Path & specCount & "." & childCount & ".txt" )
Analysis.Spectra.Item(specCount).ChildSpectra.Item(childCount).export outFileName,
    daXY
form.close

'For Each Spectrum in Analysis.Spectra
'    For i = 1 to Spectrum.ChildSpectra.Count
'        Spectrum.ChildSpectra(i).
'    next
'Next
```

In this example, a new chromatograph is defined which accounts for a pre-defined background. A selection of that chromatogram is averaged to produce a spectra, a subset of the spectra is defined, smoothed, and deconvoluted. After deconvolution, the deconvoluted spectra (which is a child of the raw spectra) is exported in a text-based file format.

Although this script uses features which are "present" in DataAnalysis, they are not functional: the AddRangeSelection method does not work. In order to work around such limitations, user-interface scripting performs such features within DataAnalysis as though the user were interacting with the keyboard and mouse: this provides reproducibility and speed:

```
#z::
WinGetTitle, DA_window, A
;MsgBox %DA_window%
WinMove, %DA_window%, , 60,60, 1100, 680
return
#x::
MouseGetPos, StartMouseLocX, StartMouseLocY
MouseClickDrag, left, %StartMouseLocX%, %StartMouseLocY%, 200, %StartMouseLocY%
return
#c::
MouseGetPos, MouseX, MouseY
MouseClick, left, %MouseX%, %MouseY%
;Send {CTRLDOWN}{+}{CTRLUP}{DOWN 2}{CTRLDOWN}{+}{CTRLUP}
MouseClick, left, 650, 626
MouseClick, right, 650,626
Send {Down}{Enter}
MouseClick, right, 250,350
Send {Down}{Enter}
MouseClickDrag, left, 620, 350, 1060, 350
MouseClick, left, %MouseX%, %MouseY%
;Send {F4}
return
#n::
MouseGetPos, MouseX, MouseY
MouseClick, left, %MouseX%, %MouseY%
;Send {CTRLDOWN}{+}{CTRLUP}{DOWN 2}{CTRLDOWN}{+}{CTRLUP}
MouseClick, left, 650, 626
MouseClick, right, 650,626
Send {Down}{Enter}
MouseClick, right, 250,350
Send {Down}{Enter}
MouseClickDrag, left, 650, 350, 1060, 350
MouseClick, left, %MouseX%, %MouseY%
;Send {F4}
return
#v::
Send {CTRLDOWN}{S}{CTRLUP}
Send {Down}
Send {F4}
return
#b::
Loop 10
{
Send, {CTRLDOWN}{+}{CTRLUP}{DOWN}{DOWN}{CTRLDOWN}{+}{CTRLUP}
{DOWN}{CTRLDOWN}{+}{CTRLUP}{UP}{UP}{UP}{UP}
}
return
```

This particular code is for the "AutoHotKey" software package in Windows, but the general concept can be carried out in whatever software is desired.

processing of kinetic data

Due to the enormous complexity and size of the kinetic data, as well as the continuously-evolving kinetic models, the software packages developed to analyze the kinetic hydrolysis data are very specific to the individual experiments, and will not be applicable to general data without substantial customization. The following scripts are provided strictly as teaching examples to indicate possible solutions for handling such data.

The following example will build a cluster from raw data:

```
'''
This is to generate a large sub-sample of the data, create
a hierarchy, then split the hierarchy into a fixed number
of clusters, and save the results.
The saved file is an npz with the following keys:
'mzFilter'
'mznpz'
'clusters'
'''

from hilmer.ms.msquant import *
from hilmer.ms.experimental import *
import scipy
import pylab

rxn32,files = load_mznpz('..\\datasets\\rxn33_both_both_DC.txt.npz')

fullMZ = rxn32[:,0,0]
fullData = rxn32[:,1,:]

filter,rxn32_hierarchy = cluster_mznpz(rxn32,countLimit=10000)
# Use the filter on the full dataset to get the appropriate array upon which to use
    the hierarchy.cuts

clusters = rxn32_hierarchy.cut(100)

saveFileName = 'rxn33_both_both_DC_10000,100.npz'

scipy.savez(saveFileName, mzFilter = filter, mznpz=rxn32, clusters=clusters)
```

Note that modules are used for efficiency. The hilmer.ms.msquant provides quantitation via selective filtering (using Python slices and logical operators) of large multidimensional matrices. The remainder of the tools used for cluster analysis are very similar in style: data loading and plotting is performed using Scipy and Matplotlib. These software libraries dramatically reduce the amount of code that needs

to be custom-written, and typically produce far greater speed than pure Python implementations.

Physical signal modulation

The physical signal modulation code for spectra alignment is based around the process for calculating splines to represent discrete spectra. By using Python and Scipy, this is relatively trivial:

```
spline = scipy.interpolate.splrep(x_vector,y_vector)
resampled_y = scipy.interpolate.splev(spline,x_vector2)
```

However, the exact implementation of the PSM code is extremely lengthy, and cannot be efficiently reproduced here. To make it available to the general public, this software is being bundled into a package called "PySpecTools", which is hosted at http://sourceforge.net. The schematic of the processing is summarized in figure A.1

Richardson-Lucy deconvolution

Our adaptation of the Richardson-Lucy algorithm follows the same pattern of using a Python interface to Fortran modules via Scipy. The original version was pure Python and completely functional, but was orders of magnitude slower, due to the optimization of the convolution operation in Scipy.

```
import pylab
import scipy
import random
from numpy import round as numpy_round
import scipy.interpolate
import hilmer.math.functions


# i = observed loc
# j = actual loc

def IRF_fxn(IRF_spline, lowLimit, highLimit):
        # Be careful about sending vectors to this function.
        def IRF_instance(i,j):
```

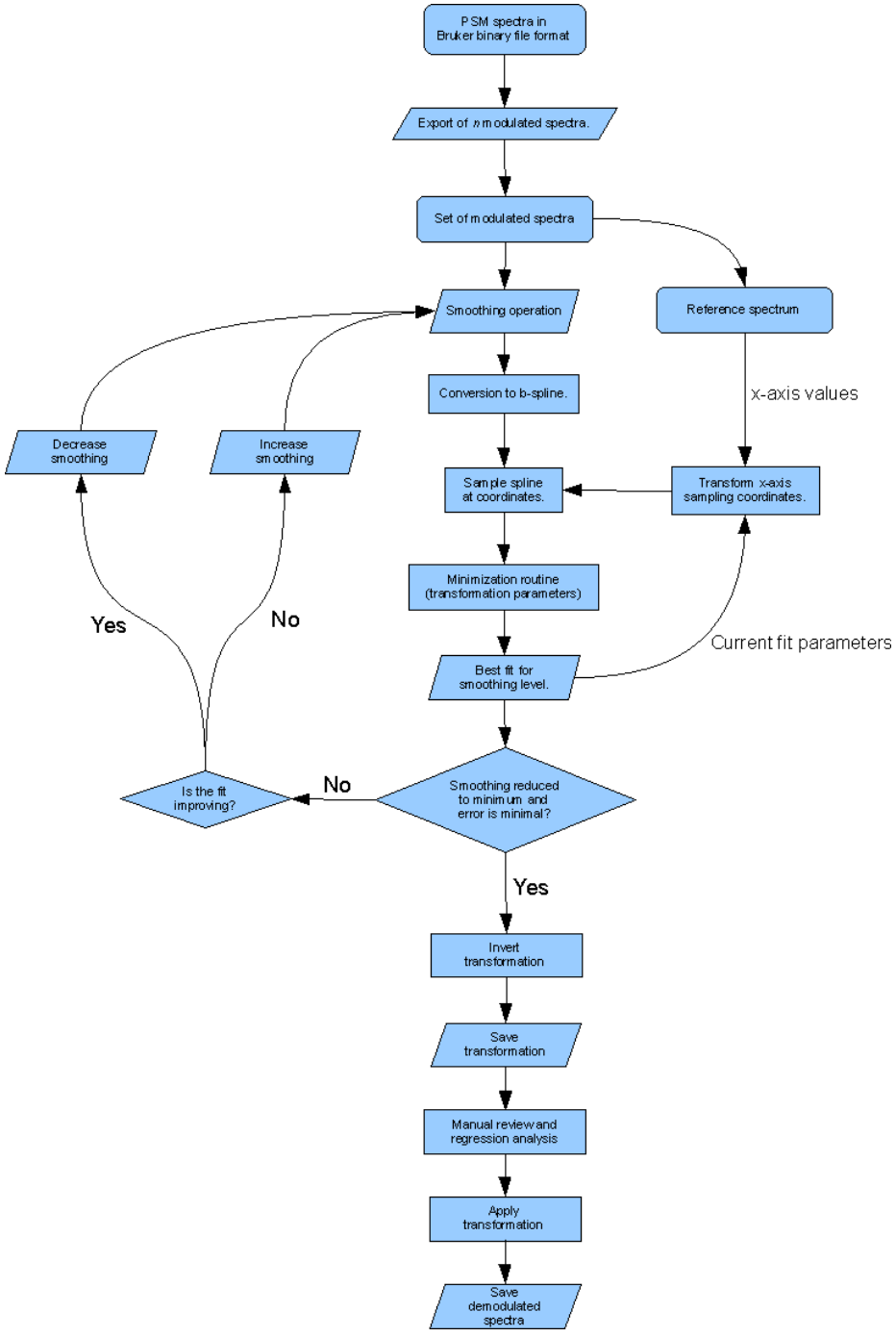


Figure A.1: Overview of the PSM software code. Spectral alignment is performed by choosing a reference spectra, turning all other spectra into splines, then using the splines to evaluate the non-reference spectra at arbitrary x-axis locations. In order to find global minima for the alignment process, the spectra are iteratively smoothed to ensure robust alignment

```
                peakX = i-j
                if (peakX>highLimit) or (peakX<lowLimit):
                        # This was returning 0 rather than 0., previously
                        return 0.
                else:
                        return scipy.interpolate.splev(peakX,IRF_spline)
                #end
        # end

        return IRF_instance
# end

def IRF_fxn_beta(IRF_spline, lowLimit, highLimit, i, j):
        if (peakX>highLimit) or (peakX<lowLimit):
                return 0.
        else:
                return scipy.interpolate.splev( (i-j), IRF_spline )
        #
#

if 0:
        p = IRF_fxn(IRF_spline, lowLimit, highLimit)

def cBold_i(U_n0, p, i): # Uj 2D array, p function, i value
        cBold_i_sum = 0
        for uj in U_n0:  # going through the DC spectra
                j = uj[0]  # m/z index
                y = uj[1]  # intensity

                pij = p(i,j)  # get the IRF value

                cBold_i_sum += y*pij
        # next uj

        return cBold_i_sum
# end cBold_i

def U_n1(U_n0, c, p, uj):  # prev Uj 2D array, Ci vector, p function, uj m/z-y pair

        Uj_sum = 0
        j = uj[0]  # m/z value from the uj pair

        for ci in c:  # going through the observed spectra
                #print '\t',
                #print ci
                i = ci[0]  # m/z index
                y = ci[1]  # intensity

                # cBold_i sums over j, so needs full u(sub-j)
                cBold = cBold_i(U_n0, p, i)

                pij = p(i,j)

                Uj_sum += (y/cBold)*pij
        # next ci

        #print uj[1]*Uj_sum

        return uj[1]*Uj_sum
# end Uj_n1

def richardson_lucy( c, p, u, str_comment):  # observed 2D array, p function, current
    DC 2D estimate
```

```
        uLen = len(u)
        jSet = range(uLen)
        #random.shuffle(jSet)

        for j in jSet:  # gets m/z-y pairs from current DC estimate
                uj = u[j]
                print 'working on ',uj, j/float(uLen), str_comment
                uj_1 = U_n1( u, c, p, uj )
                u[j,1] = uj_1  # only setting the y-value of the x,y pair
        # next j
        print 'done with routine'
        return u
# end richardson_lucy()

def richardson_lucy_discrete( c, p, u, irfOffset, str_comment):

        '''
        Assuming that c,p,u have already been regularized,
        and that p has been made symmetric.
        '''

        #print str_comment

        c_bold_i = discrete_convo_same( u, p, irfOffset )
        c_ratio = c/c_bold_i

        c_convo = discrete_convo_same( c_ratio, p, irfOffset )

        u_1 = u*c_convo

        return u_1
#

def build_kernel(array2D, center, deltaX, shift_to_zero=False, normalize=True):
        '''
        This will crop and resize an array so that
        a given point will be in the center (using
        Scipy/Numpy convention) of the array as
        used as a convolution kernel.

        Doing so prevents shifting when used in
        interative algorithms.

        It will NOT work for small kernels, since
        I'm cropping off some values to deal with
        the spline extrapolation, as well as the
        fact that spline creation assumes a window
        of 3 or so.
        '''

        '''
        This has been checked with real data, but not with
        test cases or worst-case tests.
        '''

        #pylab.plot(array2D[:,0],array2D[:,1])
        #pylab.show()

        spline = scipy.interpolate.splrep(array2D[:,0],array2D[:,1])

        lowLimit = array2D[4,0]
        highLimit = array2D[-5,0]

        #print 'lowLimit',lowLimit
```

```
        #print 'highLimit',highLimit

        high_range = highLimit-center
        low_range = center-lowLimit

        if high_range <= low_range:
                max_range = high_range
        else:
                max_range = low_range
        #

        xVector_high = scipy.arange(0.,max_range,deltaX)
        xVector_low = xVector_high[1:][::-1]*-1.

        #print 'xVector_high',xVector_high
        #print 'xVector_low',xVector_low

        xVector = scipy.concatenate((xVector_low,xVector_high))+center

        #pylab.plot(xVector)
        #pylab.show()

        yVector = scipy.interpolate.splev(xVector,spline)

        if shift_to_zero==True:
                xVector -= center
        #

        if normalize==True:
                yVector /= scipy.sum(yVector)
        #

        return scipy.vstack((xVector,yVector)).T
#

def regularize_xAxis(array2D, resample_multiplier=1.):
        '''
        This will make sure that the interval on the x-axis
        is exactly the same for the whole range.

        It uses spline interpolation for the majority of the
        data, but uses linear interpolation for the low and
        high limits to avoid the extrapolation issues.
        '''


        '''
        This has been checked with real data, but not with
        test cases or worst-case tests.
        '''

        low_spline_limit = array2D[9,0]
        high_spline_limit = array2D[-10,0]
        full_range = array2D[-1,0]-array2D[0,0]

        spline = scipy.interpolate.splrep(array2D[:,0],array2D[:,1])

        xVector = scipy.linspace(array2D[0,0],array2D[-1,0],len(array2D[:,0])*
            resample_multiplier)
        #print xVector
        #print 'xVector.shape', xVector.shape
        #print 'xVector',xVector
        #print 'spline',spline
        splineY = scipy.interpolate.splev(xVector,spline)
```

```
        linear_function = scipy.interpolate.interp1d(array2D[:,0],array2D[:,1])
        linearY = linear_function(xVector)

        combined_yVector = scipy.zeros(xVector.shape)
        for n in range(len(combined_yVector)):
                xVal = xVector[n]
                splineVal = splineY[n]
                linVal = linearY[n]

                if (xVal >= low_spline_limit) and (xVal<=high_spline_limit):
                        combined_yVector[n] = splineVal
                else:
                        combined_yVector[n] = linVal
                #
        #

        reg_array = scipy.vstack((xVector,combined_yVector)).T
        xDelta = full_range/(len(xVector)-1)

        return reg_array, xDelta
#

def deconvolute(irf, obs, irfOffset, iterations, resample_multiplier=1., str_comment=
    ''):

        resample_multiplier = float(resample_multiplier)

        irf_indices = scipy.argsort(irf[:,0])
        irf = irf.take(irf_indices,axis=0) # make sure it's sorted

        obs_indices = scipy.argsort(obs[:,0])
        obs = obs.take(obs_indices,axis=0) # make sure it's sorted

        if resample_multiplier != 1.:  # skipping this in case axes are atypical
                irf_spline = scipy.interpolate.splrep(irf[:,0],irf[:,1])
                obs_spline = scipy.interpolate.splrep(obs[:,0],obs[:,1])
                irf_xvector = scipy.linspace(irf[0,0],irf[-1,0],irf.shape[0]*
                    resample_multiplier)
                obs_xvector = scipy.linspace(obs[0,0],obs[-1,0],obs.shape[0]*
                    resample_multiplier)
                irf_yvector = scipy.interpolate.splev(irf_xvector,irf_spline)
                obs_yvector = scipy.interpolate.splev(obs_xvector,obs_spline)

                irf = scipy.vstack( (irf_xvector,irf_yvector) ).T
                obs = scipy.vstack( (obs_xvector,obs_yvector) ).T
        #

        obs_reg, xDelta = regularize_xAxis(obs, resample_multiplier=1.)
        irf_symmetric = build_kernel(irf, irfOffset, xDelta, shift_to_zero=False,
            normalize=True)

        if 0:
                pylab.plot(obs_reg[:,0],obs_reg[:,1],'r')
                #pylab.plot(irf_symmetric[:,0],irf_symmetric[:,1],'r')
                pylab.show()
        #

        DCn = obs_reg.copy()
        DCn[:,1] = 0.1

        for n in range(int(iterations)):
                print str_comment,str(n)
                # The irf has already been shifted.
```

```
                #DCn = richardson_lucy_discrete(obs_reg,irf_symmetric,DCn,irfOffset,
                    str(n))
                DCn = richardson_lucy_discrete(obs_reg,irf_symmetric,DCn,irfOffset,
                    str_comment+'_'+str(n))
        #
        return DCn
#
```

Note that in this program, the bulk of the code is related to performing data "cleaning" prior to the actual deconvolution: the Richardson-Lucy algorithm itself can be implemented in less than 10 lines of code by way of Scipy:

```
def richardson_lucy_discrete( c, p, u, irfOffset, str_comment):

        '''
        Assuming that c,p,u have already been regularized,
        and that p has been made symmetric.
        '''

        #print str_comment

        c_bold_i = discrete_convo_same( u, p, irfOffset )
        c_ratio = c/c_bold_i

        c_convo = discrete_convo_same( c_ratio, p, irfOffset )

        u_1 = u*c_convo

        return u_1
#
```

This particular routine will perform only a single iteration: between 10 and 1000 iterations is usually sufficient to achieve proper convergence.

Fitting kinetics data

Of all the software needs discussed thus far, data fitting is the topic which finds the best support via commercial software channels. There are many excellent tools

available for data plotting and curve fitting, so it is difficult to find it necessary to write custom software for the purpose. However, differential equations without explicit forms tend to be more poorly supported than other systems of equations, and the high-order complexity of virus capsids introduces complications. For example, capsids may become destabilized after a certain degree of cleavage, which necessitates that kinetic parameters for all the functions be altered mid-simulation. The complexities of proteins with cross-links also make any kinetic models extremely intricate: for the Cp149 protein, a single disulfide and two cleavages produce 24 possible trends which must be computed, and the cleavage locations make this a best-case scenario. In order to simulate arbitrary protein sequences with larger numbers of cleavages and more potential cross-links, it is essential to be able to generate kinetic equations automatically via software. Although we have not yet completed that phase of the software, our model is capable of accepting such inputs.

The following code is from the kinetic fitting software. For clarity many functions have been removed.

```
def generic_cleavage_trend(x,interior_kList,exterior_kList,C0List=scipy.array((0)),
    pathCount=1.):
        '''
        Refer to PMID:10935975 for details.

        Hilmer modifications consist of allowing pre-existing concentrations/
        probabilities of a terminus already existing: this would be the case
        for N- or C-terminal peptides or any other pre-existing cleavage.

        For this modification to work the order of the CO factors needs to
        match the order to the exterior_kList.  This is due to the grouping
        below.  See comments for specific location.  Grouping termini would
        probably be more logical than grouping the kCleave factors and
        C0 factors as it currently is set up.

        This should work on both single values of x as well as lists/arrays.
        '''
        C0List = scipy.array(C0List)
        if C0List.shape[0] == 0:
                C0List = scipy.zeros((len(exterior_kList)))

        # This isn't really the right way to handle this.
        if len(C0List) != len(exterior_kList):
                return 'error! The C0List must match the exterior_kList!!'
        else:
                #print C0List.shape
```

```
                kTerm1 = exterior_kList[0]
                kTerm2 = exterior_kList[1]
                C01 = C0List[0]
                C02 = C0List[1]

                # This is the source of the problem necessitating the proper order of
                    the C0 factors and the kTerms.
                pProduce1 = (1.-scipy.exp(-kTerm1*x)+C01)
                pProduce2 = (1.-scipy.exp(-kTerm2*x)+C02)


                pSurvive = scipy.exp(-scipy.sum(interior_kList)*x)
                #print 'Pre-existing quantity =',C0
                #print 'probability of surviving internal cleavage =',pSurvive
                #print 'Probability of production via terminal cleavage1 =',pProduce1
                #print 'Probability of production via terminal cleavage2 =',pProduce2

                #print 'Prob Sum = ', (pProduce+C0)

                #Ct = pSurvive*(pProducePairwise)
                Ct = pSurvive*(1.-(1.-(pProduce1*pProduce2))**pathCount)

                #print Ct
                return Ct
        # end if
# end def generic_cleavage_trend(interior_kList,exterior_kList):

def trend_05(x,k1,k2,k3,k4,kX1,kX2,kX3,kX4,C0List=[1.,1.]):
        return generic_cleavage_trend(x,[k1,kX1,k2,kX2,kX3,k3,kX4,k4],[0.,0.],C0List)
# def trend_05
```

The actual fitting routine is not shown because it is highly explicit and customized for the Cp149 system, as well as being very lengthy. As with the MS data clustering routines, it is based around combining matrices, vectors, etc via Scipy. The minimization routine itself is very simple:

```
import scipy
import scipy


def example_function(x):
        return x**2


def error_function(x,a):
        return example_function(x)-a


estimate = 4.
target = 5.
solution = scipy.optimize.leastsq(error_function,estimate,args=(target))
```

The leastsq routine can be replaced with one of many minimization algorithms available in the optimize module. Since nearly all of these routines are wrappers around Fortran routines, some work better than others with highly abstracted object oriented programming. The leastsq routine takes three inputs: the first is the function to be minimized: leastsq assumes that the function will generate a vector, and it attempts to minimize the sum of squares for that vector. The second argument is the initial estimate, which does not have to be a single value: it easily tolerates vectors and other list-type objects, but higher-order objects (matrices etc) can cause unexpected results. The final argument allows objects to be passed to the error function. As with the initial estimate, these can be vectors but higher-order objects are not recommend. However, Python will handle global objects and other means of passing data between functions, so this is not a limitation in practice.