In Silico Tools for the Development of Biotherapeutics

by:

Timothy Michael Lauer

B.S. Chemical Engineering
University of Massachusetts Amherst, 2009

SUBMITTED TO THE DEPARTMENT OF CHEMICAL ENGINEERING
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN CHEMICAL ENGINEERING
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2016

Signature of Author: Signature redacted

Timothy M. Lauer
Department of Chemical Engineering
October 9, 2015

Certified by: Signature redacted

Bernhardt L. Trout
Raymond F. Baddour, ScD (1949) Professor of Chemical Engineering
Thesis Supervisor

Accepted by: Signature redacted

Richard D. Braatz
Edwin R. Gilliland Professor of Chemical Engineering
Chairman, Committee for Graduate Students

In Silico Tools for the Development of Biotherapeutics

By:

Timothy Michael Lauer

Submitted to the Department of Chemical Engineering on October 9, 2015
In Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

ABSTRACT

The production of a new drug is an inherently risky process. While there are many causes for failure, a number of these are due to the many potential degradation pathways a protein can undergo. Often these reactions are slow and negligible. However, there are times where any one of these reactions can be significant enough to delay or prevent a drug's development. Testing for all of these degradation routes can be difficult in the early stages of drug development due to small amount of available protein and the long times needed to sample these reactions. In order to reduce the risk to the drug development process, *in silico* methods have been developed to predict the likelihood of these reactions, without the need for any material. This work focuses on two degradation routes; the aggregation pathway and the acid-catalyzed, non-enzymatic hydrolysis of peptide bonds. Aggregation of antibodies can be a limiting factor for liquid formulations; however, two major factors control this reaction: the surface hydrophobicity and the protein charge. These two factors were combined into a new tool, called the Developability Index, to predict protein aggregation rates. This tool was successfully applied to both antibody and individual antibody domains.

The non-enzymatic, acid-catalyzed hydrolysis of an amide bond following an aspartic or glutamic acid residue is controlled by a wider range of factors. These include: the secondary structure, the surface exposure of the amide bond, relative orientation of the sidechain, and the availability of the sidechain. These four factors impact the first two steps of the hydrolysis mechanism: the addition of the proton to the peptide bond and the addition of the sidechain to the peptide backbone. The secondary structure and surface exposure of the peptide bond impact the ability of the proton to add to the peptide bond, and thus start the reaction, while the orientation and availability of the sidechain impact the ability of the sidechain to cyclize and form a ring. These factors can be combined to produce a method to predict the reactivity of peptide bonds with high accuracy.

Thesis Supervisor: Bernhardt L. Trout
Title: Raymond F. Baddour, ScD, (1949) Professor of Chemical Engineering

Table of Contents

1. INTRODUCTION

While there is a great deal of controversy about the exact cost of making a new drug, most agree on one fact: it is expensive. One estimate of the cost per new chemical entity is $802 million (*1*), another estimate places the cost at $1.3 billion for biopharmaceuticals (*2*), while yet another estimate places it between $0.5 and $2 billion depending on the company developing the drug and type of therapy being produced (*3*). Not only is this process costly, but it is a long process that often fails. One estimate places the time needed to produce a new drug at more than a decade, with only 5% of attempts succeeding (*4*). As the long time and high failure rate contribute to the high cost of new drugs, several strategies have been proposed to decrease these, one of the frequently considered methods for reducing the cost of the process is through the use of *in silico* prediction methods (*4*).

In silico prediction methods/tools encompass a wide range of computational tools that can be used to predict protein properties based on some computational models. These predicted properties can be used in conjunction with experimental results to increase the knowledge about the behavior of the protein before substantial money is invested in its production. As a group, these tools often have a number of useful properties including: no material needed to make prediction, high throughput, rapid results, and these tools can even suggest sources of instability in addition to predicting the likelihood of instability. There are several issues with these tools; the two most significant are that they require expertise that companies may be lacking and accurate *in silico* tools are needed. These tools require either large experimental data sets to validate or a strong foundation in the physics of the reaction.

There are many potential degradation routes for proteins. In general, these stabilities are classified into two main categories: chemical instabilities and physical instabilities. Chemical instabilities are those that involve creating or destroying chemical bonds; these include such pathways as oxidation, hydrolysis, proteolysis, and deamidation (*5*,*6*). Physical instabilities are those pathways that do not alter the chemical bonds of the proteins, such as aggregation, denaturation, surface adsorption, and precipitation (*5*,*6*). While all proteins are susceptible to all of the degradation routes, they may not be significant in all cases.

This study will present *in silico* tools to predict two of these pathways, the aggregation pathway (a physical instability pathway) and the hydrolysis pathway (a chemical instability).

2. DEVELOPABILITY INDEX: A RAPID *IN SILICO* TOOL FOR THE SCREENING OF ANTIBODY AGGREGATION PROPENSITY

2.1. Introduction

Monoclonal antibodies (mAbs) represent a growing portion of pharmaceuticals. Currently, there are over twenty antibody-based drugs on the market (*7*), with many more antibody-based drugs at various stages of development. MAbs have several advantages over other types of drug substances, including high specificity for their target and long serum half-life (*8*). However, there are a number of factors that impede their use, such as restriction to antigens that are either in serum or on the cell surface and the high dose that is frequently required for treatment. The high mAb concentrations required to treat patients leads to challenges in their long-term storage, in particular the risk of aggregation. For mAbs, and therapeutic proteins in general, aggregation is an important degradation pathway. It can compromise product integrity, and the aggregates may elicit an immunological response (*9*).

One of the challenges in the development of an antibody-based drug is selecting a suitable lead candidate among many possible mAbs that will bind to the target antigen. Among others, aggregation propensity is a property that is often quantified during the development phase to select mAb with lower aggregation propensity. Because there are many possible candidates to be used for a potential drug, each of these mAbs is tested for aggregation propensity to determine which would be the best candidate for use as a drug. The current practice in the industry is to infer the aggregation propensity of the mAb via a number of biophysical experiments; however, these tests need to be done quickly with small amounts of protein, so it is difficult to truly capture the long-term stability properties of a mAb, particularly if the mAb is to be stored at a high concentration. In this work, we present a novel, high-throughput, *in silico* tool to predict relative stability, with respect to aggregation, of a mAb based on its structure.

Several tools already exist to predict protein aggregation (*10*), (*11*), (*12*), (*13*). The majority of these tools are based on the protein's primary sequence and have been developed based on the knowledge gained from amyloid formation of short peptides. These sequence-based tools generate predictions using a number of amino acid characteristics, such as their propensity to form different secondary structures, charge, polarity, and solubility. A few of these methods have been applied to mAbs (*14*); however, minimal information is available in the open literature on the validation of these tools. Moreover, none of these tools have been applied to screen and rank mAbs according to their aggregation propensity.

There are many properties that affect protein aggregation, including hydrophobicity, charge, propensity to form β-sheets, and the propensity to form α-helical structures. Hydrophobicity is thought to be of major importance during protein aggregation (*15*). Recently, a tool called Spatial-Aggregation-Propensity (SAP) was developed in our group to identify the aggregation prone regions of proteins (*13*), (*16*), (*17*), (*18*), (*19*). SAP is a measure of the local hydrophobicity of surface patches (either static or dynamically exposed), which identifies regions responsible for hydrophobic interactions. Electrostatic interactions are also quite important as electrostatics can have an important role in solution phase reactions. To incorporate the competing effects of electrostatic interactions and hydrophobic interactions, we developed a new parameter, termed the Developability Index (DI). The DI provides a rapid *in silico* predictive tool to rank mAbs according to their aggregation propensities (see Figure 2-1 for more details).

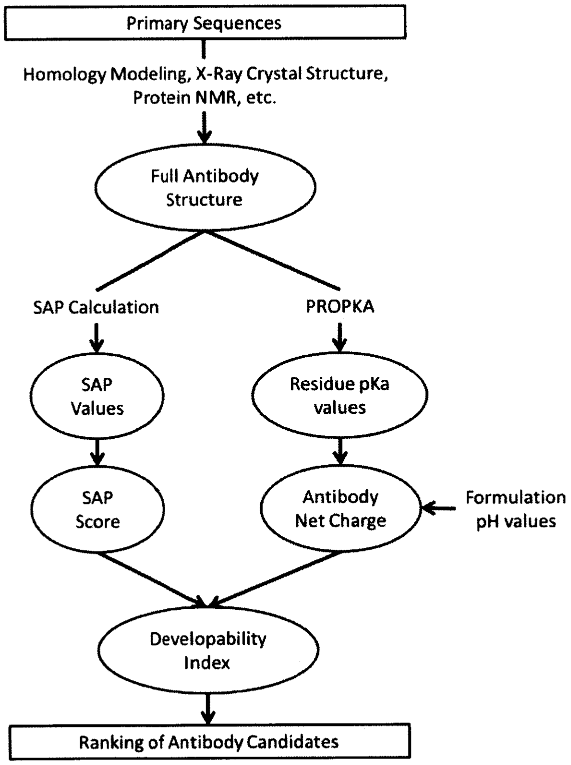

Figure 2-1 - The procedure for the calculation of the Developability Index (DI) from a primary sequence. The primary sequence of the mAb, homology modeling (See Section 2.2.2) or other methods can be used to determine the structure of the mAb. This structure can then be used to calculate the net charge (See Section 2.2.5) through the use of PROPKA. Concurrently, the SAP Score of the mAb (See Section 2.2.3) can also be calculated. Using the value of the SAP Score and the net charge, the DI can then be computed and used to rank mAbs (See Section 2.3.3).

$$DI = [Antibody\ SAP\ Score] + \beta \times [Antibody\ Net\ Charge]^2$$

Equation 2-1

The DI, mathematically defined in Equation 2-1 is the combination of a mAb's SAP score and the square of the mAb's net charge. There are many functions that can combine a mAb's SAP score and net charge into a single value, and any of these functions are possible. A high value of DI is indicative of higher aggregation propensity and vice-versa.

In Equation 2-1, the SAP Score accounts for hydrophobic interactions during the aggregation process, while the mAb net charge accounts for the electrostatic interactions during aggregation. The weighting factor, β, can be determined through data regression of long-term stability data with SAP Score and net charge. This factor is necessary due to the numerical difference in the scales of the SAP Score and the square of the net charge as well as the possibility that either of the two variables may have a dominant role in the determination of the aggregation rate. Additionally, the weighting factor, β, may be system dependent and could vary with formulation conditions, such as concentration of salts or other additives.

One challenge of using SAP and the structure-based net charge to measure protein-protein interactions is that both require a known structure of the protein. However, there are a number of freely available tools that can be used to create reasonably accurate models of Fv regions of mAbs from their primary sequence, including WAM (***20***), PIGS (***21***), and Rosetta (***22***). While these modeling algorithms may take several hours, user input is not required once the modeling has begun, and multiple mAbs can be run in parallel. As there are a number of tools and procedures to generate a reasonably accurate structure from the sequence for mAbs, the use of structure-based variables to calculate DI is not considered limiting to its use.

2.2. Methods

2.2.1. Long-Term Stability Data

To determine the experimental aggregation propensity, defined here as the rate constant for the aggregation reaction, long-term stability studies were performed on nine IgG1 mAbs, one IgG2 mAb, and two IgG4 mAbs. These studies consisted of storage of mAbs for long times, up to two years, under controlled conditions. The mAb samples were stored in controlled environments (either 25 °C at 60% relative humidity or 40 °C at 75% relative humidity) until a sample was removed for size-exclusion chromatography.

There was some variability in the data due to slight differences in procedures over the time that the different stability studies were started and the individual performing the study. These differences included slight variability in the formulations, variability in the time of measurement (i.e. additional measurements at different times), and variability in the initial concentrations of mAb and aggregates. All mAbs were stored in L-histidine solutions at a concentration of either 10 mM or 20 mM. The initial mAb concentration also varied across samples and ranged from 40 mg/mL to 150 mg/mL; similarly, the initial percent of aggregates varied from 0.3% to 3%. Finally, the pH of the formulation varied between experiments, ranging from 5.5 to 6.5. Further details on the experimental conditions are available in the supplemental section 10.6.

To determine a measure for the stability, i.e., formation of aggregates, a kinetic model was proposed. For simplicity and relative accuracy, a second order model (Equation 2-2) is assumed.

$$\frac{dC_M}{dt} = -k_{Agg}{C_M}^2$$

Equation 2-2

Here, C_M is the concentration of non-aggregated mAbs, k_{Agg} is the aggregation rate constant (aggregation propensity), and t is time.

This model does not capture all of the aspects of the complex aggregation process. While more complicated models exist to describe protein aggregation (***23***), this model captures the data suitably well given the accuracy of the long-term measurements. The initial mAb concentration is included in the model because many of the mAbs are studied at multiple concentrations and the aggregation rate appears to be dependent on the initial concentration of mAb. The amount of aggregated product and drug product are measured using size-exclusion chromatography. The values for the kinetic rate constants are determined using ordinary least squares linear regression, using all available data for each mAb. Using the experimental data at 25 °C and 40 °C, a rate constant is calculated for each temperature.

Due to the uncertainty in the experimental values for the aggregation rate constants, we restricted ourselves to predict the stability class of each mAb instead of the actual aggregation rate. The mAbs were clustered into three groups, low stability (high aggregation propensity), medium stability, and high stability (low aggregation propensity), based on the value of their aggregation rate constant. The cluster centroids were determined by minimizing the sum of the square of the distance between the data and the nearest centroid using MATLAB's® *kmeans* algorithm. This procedure was performed independently for the two temperature data sets so that the classifications at either temperature were not influenced by the other.

2.2.2. Homology Modeling

As these mAbs are in various stages of development, the structures of these mAbs are not known experimentally. In the past, it had been found that the use of the WAM modeling tool (***20***), a mixture of homology and canonical structure modeling, generated a structure for the aglycosylated mAb to an acceptable accuracy (***17***). This information was used in conjunction with the known crystal structure (PDB ID: 1HZH (***24***)) for an IgG1 mAb with a kappa light chain, which serves as the basis for the full body structure of all IgG1 mAbs. The 1HZH served as the basis for the constant regions of all the mAbs used in this work, and it was also used as the basis for the constant regions of the κ side chains for mAbs mAb1, mAb2, mAb3, mAb5, and mAb6. To model the constant regions of light chains for those mAbs with λ light chains, mAb4 and mAb7, the model 1ZVO from the RCSB database was used (***25***), which also has a λ light chain. The structures generated were used for both the SAP calculations and the mAb net charge calculations. A list of the properties of the mAbs used in this study is given in Table 2-1 in Section 2.3.3, including SAP Scores, net charges, aggregation propensities, the IgG subclass, and the type of light chain.

In addition to the seven IgG1 mAbs, five other mAbs were also tested to validate the DI. While the constant light regions are similar between all mAbs, different models are used to model the constant regions of the heavy chains due to the different structures of the IgG subtypes. The two IgG1 mAbs, mAb8, and mAb9, were modeled using the same method as the other IgG1 mAbs, mAb1 to mAb7. The two IgG4 mAbs, mAb10 and mAb11, were modeled using the structure available for PDB ID: 1ADQ (***26***). The constant region of the heavy chain of the last mAb, mAb12, was modeled using PDB ID: 1FC1 (***27***).

2.2.3. SAP Score and SAP definition

The SAP tool was developed to identify regions of hydrophobic patches on the surface of a protein. In the past papers, the SAP tool has been applied to a number of proteins, including mAbs (***16***,***13***,***17***,***19***,***18***). These papers showed that SAP could be used to predict aggregation-prone regions of a protein. SAP is defined as:

$$SAP_{atom\ i} = \sum_{\substack{Simulation \\ Average}} \left\{ \sum_{\substack{Residues\ with\ at \\ least\ one\ atom \\ within\ R\ of\ atom\ i}} \left[\frac{\begin{array}{c} SAA\ of\ side\ chain\ atoms \\ within\ radius\ R \end{array}}{\begin{array}{c} SAA\ of\ side\ chain\ atoms \\ of\ fully\ exposed\ residue \end{array}} \times \begin{array}{c} Residue \\ Hydrophobicity \end{array} \right] \right\}$$

Equation 2-3

where

1) *SAA* is the 'solvent accessible area' of side chain atoms contained within radius R from the central atom.
2) *SAA of side chain of fully exposed residue* (for amino acid 'X') is obtained by calculating the SAA of the side chain of the middle residue in the fully extended conformation of tripeptide 'Ala-X-Ala'.
3) *Residue Hydrophobicity* is obtained from the hydrophobicity scale of Black and Mould (***28***). The scale is normalized such that glycine has a hydrophobicity of zero, the most hydrophobic residue (i.e., PHE) has a value of 0.5, and the least hydrophobic residue (i.e., ARG) has a value of -0.5. Hydrophobic residues have residue hydrophobicity values greater than 0. Residue hydrophobicity values less than zero are less hydrophobic than Glycine.
4) The simulation average is not a requirement. Static SAP also gives good results (see Figure 2-2 and Section 2.3.1 for more details).

One difficulty in using SAP for ranking mAb aggregation is that the original SAP method generates a value for each individual atom, not a single value for the protein. To convert the SAP values for the atoms of a mAb into a single value for the mAb, the SAP Score was developed. The SAP Score of a mAb is determined by the sum of all positive SAP values of the atoms in the CDR of the mAb. The definition of the CDR that is used throughout this work is the definition used by the WAM homology algorithm (***20***).

$$SAP\ Score = \sum_{\substack{All\ atoms\ in\ CDR \\ with\ SAP\ Value>0}} (SAP_{atom\ i})$$

Equation 2-4

The summation only includes the CDR because the CDR are the regions where the greatest variation in IgG mAb is observed. The benefit of only considering the atoms in the CDR is that SAP only needs to be calculated for a relatively small section of the mAb. The inclusion of only those atoms in the CDR is also limiting in that it does not include the effect of mutations in the framework or in the constant regions. However only including the CDR atoms would make comparisons across subtypes difficult, such as comparing an IgG1 mAb to an IgG2 mAb. Other differences, such as those in the framework of the Fv region, are not accounted for directly by SAP. These differences may affect the SAP values in the CDR if the affected residue is sufficiently close to the CDR. However, including regions beyond the CDR do not appear to affect the ranking of mAbs by SAP, for more details see supplement Section 10.5.

There were two changes to the SAP procedure compared to the previous work[7]; one is the selection of a slightly different radius, and the other is the use of only the static structure in the determination of SAP. A SAP radius of 5 Å was used because it is approximately averaging over the nearest neighbors. Other radii were attempted, such as 10 Å and 15 Å as in past work, but 5 Å was among the best performing radii (see Figure 10-2). Dynamic SAP calculations are discussed further in the next section (Section 2.2.4).

2.2.4. Dynamic SAP Calculations

In previous papers (***13***), (***16***), (***17***), (***18***), (***19***), SAP was determined over the course of a molecular dynamics (MD) simulation; however, all atom MD simulations are computationally costly. To increase the utility of the DI, SAP calculations for the DI are based solely on the static structure. By using only, the static structure, the DI can therefore be used in the early development process to rapidly screen a large number of potential mAbs. To determine if the SAP Score derived from the static structure is sufficient to determine the ranking, four Fab segment MD simulations were performed. These simulations were run for a total of 10 ns, 2 ns for equilibration and 8 ns for property calculations, using the Fab segments of each mAb in an explicit solvent.

The simulations were setup and analyzed using the CHARMM simulation package (***29***). These simulations were performed using NAMD (***30***) using the CHARMM22 fully atomistic force field (***31***) with the TIP3P solvent model for water (***32***). The charges of the histidine residues were determined based on the distance to nearby negatively charged residues. The simulation was run as an NPT ensemble, with the temperature fixed at 298 K and the pressure fixed at 1 atm. The mAb Fab was solvated in an orthorhombic box, with periodic boundary conditions in 3 directions and an 8 Å water solvation shell around the mAb fragment. Ions were added as needed to neutralize the net charge of the system, as required by the Ewald summation technique for the calculation of the electrostatic contribution. After the mAb was solvated, the energy was initially minimized with steepest descents (SD) by fixing the protein to allow the water to relax around the protein. Then, the restraints were removed, and the structure was further minimized with SD and the adopted basis Newton-Raphson. The system was then slowly heated to room temperature with 5 °C increments every 0.5 ps using a 1 fs time step. The system was then equilibrated for 2 ns before starting to compute the various properties from simulation. The configurations were saved every 0.1 ns during the simulation for further analysis.

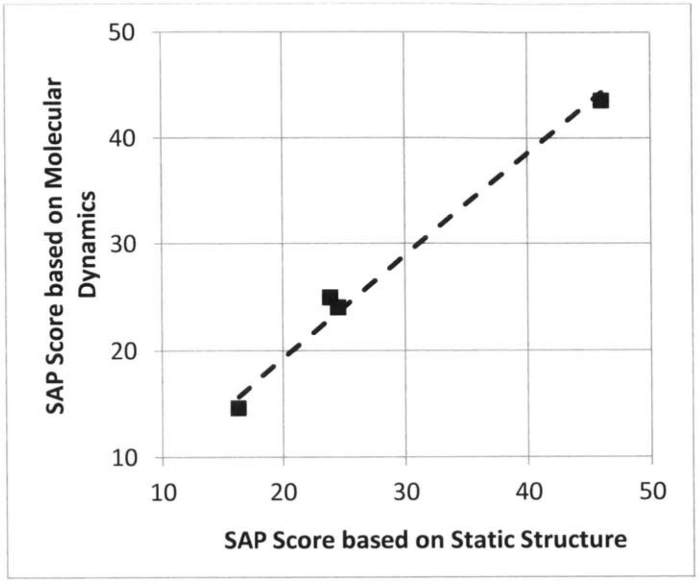

Figure 2-2 - SAP Score based on molecular dynamics simulation vs. the SAP Score based on the initial static structure for mAbs (in order of increasing SAP Score based on static structure) mAb1, mAb2, mAb10, and mAb3. The dashed line is the linear regression between the SAP Score based on the initial static structure, from homology modeling, and the 10 ns molecular dynamics simulation (R^2 of 98.8%).

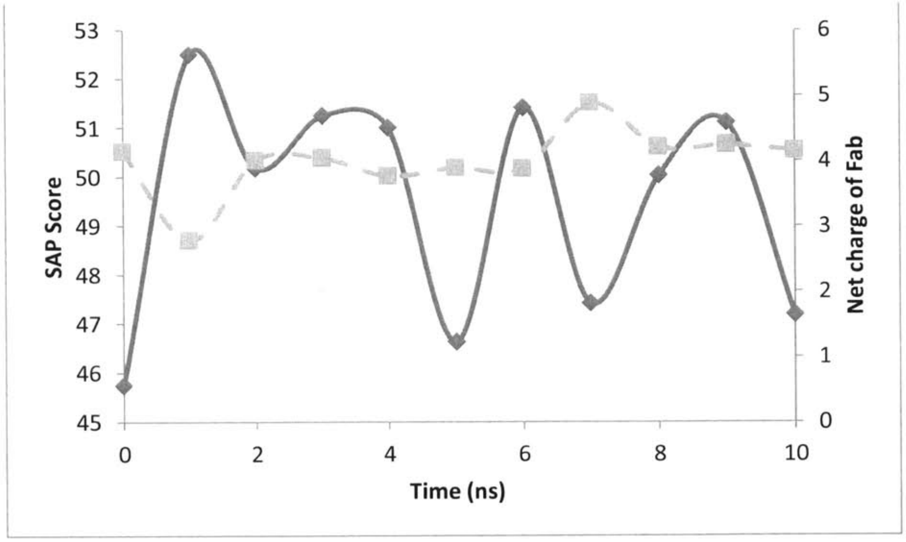

Figure 2-3 - Variations in SAP Score [solid black line; average of 49.6] and Net Charge [dashed gray line; average of 4.1] over the 10 ns molecular dynamics simulation for mAb3.

2.2.5. Antibody Net Charge

To take electrostatic interactions into account, the mAb net charge was used. The pKa of individual residues was determined using PROPKA 1.0 (*33*), (*34*) with the full mAb structure. PROPKA uses the local environment of a residue, including electrostatic interactions of nearby charged residues, hydrogen bonding, and desolvation effects, to determine its pKa value. These pKa values were then used to determine the charge of each amino acid. Because this pH range contains the pKa of free histidine, the pKa of each residue that could be charged was used to compute the partial charge using equations derived from the Henderson-Hasselbalch equation. A similar process has been used before (*35*), where all arginine, histidine, lysine, aspartic acid, glutamic acid, and tyrosine residues were considered chargeable and included in the net charge. In addition to these residues, unbound cysteine residues were also included in the net charge. The glycosyl group was not considered in the net charge as there are a number of different possible glycosylation patterns, and the distribution of these patterns is dependent on many variables outside the structure of the mAb.

In past works, it has been found that the calculated net charge and the measured net charge could vary significantly. However, it has also been found that the net charge from calculations based on the amino acid sequence results in the same ranking of mAbs as was determined from experiments (*36*).

We used the partial charge of each residue instead of the discrete charge in which a residue is either charged or not charged. The first reason for this approach was that there are several histidine residues in the mAb, particularly in the heavy chain used for the main data set. If a discrete charge is used, there is a large jump in charge in the pH range of interest because it crosses the pKa of histidine. However, this change is less severe if the partial charge is used. Another reason for the use of partial charge is that it allows for greater differentiation between similar residues and similar mAbs. For instance, at a pH of 6, both histidine and arginine are positively charged, but as the pH nears the pKa of histidine, only a fraction of those in solution will be charged, while virtually all the arginine would still be charged. This differentiation makes comparison between similar mAbs possible.

Another effect that the variable protonation state of histidine has is that it requires the inclusion of the entire mAb structure in the net charge calculation. In the case of the SAP Score, the SAP of the constant regions is the same for all mAbs. In the case of the net charge, the variable protonation state of histidine can substantially change the net charge of the mAb because of the histidine residues in the constant region of the heavy chain and the range of pH values tested.

2.2.6. Determination of β

To determine the relative importance of mAb net charge and the SAP Score to the aggregation propensity, possible values of β were iterated through, from hydrophobic controlled (a β of zero) to charge controlled (high β). At each attempted value for β, the DI value for each mAb was computed using Equation 2-1, and the centroids were scaled to the centroids of the fitted aggregation data. The magnitude of the scaling was determined by minimizing the root mean square distance between the DI values and the nearest centroid. The centroids were scaled by Equation 2-5:

$$\mu_{DI} = \mu_k \times c$$

Equation 2-5

where μ_K and μ_{DI} are the centroids for the aggregation rate constants and the centroids for the DI, respectively. The factor c is a scaling factor. The function was minimized using the MATLAB function *fmincon*. The mAbs were then put into clusters based on their DI values and the scaled cluster centroids. After the new clustering was determined, it was compared to the actual clustering, and the number of correctly classified mAbs was computed. The optimal range of β was the range that resulted in the largest number of correct classifications.

This fitting was done with a few assumptions mainly that the formulation conditions and the constant regions are identical or similar enough among all the mAbs to be directly comparable. To address these concerns, the mAbs used to determine the value for β were selected carefully. To maintain similar formulation conditions, only stability studies with the same additives were used, although the concentration of these species did vary slightly. Similarly, only long-term stability studies containing data for both 40 °C and 25 °C were considered; at lower temperatures, the aggregation rate is often not substantial enough to be measurable accurately in a reasonable period of time. Of the mAbs with stability reports that met these requirements, only those that were IgG1 with identical constant regions for the heavy chain were considered. The constant region of the light chain followed one of two possibilities: either κ or λ. There is only one isotype of either κ or λ.

2.3. Results

2.3.1. SAP: Static Structure vs. Dynamic Simulations

After the 10 ns simulations were performed, the SAP values of each atom were computed over the final 8 ns simulation to allow the fragment to equilibrate. While the magnitude of these numbers differed slightly, the values based on the static structure were close to the value from the simulation. The SAP Score based on the static structure was, on average, 4% greater than the simulation-based SAP Score. As can be seen in Figure 2-2, there is a correlation between SAP derived from the static and SAP derived from the dynamic simulations, so it was acceptable to use only the initial static structure generated from WAM in the computation of the SAP score for the mAb.

To see how the DI would vary over the course of a simulation, a structure was generated every 1 ns of the simulation and used to calculate the net charge and the SAP Score. The net charge varied only slightly over the course of the simulation and only varied significantly at two points (Figure 2-3). At these points, a histidine residue temporarily gained or lost a hydrogen bond to another nearby residue, which greatly impacted the stability of the protonated state. This effect was noticeable at a pH of 7, which is close to the model value for the pKa of histidine residues. There were larger variations of the SAP Score over the course of the simulation, up to ±6% of the mean value. While both the SAP Score and the net charge varied over the course of the simulation, the change in the DI was approximately the change in the SAP Score, as the change in the net charge was small compared to the change in the SAP Score.

2.3.2. Sensitivity Analysis for SAP Score

To determine the DI of any given protein a number of variables must be set. These variables are typically involved in the computation of the SAP Score. In particular, the cutoff value for the atoms in the SAP Score, the summation of SAP values for atoms in the CDR, and the SAP radius can be selected arbitrarily. To test the sensitivity of the resulting classifications to these variables, they were varied and the changes in classifications were noted. Furthermore, during the determination of the numeric value of the SAP Score used in the computation of the DI, a minimum threshold value to include was defined. Originally, this was set at zero.

$$SAP\ Score\ of\ Antibody = \sum_{\substack{All\ atoms\ in\ CDR \\ SAP\ Value > cutoff}} (SAP\ Value_{atom\ i})$$

Equation 2-6

Changes of less than ± 0.1 in the cutoff value had no effect on the final mAb classifications (see Figure 10-2). This range of ± 0.1 contained the majority of SAP values; approximately 60% of the observed SAP values fell within this range (see Figure 10-3).

The radius used to compute the SAP Score did have a substantial effect on the classifications (see Figure 10-1). A SAP radius of 5 Å was used throughout the rest of this work because it resulted in the largest number of correct classifications and a SAP radius of 5 Å roughly corresponded to the average of hydrophobicity over a residue's nearest neighbors. However, a range of SAP radii from 4 to 7 Å performed as well as 5 Å.

It is possible to compute the SAP Score of any region or portion of a protein. In this case, as the remaining parts of the protein are constant or close to constant across the mAbs, only those atoms in the CDR of the mAb were included because the CDR contains the majority of the differences between the different mAbs. Overall, the SAP Score parameter was robust, and a minor change in its parameters did not lead to major changes in the classifications of the mAbs.

2.3.3. IgG1 Data Regression

A set of seven IgG1 mAbs were used to determine the optimal value of the weighting factor, β, in Equation 2-1. The fitting was performed independently for both the 40 °C and the 25 °C point. The resulting DI classifications and the results from experiments are shown in Figure 2-4. The values for the fitted parameters and the cluster values for the aggregation rates are in Table 2-2. For a physical meaning of these values, see Figure 2-5. In general, the classifications were quite good. However, it was unclear whether deviations were due to the DI or inaccuracies in the experimental data.

DI Classifications of Long-Term Stability				
mAb	25 °C		40 °C	
	Expt.	DI	Expt.	DI
mAb1	🟩	🟩	🟩	🟩
mAb2	🟩	🟩	🟩	🟩
mAb3	🟨	🟨	🟥	🟥
mAb4	🟩	🟩	🟩	🟩
mAb5	🟩	🟩	🟩	🟨
mAb6	🟥	🟩	🟨	🟨
mAb7	🟩	🟩	🟨	🟨

🟥🟨🟩

Increasing Stability →

Figure 2-4 - Experimental and DI-based classifications for long-term stability against aggregation for the 7 mAbs used in the regression (mAb1, mAb2, mAb3, mAb4, mAb5, mAb6, and mAb7). Green represents high stability, yellow represents medium stability, and red represents low stability.

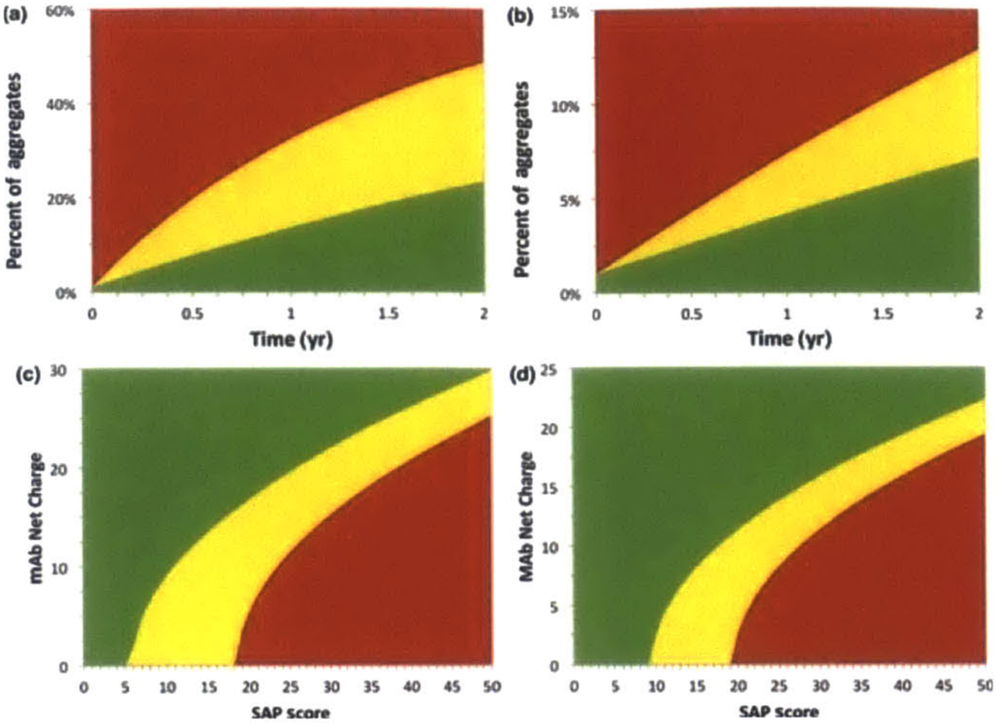

Figure 2-5 - Graphical representations of the stability classifications. A) and B) show the percent of aggregates in solution as a function of time assuming an initial mAb concentration of 100 mg/mL with 1% of the total protein in the aggregated state. The different classifications were determined by the aggregation propensity values. A) shows the classifications of a mAb solution if stored at 40 °C and B) if the solution had been stored at 25 °C. C) and D) show the SAP Score and Net Charge regions that map to the different stability classifications. C) is the classifications based on the 40 °C data. D) is based on the 25 °C data. In A), B), C), and D), the red region is the region or predicated region of low stability, the yellow region is the region of medium stability, and the green region is the region of high stability.

Table 2-1 - List of the mAbs used in this study and their properties, including their IgG subclass, the type of light chain, the values of their aggregation propensities at both 40 °C and 25 °C from long-term stability experiments, and the computed values of their SAP Score and net charge at the formulation pH of the studies used to determine the aggregation propensities.

mAb Label	IgG subclass	Light chain type	Aggregation Propensity (25°C) [mL/(mg*month)]	Aggregation Propensity (40^{0}C) [mL/(mg*month)]	SAP Score	Net Charge[†]
mAb1	IgG1	κ	1.77 x 10^{-5}	7.17 x 10^{-5}	16.3	19.4
mAb2	IgG1	κ	9.51 x 10^{-6}	6.42 x 10^{-5}	23.9	20.1
mAb3	IgG1	κ	4.55 x 10^{-5}	5.96 x 10^{-4}	46.0	18.1
mAb4	IgG1	λ	5.54 x 10^{-6}	5.40 x 10^{-5}	24.1	28.9
mAb5	IgG1	κ	1.11 x 10^{-5}	5.62 x 10^{-5}	40.6	24.5
mAb6	IgG1	κ	6.94 x 10^{-5}	2.15 x 10^{-4}	35.6	23.9
mAb7	IgG1	λ	7.74 x 10^{-6}	1.58 x 10^{-4}	44.0	25.8
mAb8	IgG1	κ	1.21 x 10^{-5}	1.19 x 10^{-4}	35.2	24.4
mAb9	IgG1	κ	8.79 x 10^{-6}	8.01 x 10^{-5}	32.4	25.8
mAb10	IgG4	κ	1.05 x 10^{-5}	2.54 x 10^{-4}	24.6	15.5
mAb11	IgG4	κ	2.74 x 10^{-5}	2.12 x 10^{-4}	42.3	18.1
mAb12	IgG2	λ	1.15 x 10^{-5}	1.42 x 10^{-4}	28.7	27.9

Table 2-2 - Values for fitting parameters at 40 °C and 25 °C, as derived from the regression of 7 IgG1 mAbs (mAb1, mAb2, mAb3, mAb4, mAb5, mAb6, and mAb7). β is the regressed parameter β in Equation 2-1. The given DI values and aggregation propensities correspond to the different classifications (high, medium, and low stability) at 40 °C and 25 °C.

		Values for 40 ^{0}C	**Values for 25 ^{0}C**
β		0.0498	0.0815
DI values	High Stability	DI < 5.8	DI < 9.4
	Medium Stability	5.8 ≤ DI ≤ 18.4	9.4 ≤ DI ≤ 19.4
	Low Stability	18.4 < DI	19.4 < DI
Aggregation Propensity [mL/(mg*month)]	High Stability	k < 1.24 x 10^{-4}	k < 2.79 x 10^{-5}
	Medium Stability	1.24 x 10^{-4} ≤ k ≤ 3.91 x 10^{-4}	2.79 x 10^{-5} ≤ k ≤ 5.74 x 10^{-5}
	Low Stability	3.91 x 10^{-4} < k	5.74 x 10^{-5} < k

2.3.4. Functional forms for the Developability Index

Many functions that include both SAP and electrostatics are possible for DI. The primary method to differentiate the functions is the number of classifications that differ from the experimental classifications at 40 °C. Other factors are also considered, such as their actions at different limits, high charge, or numeric stability.

Two of the best performing functions were:

$$DI = [Antibody\ SAP\ Score] + \beta \times [Antibody\ Net\ Charge]^2$$

Equation 2-7

$$DI = [Antibody\ SAP\ Score] + \beta \times [Antibody\ Net\ Charge]$$

Equation 2-8

Both resulted in one deviation from the 40 °C data. In addition, both resulted in the same deviation; mAb5's stability was underestimated by predicting it to have medium stability although it was observed to have high stability. As both were equally accurate, Equation 2-7 was used as the overall function for the DI because it was slightly more numerically stable compared to Equation 2-8. Another benefit of using Equation 2-7 was that the square allowed for the pH range to extend above the isoelectric point, although this case did not arise in the available data set, as the electrostatic portion depends on the magnitude of the charge, not on the sign of the charge. The following functions were also considered:

$$DI = e^{[Antibody\ SAP\ Score] + \beta \times [Antibody\ Net\ Charge]^2}$$

Equation 2-9

$$DI = e^{[Antibody\ SAP\ Score] + \beta \times [Antibody\ Net\ Charge]}$$

Equation 2-10

Neither of these equations performed as well as the previous two functions, as these both deviated from the data for two test mAbs. However, these functions are able to generate the same predictions using the SAP Score of the Fv region as the predictions generated using the CDR. Following a similar idea, the following equation could be used:

$$DI = \frac{e^{\beta \times [Antibody\ SAP\ Score]}}{[Antibody\ Net\ Charge]^2}$$

Equation 2-11

This equation resulted in the same classifications as Equation 2-7 and its linearized form, Equation 2-8. Because Equation 2-10 did not perform any better than Equation 2-7, it was not used for further DI calculations. As was true with Equation 2-9 and Equation 2-10, Equation 2-11 made the same predictions with either the SAP Score of the CDR or the SAP Score of the Fv region. Other means can be used to account for the hydrophobicity and the electrostatic interactions. The first of these is the replacement of the SAP Score in Equation 2-7 by the effective hydrophobicity, the residue hydrophobicity multiplied by the ratio of solvent accessible surface area to a reference surface area. The positive effective hydrophobicity values are summed over the CDR and then used as the basis for a new fitting.

This case resulted in three deviations from experiments. Thus, the use of relative hydrophobicity performed more poorly than SAP because it does not take into account neighboring residues; two adjacent hydrophobic residues will have a greater effect on aggregation than two separated residues. The second case replaces the mAb net charge in Equation 2-8 with the isoelectric point (pI). The isoelectric point is determined using the structure and the set of pKa values from PROPKA. The pH is varied until the net charge on the whole mAb is zero, which resulted in two deviations from experimental values. The last two possibilities for ranking the mAbs are the DI based solely on SAP or solely on charge. If only net charge is considered, mAb1 and mAb2 would be classified as low-stability mAbs, but due to their low SAP Score, they are actually high-stability mAbs. Moreover, the use of the SAP Score alone resulted in five correct classifications at 40 °C. Thus, the SAP Score only, without any fitted parameter, was able to predict the long-term stability of mAbs to a good accuracy, and this accuracy could be further improved by combining the charge and the SAP Score (as in Equation 2-7).

There are many functions that can be used to compute a value for the DI. These can either be simple linear combinations of terms, as is the case in Equation 2-7 and Equation 2-8, or can be more complex ones, as is the case in Equation 2-9, Equation 2-10, and Equation 2-11. There are also other possible parameters that could have been used in place of the SAP Score and the net charge, but they do not perform as well as the SAP Score and the net charge. Lastly, neither the SAP Score nor the net charge is sufficient on their own to explain the observed differences in aggregation rate.

DI Classifications of Long-Term Stability

mAb	Expt.	Equations: 7, 8, 11	Equations: 9, 10	Relative Hydroph-obicity	SAP + βxpI	SAP Only	Charge Only
mAb1	🟩	🟩	🟩	🟩	🟩	🟩	🟨
mAb2	🟩	🟩	🟩	🟩	🟩	🟩	🟩
mAb3	🟥	🟥	🟥	🟨	🟨	🟨	🟨
mAb4	🟩	🟩	🟩	🟩	🟩	🟩	🟩
mAb5	🟩	🟨	🟩	🟨	🟨	🟨	🟨
mAb6	🟨	🟨	🟩	🟩	🟨	🟨	🟨
mAb7	🟨	🟨	🟩	🟨	🟨	🟨	🟨

Figure 2-6 - Classifications of the stabilities for 7 mAbs based on different functions to determine the DI. The first column is the value given by long-term stability studies. Later columns are classifications based on different functions for the evaluation of the DI. All data and classifications are at 40 °C. Green represents high stability, yellow represents medium stability, and red represents low stability.

2.4. Discussion

2.4.1. Validation of the Developability Index

In addition to those mAbs used to determine β, five other mAbs were tested to validate the DI tool. These five mAbs were not used in any way to determine any parameter or to determine a step in our method. These mAbs included two IgG1 mAbs (mAb8, mAb9), two IgG4 mAbs (mAb10 and mAb11) and one IgG2 mAb (mAb12). Four of these mAbs (mAb8, mAb9, mAb11, and mAb12) were stored in similar formulations to those used in the original set of mAbs. For each of these mAbs, DI was applied using Equation 2-1 (see Figure 2-7). The predictions were quite good given the uncertainty in the experimental data. Note that there was a bias in the data toward high-stability mAbs because some of the unstable mAb candidates were removed during the normal drug development process.

mAb	25 °C		40 °C	
	Expt.	DI	Expt.	DI
mAb8	🟩	🟩	🟩	🟩
mAb9	🟩	🟩	🟩	🟩
mAb10	🟩	🟩	🟨	🟨
mAb11	🟩	🟨	🟨	🟥
mAb12	🟩	🟩	🟨	🟩

Figure 2-7 - Predictions generated using the Developability Index for 5 test mAbs, 2 IgG1 (mAb8, mAb9), 2 IgG4 (mAb10 and mAb11), and 1 IgG2 (mAb12), compared to their experimental stability. Green represents high stability, yellow represents medium stability, and red represents low stability with respect to aggregation.

When this process was applied to five additional mAbs (mAb8, mAb9, mAb10, mAb11, and mAb12), the predictions worked well. In the case of the IgG1 mAbs, mAb8 and mAb9, both mAbs' aggregation propensity were correctly predicted at both temperatures. Of the two IgG4 mAbs, mAb10 and mAb11, mAb10 was correctly predicted at both 40 °C and 25 °C. However, mAb11's stability was underestimated at both temperatures. This result was reasonable because IgG4 mAbs are more prone to aggregation than IgG1 mAbs (***37***). In this case, the difference in stability of an IgG1 and IgG4 could at least be partially explained by the difference in the number of chargeable residues in the heavy chain, as an IgG4 mAb has a net charge roughly 6 e lower than an IgG1 mAb with the same variable region.

The IgG2 mAb, mAb12, was stored in the standard L-histidine buffer for the stability studies. In this case, it was predicted to have high stability at both 40 °C and 25 °C. However, experimentally, it had medium stability at 40 °C, while the prediction for 25 °C was correct. As was the case with the IgG4, there was some evidence that IgG2 mAbs were also less stable than IgG1 mAbs (***38***), (***39***). This fact could explain the less accurate results obtained at 40 °C: the DI predicted high stability, but experiments showed it to have medium stability. Because the constant regions of the heavy chain and the light chain were assumed to be constant among all the subtypes during the calculation of the DI, except for the calculation of the net charge, the hydrophobicity of these uncounted regions might be the cause of these incorrect predictions.

While DI gives the general trend of stabilities (Figure 2-7), there is a degree of error. There are a number of differences that exist between the different subtypes of IgG mAbs that could lead to reduced accuracy of prediction. These differences include different glycosylation patterns between mAbs and different cysteine bond shuffling, among other factors. All of these could lead to differences in the intrinsic rate of aggregation for the particular class of mAb or even to differences in the aggregation pathways accessible in solution. Future work could be performed to better understand the differences in the aggregation rates of IgG2 and IgG4 mAbs compared to IgG1 mAbs and to adjust the DI to better account for the inherent differences between the IgG subtypes.

2.4.2. pH dependence of the Developability Index

One of the strengths of the DI is that it is pH dependent. This dependence comes from the effect that pH has on the charge of a residue. This fact is important because the range of possible pH values for a liquid formulation leads to a very large range in net charge. As can be seen in Figure 2-8, for mAb5, there is a wide range of possible mAb net charges. Most of the drop in pH seen in this range (i.e., from 5 to 8) was due to the deprotonation of the histidine residues, of which there are nine in the constant region of the IgG1 heavy chain. While the pKa of most histidine residues falls between pH 6 and 6.5, the curve is smooth and does not show a particularly sharp drop in that range due to the use of partial charges.

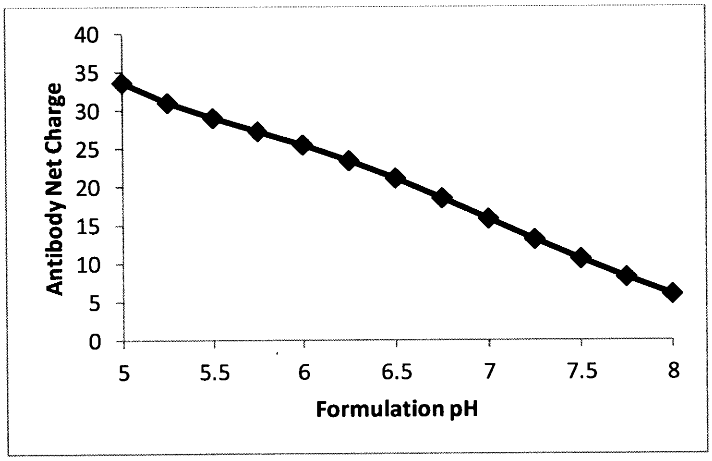

Figure 2-8 - Variation of mAb net charge as a function of the formulation pH for the moderately charged mAb mAb5

This effect of pH on the net charge can have a large effect on the predicted stability of mAbs. In Figure 2-9, this effect can be seen for two mAbs, one with low SAP and low charge, and another with high SAP and low charge. As shown, even a high SAP protein can be stabilized by greatly shifting the pH from the pI. This dependence can be useful in determining the optimal pH or possibly as an initial estimate for the optimal pH that a mAb should be stored at without any experiments. However, the DI does not account for the possible change in other degradation pathways that may be catalyzed by the change in pH.

The dependence on pH also leads to a difficulty that arises during the screening process, where two mAbs might be classified the same at one pH and vary substantially at others. For instance, as shown in Figure 2-9, at a pH of 5, both mAbs are highly stable, but as the pH increases to 6.5, one mAb still has high stability, and the other mAb has changed to the low-stability class. For comparison between mAbs during development, the dependence of stability on pH could be easily overcome by calculating the DI for the mAbs at several pH values.

mAb1	🟩	🟩	🟩	🟩	🟩	🟨	🟨
mAb3	🟩	🟨	🟨	🟥	🟥	🟥	🟥
pH	5.0	5.5	6.0	6.5	7.0	7.5	8.0

Figure 2-9 - The predicted long-term stability for two mAbs, mAb1 (a low SAP, moderate charge mAb) and mAb3 (a high SAP, low charge mAb), at various pH values. All predictions were made at 40 °C, and the definition of the stability classes is given in Table 2-2.

2.4.3. Application of the Developability Index in Protein Development

One of the difficulties in the development of a mAb-based drug is that there are many possible mAbs that will bind to the target antigen. If there are many possible candidates to be used for a potential drug, each of these mAbs needs to be tested to determine which mAb is the best candidate. Currently, initial screening is done through a series of experiments that measure various properties of a candidate mAb. However, these tests need to be done quickly with small amounts of protein so that they cannot truly capture the long-term stability properties of a mAb. The DI was therefore developed to predict the long-term stability properties of mAbs.

There are two distinct roles that the DI can fill in the development of mAb-based pharmaceuticals. The DI can be used to prioritize mAb candidates for experimentation and as a tool to guide the mutation of unstable mAbs. The prioritization of mAbs can be done several ways, such as by the removal of low-stability candidates from the candidate pool or by ranking the mAbs and using this rank to prioritize experiments. This prioritization allows for the earlier identification of more stable drug candidates. While it is true that one (or more) of the discarded mAbs may be stabilized to an acceptable level by the judicious choice of excipients, some of these unstable mAbs may not be sufficiently stabilized for use in a drug. The benefit of the DI is that it would allow for a more stable mAb to be selected in the discovery phase and would help to minimize the time and money needed during the formulation development phase.

The DI also allows for selection of mutants with increased stability relative to the original mAb. These mutations are determined either through locating high SAP residues and mutating these residues or through the mutation of a residue that is not charged to a residue that can be charged. Through both of these methods, many potential mutations can be identified, and the mutations can then be ranked by the DI to select the best without any experiments.

One of the benefits of the DI is that it does not depend on any information about the protein other than the structure, which can be determined from the primary sequence using homology modeling. With the mAb structure, both SAP calculations and charge calculations can be performed, and the DI can be calculated. While this tool can help screen mAbs, it can also be used to help guide mutations of less-stable mAbs to improve their stability. The DI could also be used to evaluate the risk of the future development of specific mAbs. For instance, if the current mAb were to have a high DI value, it would indicate the possibility that more work would be required to stabilize the mAb during the formulation development stage.

It is important to note two of the limitations to the current DI tool: the applicability of the DI to only mAbs and the neglect of other degradation pathways. Currently, this DI tool has only been validated for mAbs, but it may be directly applicable to other classes of proteins, the primary challenge to which would be obtaining an accurate 3-D structures. Another limitation is that this tool only accounts for aggregation. However, there are a number of other factors that also need to be considered when selecting a mAb from a set, such as protein expression, mAb purification, interactions with the formulation, and other degradation pathways. These other degradation pathways include oxidation of Methionine residues, deamidation reactions, hydrolysis, and bonding of cysteine residues between mAbs (*5*,*6*). The effects of these other factors need to be considered separately from DI.

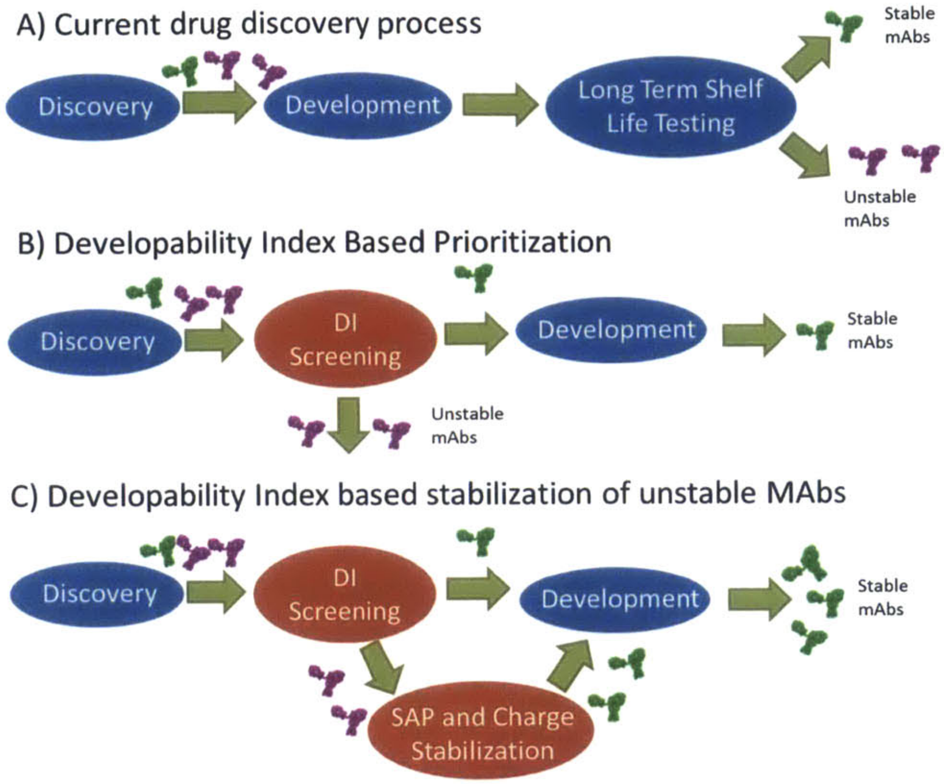
A) Current drug discovery process
Discovery
Development
Long Term Shelf Life Testing
Stable mAbs
Unstable mAbs
B) Developability Index Based Prioritization
Discovery
DI Screening
Development
Stable mAbs
Unstable mAbs
C) Developability Index based stabilization of unstable MAbs
Discovery
DI Screening
Development
Stable mAbs
SAP and Charge Stabilization

Figure 2-10 - A comparison of the proposed method of mAb development using the DI and the existing method of mAb-based drug development during the discovery phase of drug development. Part A represents the current state of drug discovery, where many mAbs are developed through a range of experiments to separate stable mAbs (in green) and unstable mAbs (in purple). Part B of the figure shows DI screening for prioritization, where the unstable mAbs are removed, and the stable mAbs move on. Part C shows DI removing unstable mAbs from the candidate pool and then using SAP and the net charge to select mutants of these mAbs for greater stability, which can then be considered for drug development.

2.5. Conclusion

The Developability Index (DI) is a tool that allows for the rapid screening of mAbs for their aggregation propensity without any experimental data. The DI is based on two parameters: SAP Score of the CDR and net charge of the full-length mAb, to account for hydrophobicity and electrostatic interactions, respectively. Both of these are dependent on the tertiary structure of the mAb, which is a required input either from experimental data or homology algorithms. Here, we present the details of the algorithm and its implementation. In addition, we validate the DI on a number of mAbs. Seven were used to determine the β parameter; another five were used to test the accuracy of the DI. The accuracy was very good, considering the uncertainty in the experimental data. Therefore, the DI could be applied in the discovery and early development phases to select for stability among candidates. It could also be used to estimate the risk in going forward with a particular candidate. Finally, it could help guide mutations of a mAb to stabilize it against aggregation.

3. APPLICATION OF SAP AND DI TO ANTIBODY FRAGMENTS

3.1. Introduction

Monoclonal antibodies are successful drugs because they exhibit significant therapeutic potential with few side effects. However, antibodies are less stable than low molecular weight chemical compounds and are prone to chemical and physical degradation (***40***). Degraded antibodies can aggregate (***41***), and protein aggregates can exhibit low efficacy and trigger immunogenic responses (***9***,***42***). The generation of higher quality antibody therapeutics requires understanding the mechanisms underlying aggregation and establishing production technologies that can strictly control aggregate formation.

The tendency of a protein to aggregate likely depends on its conformational and colloidal stabilities (***43***,***44***). Proteins unfold upon exposure to stresses such as heat, pH, and agitation, and then often aggregate (***45***)), suggesting that conformational stability towards such stresses may decrease the propensity to aggregate. However, some_proteins remain monomeric and monodisperse in the unfolded state, suggesting that the colloidal stability of the unfolded state would impact the propensity of a protein to aggregate. The conformational and colloidal stabilities of a protein likely depend both on the protein (amino acid composition, sequence, and structure) and environment (buffer, salts, and other solvent components) (***46***,***47***,***48***). Consequently, assessing the propensity and exploring the mechanisms underlying aggregation require the systematic investigation of both the conformational and colloidal stability of a protein in a wide range of solution conditions.

Immunoglobulin G (IgG) is a multi-domain protein exhibiting complex molecular behavior, but recent studies have reported a relationship between the aggregation reaction of the antibody and the conformational and colloidal stabilities of its domains. Calorimetry experiments showed that the CH2, CH3, and Fab regions unfold at different temperatures (***49***). Enk et al. reported that thermally-unfolded aglycosylated CH2 region led to aggregation of Fc, and that the presence of anions destabilized the CH2 region and accelerated the aggregation reaction (***49***). Kim et al. reported that aggregation rates for intact antibody were strongly influenced by the conformational stability of the Fab region (***50***). Furthermore, Buchner and coworkers showed that murine IgG1 domains (i.e., whole-IgG1, Fab, CH3, VH, VL, CH1, and CL) form molten-globule-like intermediate structures under specific acidic conditions (pH 2 and 100 mM NaCl; CL: pH 2, 175 mM NaCl) (***51***,***52***,***53***,***54***). These intermediate structures, also called the alternatively folded state (AFS), exhibit molecular properties unique from both native and random coil conformations. Moreover, antibody domains in the AFS generally oligomerize, suggesting that the AFS is involved in antibody aggregation mechanisms. Taken together, the evidence to date suggests that each antibody domain has the potential to induce antibody aggregation, and that aggregation occurs through complicated interactions between multiple antibody domains, each of which may have different conformational and colloidal stabilities. Since understanding antibody aggregation mechanisms is clearly challenging, we propose that an extensive investigation of the conformational and colloidal stabilities of individual antibody domains is a useful approach towards understanding antibody aggregation mechanisms.

3.2. Materials and Methods

3.2.1. Protein Preparation

Gene sequences for four constant domains (i.e., CH1, CH2, CH3, and CL) were designed based on the amino acid sequence of human IgG1 containing kappa light chain. The N- and C-terminal amino acid residues of each domain were determined from IgG1 crystal structures (PDBID: 1N8Z and 3D6G). The C-terminal cysteine residues of CH1 and CL were deleted to prevent undesired dimerization. Each domain had a His-tag sequence at its N-terminus. Codons were optimized for *Escherichia coli* expression, and cleavage sites for restriction enzymes were extended to the 5'- and 3'-ends. CH1 and CL gene fragments were synthesized using overlap extension polymerase chain reaction (PCR). The CH2 and CH3 gene fragments were obtained by PCR amplification of the pFUSE-hIgG1-Fc1 plasmid (Invtrogen). The CH1, CH2, and CL gene fragments were digested with NdeI/EcoRI and ligated into pET-22b(+) (Novagen). The CH3 gene fragment was digested with NcoI/BamHI and ligated into pET-16b (Novagen). *Escherichia coli* strain Origami^{TM} B (DE3) (Novagen) was transformed with plasmid vectors coding each domain and cultured in Luria-Bertani media containing 100 μg/ml ampicillin, 20 μg/ml kanamycin, and 20 μg/ml tetracycline. Recombinant gene expression was induced by the addition of isopropyl β-D-1-thiogalactopyranoside to a final concentration of 0.5 mM at 25 °C. After overnight culture, cells were centrifuged and sonicated. Domain proteins were purified from the cell lysates using His GraviTrap^{TM} (GE Healthcare). Purified domain proteins were solubilized in 20 mM MES pH 6.0 buffer by dialysis and applied to a Resource S cation exchange chromatography column (GE Healthcare) equilibrated with the same buffer as the sample. Purified proteins were concentrated by ultrafiltration and applied to a Superdex 75 (10/300) gel filtration chromatography column (GE Healthcare) equilibrated with 20 mM citrate-phosphate buffer, pH 7.0, containing 150 mM NaCl. The purities of the samples were confirmed by tricine SDS-PAGE, and by mass spectrometry using an Axima-TOF^2 (Shimazu) and 4700 Proteomics Analyzer (Applied Biosystems). Cytochrome c (m/z, 12,362) and apomyoglobin (m/z, 16952) were used as internal TOF mass standards.

CH1-CL heterodimer protein was synthesized by CH1 and CL co-expression in *Escherichia coli.* Based on the report by Corisdeo and Wang (*55*), a CH1-CL co-expression gene fragment was designed, placing the CL gene sequence before the CH1 gene sequence, and inserting spacer DNA and an additional ribosomal binding site between the two genes. CH1 and CL in the CH1-CL co-expression gene fragment both contained a C-terminal cysteine residue to allow intermolecular disulfide bond formation. The CH1-CL co-expression gene fragment was obtained by PCR amplification, digested with NdeI/EcoRI, ligated into pET-22b(+) (Novagen), then used to transform *Escherichia coli* strain Origami$^{\text{TM}}$ B (DE3) (Novagen). Protein expression and purification were performed as described for the other domains.

Forty-nine sets of protein solutions (seven pH values (pH 2-8, at one unit intervals), each at seven salt concentrations (NaCl 0-300 mM, at 50 mM intervals)) were prepared by dialysis against 20 mM glycine-HCl buffer pH 2 or 20 mM citrate-phosphate buffer pH 3-8 containing 0-300 mM NaCl. Dialysis was performed for 18-20 hours at 4 °C using a micro dialyzer (Toru-kun TOR-3K, Nippon Genetics).

3.2.2. Circular Dichroism Spectroscopy

Circular dichroism (CD) measurements were carried out using a J-805 spectropolarimeter (Jasco). Far-UV CD spectra were recorded from 195 nm to 260 nm at 1 nm intervals at 20 °C with 50 μM protein in 0.2 mm pathlength quartz cuvettes. All spectra were corrected by subtracting the buffer spectrum. Two independent measurements were made for the pH 2 and pH 3 samples, and one for all other solution conditions.

3.2.3. Empirical Phase Diagrams

Empirical phase diagrams (EPDs) of each antibody domain were drawn according to the reports of Middaugh and co-workers (***48***,***56***). First, the CD and fluorescence spectra were normalized using the following equation:

$$\tilde{x}_{i,\text{pH,NaCl}} = \frac{\left(x_{i,\text{pH,NaCl}} - \bar{x}_i\right)}{s} \qquad (1.)$$

where $\tilde{x}_{i,\text{pH,NaCl}}$ and $x_{i,\text{pH,NaCl}}$ are the normalized and measured signal intensities, respectively, at wavelength *i* for each combination of pH and NaCl concentration, and $\bar{x_i}$ is the signal intensity averaged for all solution conditions at wavelength *i*. The symbol *s* denotes the standard deviation of the intensities at all wavelengths and all solution conditions. Next, data matrix A was constructed, containing all 49 normalized spectra: data matrix A for the CD data consisted of 66 columns and 49 rows, while data matrix A for the fluorescence data consisted of 151 columns and 49 rows. Singular value decomposition (SVD) of these data matrices was conducted using the following equation:

$$A = USV^{\mathrm{T}} \qquad (2.)$$

where U is the left singular vector whose columns contain orthonormal eigenvectors of the column space information in A. Each column of U contains the significant spectral fractions. S is the singular value which quantifies the relative importance of each vector in U and V. V is the right singular vector whose columns contain orthonormal eigenvectors of the row space information in A. Each column of V contains a titration profile for each corresponding column of U. The superscript T denotes transposition of the matrix V (*57*). SVD calculations were carried out using IGOR Pro (Wavemetrics).

Each EPD was drawn using the three most significant right singular vectors (Figure 3-3). If insignificant vectors apparently due to noise data were found, one or two significant right singular vectors were selected and used for the EPD. Noise in the data was judged using the values of a contribution ratio and an autocorrelation function calculated from the singular values and the singular vectors, respectively.

To visualize the EPDs using a RGB color scheme, the values of the right singular vectors were normalized using the following equation:

$$\tilde{V}_{i,\mathrm{pH,NaCl}} = \frac{(V_{i,\mathrm{pH,NaCl}} - V_{i,\mathrm{min}})}{(V_{i,\mathrm{max}} - V_{i,\mathrm{min}})} \times 256 \qquad (3.)$$

where $\tilde{V}_{i,\mathrm{pH,NaCl}}$ and $V_{i,\mathrm{pH,NaCl}}$ are the normalized and calculated values, respectively, of the i-th right singular vector in each combinations of pH and NaCl concentration. $V_{i,\mathrm{min}}$ and $V_{i,\mathrm{max}}$ are the minimum and maximum values of the i-th right singular vector among all 49 solution conditions, respectively. RGB color schemes corresponding to the normalized values of each i-th right singular vector were mapped for every solution condition, with the first, second, and third right singular vectors shown as red, green, and blue, respectively. The RGB color intensities correspond to the normalized values of the right singular vectors. No color was used for data representing noise (insignificant singular vectors). For example, the RGB representation of EPD would be black if all three normalized values were zero and white if they were 256. Similar colors for the different solution conditions indicate that the protein exists in similar conformational sub-states. The color scheme manipulation was carried out using Excel (Microsoft). Finally, to classify the 49 conformational sub-states into several major-states, a hierarchical clustering analysis was carried out using R. The resulting classified conformational major-states were separated by drawing a thick black or white line on each EPD.

3.2.4. Dynamic Light Scattering

Dynamic light scattering (DLS) measurements were conducted using a Zetasizer Nano S (Malvern). All measurements were performed at 20 °C using quartz microcells with a sample volume of 12 μL at a protein concentration of 50 μM and a scattering angle of 173°. Samples were filtered through a 0.22 μm centrifugal filter (Millipore). Two independent measurements were made for the pH 2-4 samples, and one for all other solution conditions. The translational diffusion coefficient and the particle size diameter were calculated from the autocorrelation function using Zetasizer Software (Malvern).

3.2.5. Estimation of Aggregation Propensities from Various Parameters for Amino Acid Composition, Sequence, and Surface Structure

Amino acid compositions and general protein characteristics were calculated according to the reports by Goh et al. and Thomas et al. (***58***,***59***). Aggregation prone regions (APRs) as representations of sequence characteristics were calculated using several sequence-based aggregation prediction algorithms: Aggrescan (***60***), PASTA (***61***), Zyggregator (***11***), and Tango (***10***). The calculations were performed using the default settings, except that the pH conditions for Zyggregator and Tango were set to 3.0, and the temperature and ionic strength conditions for Tango were set to 293.15 K and 0.3 M, respectively. Output scores for Na^4vSS (normalized sum of averaged aggregation propensity value, Aggrescan), best energy (PASTA), Z^{agg} (Zyggregator), and the Agg parameter (Tango) were used as parameters for estimating the aggregation tendency of the domains. The spatial aggregation propensity (SAP) and the developability index (DI) were used as representations of structural surface properties (***17***,***62***). SAP was calculated using the native structure of the domains extracted from the crystal structures of Fab and Fc (PDBID: 1N8Z (***63***) and 3D6G (***64***) respectively). The SAP score was calculated using all atoms of each domain except for the N-terminal methionine and the His-tag sequence. The SAP radius was set as 10 angstroms for all calculations. The p*K*a of individual residues in the pH 3 condition was calculated using PROPKA 3.0 (***65***). The β value for the 20 mM histidine solution, 25 °C condition was selected from known values (***62***)

3.3. Results

3.3.1. Analysis of Secondary Structure of Antibody Domains Using CD Spectroscopy

To investigate the secondary structures of the domains, the far-UV CD spectra of the proteins in 49 solvents (pH 2-8 and 0-300 mM NaCl) were measured (Figure 3-1). Acid-induced unfolding of the secondary structure was observed for each domain. The CD spectra of non-native states were affected by NaCl concentration.

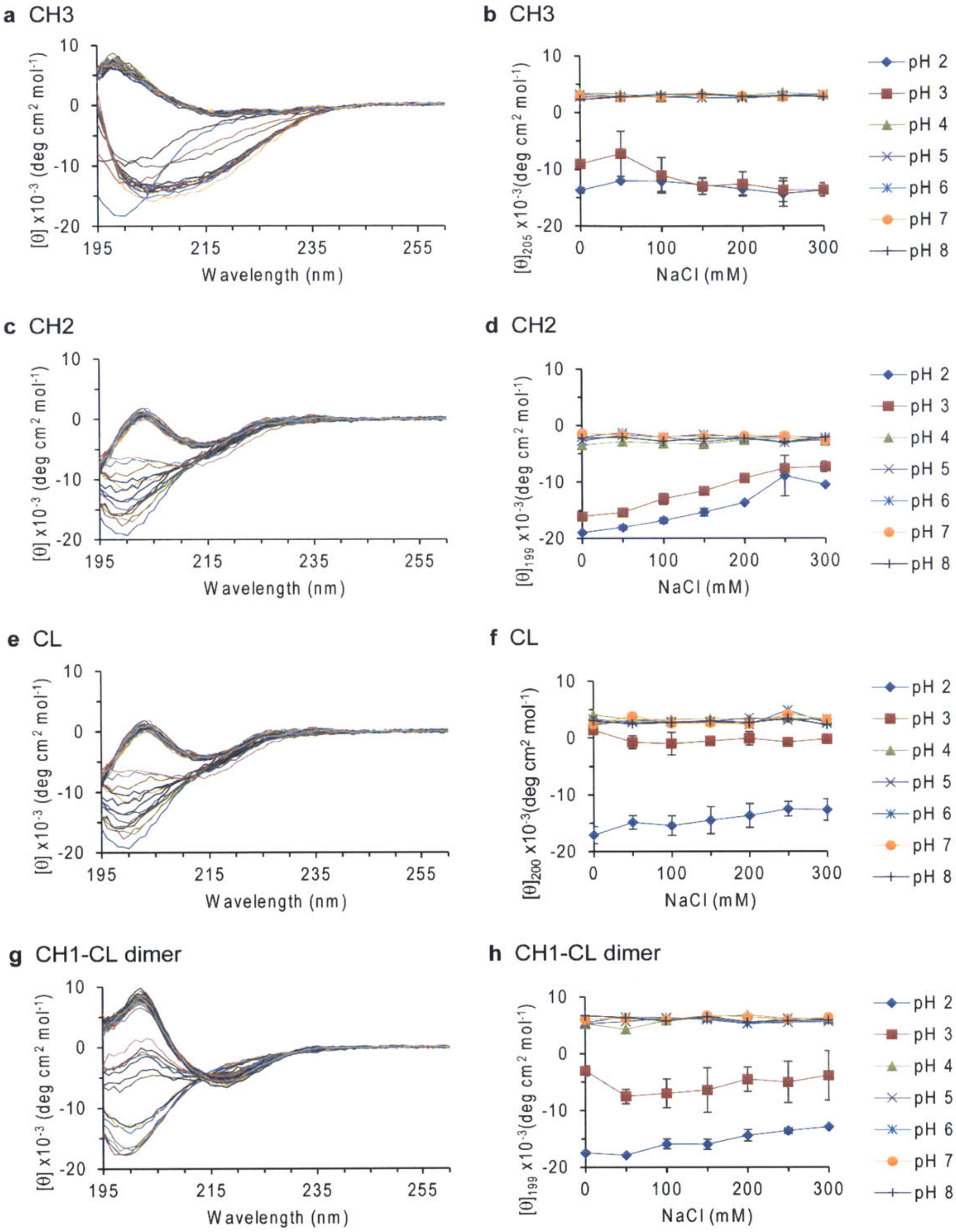

Figure 3-1 - Circular dichroism (CD) spectra of CH3 (a), CH2 (c), CL (e), and CH1-CL dimer (g) at 49 solution conditions (pH 2-8 and 0-300 mM NaCl), and molar residual ellipticities of CH3 at 205 nm (b), CH2 at 199 nm (d), CL at 200 nm (f), and CH1-CL dimer at 199 nm (h).

The CD spectra of CH3 from pH 4 to pH 8, 0 mM-300 mM NaCl, were almost identical, indicating that CH3 retains its native secondary structure under these conditions (Figure 3-1a). The spectra showed a maximum intensity at 198 nm and two minimum intensities at 218 nm and 229 nm. In contrast, at pH 2 and pH 3, 0 mM NaCl, the CD spectra were typical of a random coil, with the minimums shifting to 200 nm. The absolute intensity of the minimum at pH 2 was greater than that at pH 3 (Figure 3-1b). At both pH 2 and pH 3, the minimum shifted to longer wavelengths as the NaCl concentration increased up to 100 mM; above 100 mM NaCl, the minimum remained around 205 nm and the spectra did not change.

The CD spectra of CH2 from pH 4 to pH 8, 0-300 mM NaCl, were almost identical (Figure 3-1c) and were indicative of native secondary structure, showing maximum and minimum intensities at 203 nm and 214 nm, respectively. In contrast, at pH 2 and pH 3, 0 mM NaCl, the spectra were typical of a random coil, with the minimum shifting to 199 nm. The absolute intensity of the minimum at pH 2 was larger than that at pH 3. At both pH 2 and pH 3, the absolute intensity of the minimum gradually decreased as the NaCl concentration increased (Figure 3-1d).

CD spectra of CL corresponding to native secondary structure were observed from pH 4 to pH 8, 0-300 mM NaCl (Figure 3-1e), with maximum and minimum peaks at 202 nm and 217 nm, respectively. Although the wavelength of the maximum and minimum at pH 3 remained consistent with the native state, the absolute intensity of the maximum decreased as the NaCl concentration increased from 0 mM to 50 mM (Figure 3-1f). At pH 2, 0 mM NaCl, the spectrum was indicative of a random coil, with the minimum shifting to 200 nm. The absolute intensity of the minimum changed only slightly as the NaCl concentration increased, in contrast to the much larger changes observed with CH2 and CH3.

The CD spectra of the CH1-CL dimer from pH 4 to pH 8, 0-300 mM NaCl, were almost identical (Figure 3-1g), with maximum and minimum peaks at 202 nm and 217 nm, respectively. No significant shifts in the maximum and minimum were observed at pH 3, but their intensities changed as the NaCl concentration increased. The absolute intensity at the maximum increased as the NaCl concentration increased from 0 mM to 50 mM and decreased as the NaCl concentration increased from 50 mM to 300 mM (Figure 3-1h). At pH 2, 0 mM NaCl, the spectrum was characteristic of a random coil, with the minimum shifting to 199 nm. At pH 2 and pH 3, the absolute intensity at the minimum decreased as the NaCl concentration increased, but the decrease was small compared to those observed with CH3 and CH2 under the same conditions.

3.3.2. Analysis of the Tertiary Structure of the Antibody Domains Using Intrinsic Tryptophan Fluorescence Spectroscopy

The intrinsic tryptophan fluorescence spectra of the four domains in the 49 pH/NaCl solvent conditions were measured in order to investigate the tertiary structure of the domains (Figure 3-2). CH2 and CH3 each contain two tryptophan residues (Trp277 and Trp313 in CH2; Trp381 and Trp417 in CH3), while CH1 and CL contain only one tryptophan residue (Trp161 in CH1; Trp148 in CL). Consistent with the secondary structure data obtained using CD, acid-induced unfolding of the tertiary structure was observed for each domain. The fluorescence spectra of the non-native states were also affected by NaCl concentration.

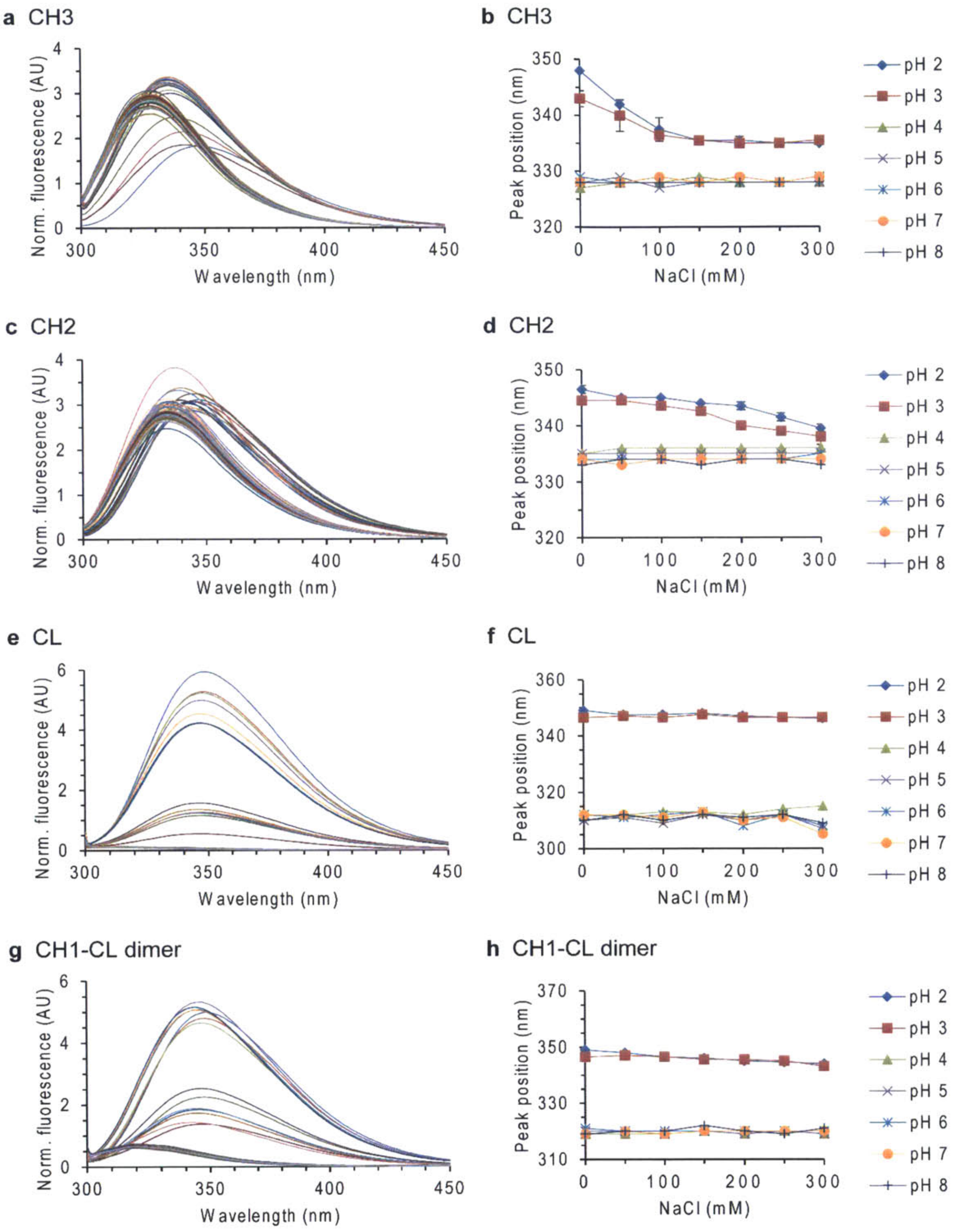

Figure 3-2 - Intrinsic tryptophan fluorescence spectra of CH3 (a), CH2 (c), CL (e), and CH1-CL dimer (g) at 49 solution conditions (pH 2-8 and 0-300 mM NaCl), and wavelengths at the peak maximum of CH3 (b), CH2 (d), CL (f), and CH1-CL dimer (h).

The fluorescence spectra of CH3 from pH 4 to pH 8, 0-300 mM NaCl, were essentially identical, indicating that CH3 retains its native tertiary structure under these conditions (Figure 3-2a). A fluorescence maximum was observed at 328 nm. At 0 mM NaCl, pH 2 and pH 3, the maximum fluorescence shifted to 348 and 343 nm respectively (Figure 3-2b), whereas the maximum wavelength decreased as the NaCl concentration increased. Above 100 mM NaCl, the maximum fluorescence remained unchanged at around 335 nm.

The fluorescence spectra of CH2 from pH 6 to pH 8, 0-300 mM NaCl, were almost identical (Figure 3-2c). Spectra corresponding to the native tertiary structure showed maximum fluorescence at 334 nm. Little change was observed at pH 4 and pH 5 (see section below). At pH 2 and pH 3, 0 mM NaCl, the maximum fluorescence shifted to 347 and 345 nm, respectively (Figure 3-2d), but shifted to shorter wavelengths as the NaCl concentration increased. The fluorescence maximum was 340 nm and 338 nm at pH 2 and pH 3, respectively, at 300 mM NaCl.

The fluorescence spectra of CL corresponded to the native tertiary structure from pH 4 to pH 8, 0-300 mM NaCl (Figure 3-2e), with a fluorescence maximum at 311 nm. The fluorescence intensity of the native state was very low because CL has only one tryptophan residue whose fluorescence is quenched by an adjacent intrinsic disulfide bond (31). At pH 2 and pH 3, there was no fluorescence quenching and the maximum shifted to 349 and 347 nm, respectively (Figure 3-2f). The wavelength maximum obtained at pH 3 was little affected by NaCl concentration. The fluorescence intensity at pH 2 was higher than at pH 3 and shifted to shorter wavelength as the NaCl concentration increased, reaching 346 nm in 300 mM NaCl.

The fluorescence spectra of CH1-CL dimer from pH 4 to pH 8, 0-300 mM NaCl, were almost identical (Figure 3-2g), with a maximum observed at 320 nm. The fluorescence intensity of the native state was very low, possibly due to fluorescence quenching of the CH1 and CL tryptophan residues. At pH 2 and pH 3, fluorescence quenching disappeared and the maximum shifted to 349 and 347 nm, respectively. The fluorescence intensity was higher at pH 2 than at pH 3, but at both pH conditions, the maximum fluorescence shifted to shorter wavelengths as the NaCl concentration increased, to 343 nm and 344 nm, respectively, in 300 mM NaCl (Figure 3-2).

3.3.3. Evaluation of the Conformational States of the Antibody Domains Using EPD

The conformational stability of the domains was visualized and characterized by conducting singular value decomposition (SVD), and by drawing empirical phase diagrams (EPDs) using CD and fluorescence spectral data (CD-EPD and FL-EPD, respectively) (Figure

3-3). SVD analysis can remove noise from spectra, allowing extraction of more significant components contributing to spectral changes (*57*). EPD is a visualization method for easily recognizing to what degree the extracted components dominate the spectrum (*56*,*48*). Different colors on EPDs indicate dissimilar conformations.

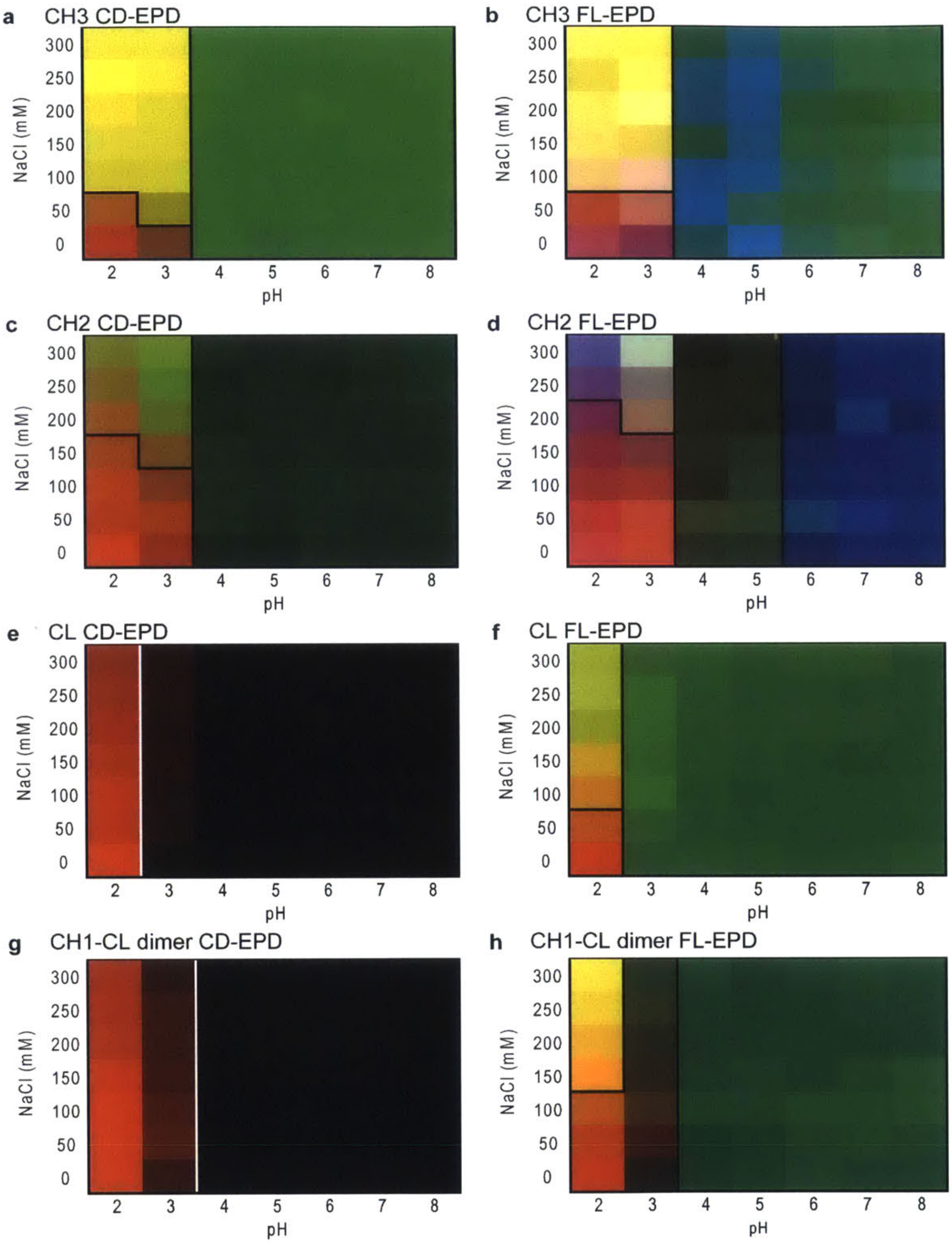

Figure 3-3 - Empirical phase diagrams (EPDs) based on spectral data from circular dichroism (CD-EPDs) and fluorescence (FL-EPDs) at 49 solution conditions. (a), (c), (e), and (g) corresponding to CD-EPDs of CH3, CH2, CL, and CH1-CL dimer, respectively. (b), (d), (f), and (h) correspond to FL-EPDs of CH3, CH2, CL, and CH1-CL dimer, respectively. Different colors in the EPDs indicate conformational differences between sub-states. Conformational major-states obtained by hierarchical cluster analysis were separated by drawing a thick black or white line on the EPDs

EPDs for CH3 based on CD (Figure 3-3a) and fluorescence (Figure 3-3b) data allowed the 49 conformational sub-states obtained at the 49 solvent conditions to be classified into three conformational major-states. The major-state corresponding to the native conformation is illustrated by the same color from pH 4 to pH 8 at all NaCl concentrations. Acid unfolding is clearly recognized by differences in color between pH 3 and pH 4. Non-native sub-states at pH 2 and pH 3 could be classified into two distinct non-native major-states that are dependent on the NaCl concentration: above 100 mM NaCl, the color remains constant, suggesting that CH3 adopts a unique conformation under these conditions.

The 49 conformational sub-states of CH2 were classified into three and four major-states from the CD- and FL-EPDs, respectively (Figure 3-3c, d). A native major-state of CD-EPD was found from pH 4 to pH 8 at all NaCl concentrations. FL-EPD showed a native major-state at pH 6-8 at all NaCl concentrations, while another major-state was apparent at pH 4-5. The presence of this alternative major-state was supported by the fluorescence peak shift from 334 nm to 335 nm (pH 4) or to 336 nm (pH 5) (Figure 3-3d) due to a partially distorted state (see Additional Discussion in Supplementary Material for details). In both the CD- and FL-EPDs, two non-native major-states were found at pH 2 and pH 3, and a gradual conformational change expressed as a color gradient was observed as a function of NaCl concentration.

The CD- and FL-EPDs of CL revealed that the 49 conformational sub-states could be classified into two and three major-states, respectively (Figure 3-3e, f). Both EPDs showed the presence of a non-native major-state at pH 2 but a native major-state at pH 3. A minor color change at pH 3 suggests the co-existence of a small number of unfolded molecules with the native-fold protein. In the FL-EPD constructed using data obtained at pH 2, the colors gradually changed depending on the NaCl concentration, and two non-native major-states were obtained.

The CD- and FL-EPDs of CH1-CL dimer allowed classification of the 49 conformational sub-states into two and three major-states, respectively (Figure 3-3g, h). Both EPDs showed the presence of a non-native major-state at pH 3 but a native major-state at pH 4. In the FL-EPD using data obtained at pH 2, the colors gradually changed depending on the NaCl concentration, and two non-native major-states were obtained.

In general, the color differences observed with CD-EPD correlated well with those of FL-EPD, suggesting that the secondary and tertiary structures changed cooperatively in all domains.

3.3.4. Evaluation of the Colloidal States of the Antibody Domains Using DLS, SEC, and BN-PAGE Measurements

To investigate the colloidal stability of the domains, we measured the particle size of the domains in the 49 solvent conditions using DLS and drew particle size diagrams (PSDs) (Figure 3-4a-d). A quantitative comparison is provided in Table 1, which shows the obtained particle sizes of a native state (pH 7, 150 mM NaCl) and two non-native states (pH 2, 0 mM NaCl, and pH 3, 300 mM NaCl). Furthermore, an oligomeric state of the domains under the same three conditions was analyzed using BN-PAGE and SEC (Figure 3-5). BN-PAGE and SEC measurements of the domains at pH 7, 150 mM NaCl, confirmed that CH2 and CL exist as monomers, whereas CH3 and CH1-CL dimer exist as dimers (Figure 3-5a, b). At pH 2, 0 mM NaCl, CH3 eluted later than the other domains on SEC (Figure 3-5c), suggesting that the CH3 homo-dimer dissociated into monomers during acid unfolding. The particle sizes obtained from DLS and the elution times from SEC at pH 2, 0 mM NaCl (Figure 3-5c) indicate that non-native CH2, CH3, and CL exist as monomers while the non-native CH1-CL dimer remained a dimer, probably due to an inter-molecular disulfide bond. Higher-oligomeric states of CH3 and CH2 were observed at pH 2 and pH 3 as the NaCl concentration increased (Figure 3-4a, b, Figure 3-5d). Interestingly, the particle size of CH2 increased at pH 4, unique among the domains (Figure 3-4b), even though its native secondary structure was retained (Figure 3-3c). DLS measurements of CH2 at pH 4 showed a multimodal particle size distribution, whereas oligomerized CH3 and CH2 at pH 2 and pH 3 provided single peaks. The particle sizes of CL and CH1-CL dimer did not change even at pH 2 and pH 3 (Figure 3-4c, d), suggesting that CL and CH1-CL dimer retain their monomeric and dimeric state, respectively, under the solvent conditions tested.

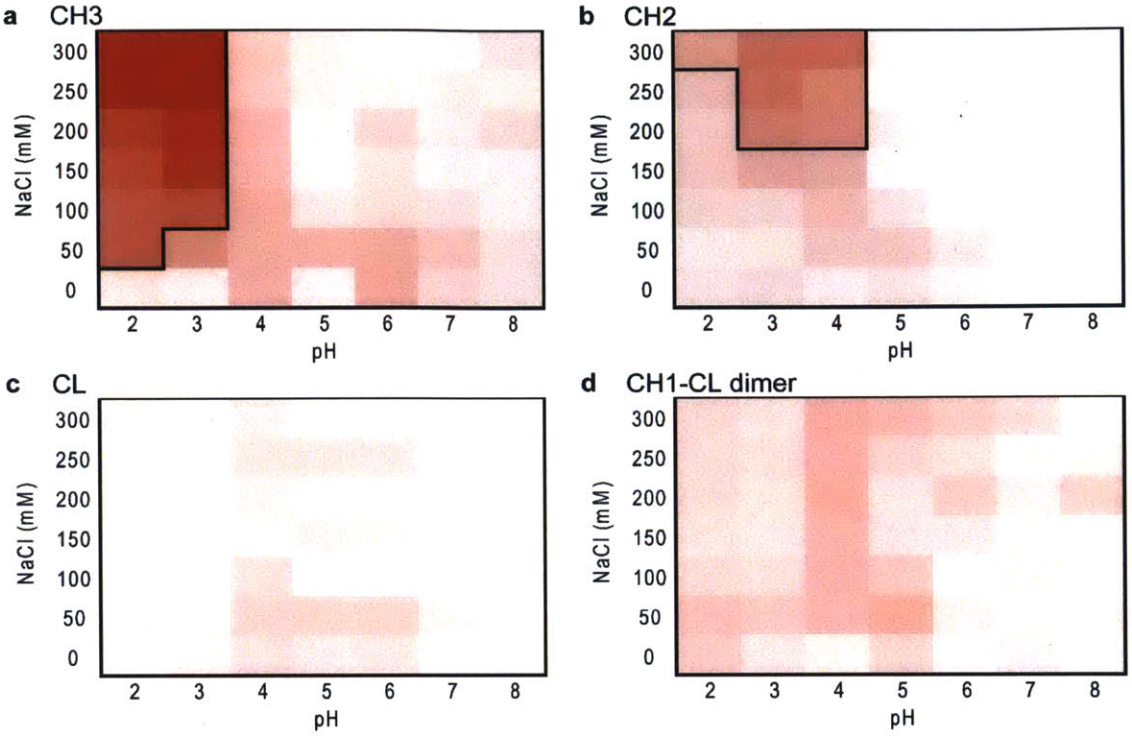

Figure 3-4 - Particle size diagrams for CH3 (a), CH2 (b), CL (c), and CH1-CL dimer (d) measured by dynamic light scattering (DLS) at 49 solution conditions (pH 2-8 and 0-300 mM NaCl). Red intensity corresponds to the relative particle size of samples. A deep red indicates large particles. Oligomerization states determined by hierarchical cluster analysis were separated by drawing a thick line on the diagrams.

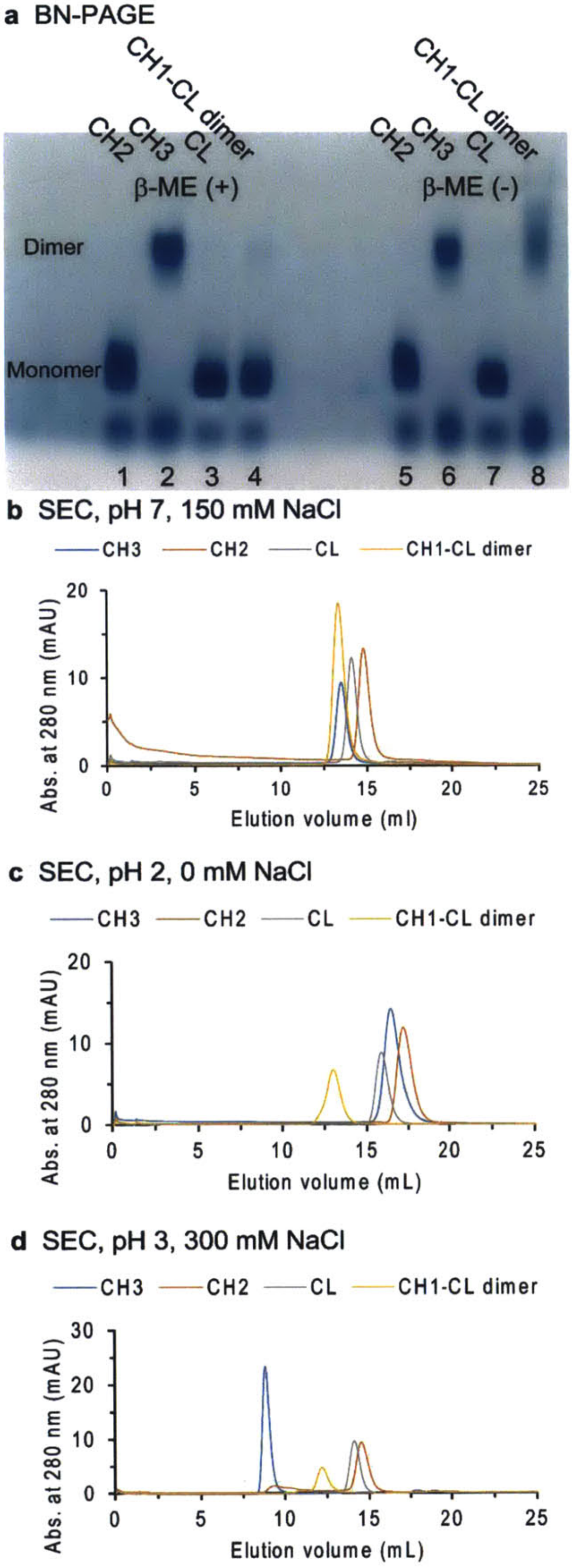
a BN-PAGE
CH2
CH3
CL
CH1-CL dimer
β-ME (+)
CH2
CH3
CL
CH1-CL dimer
β-ME (-)
Dimer
Monomer
1 2 3 4
5 6 7 8
b SEC, pH 7, 150 mM NaCl
CH3
CH2
CL
CH1-CL dimer
Abs. at 280 nm (mAU)
0
10
20
0
5
10
15
20
25
Elution volume (ml)
c SEC, pH 2, 0 mM NaCl
CH3
CH2
CL
CH1-CL dimer
Abs. at 280 nm (mAU)
0
10
20
0
5
10
15
20
25
Elution volume (mL)
d SEC, pH 3, 300 mM NaCl
CH3
CH2
CL
CH1-CL dimer
Abs. at 280 nm (mAU)
0
10
20
30
0
5
10
15
20
25
Elution volume (mL)

Figure 3-5 - Oligomeric state analyses using blue native gel polyacrylamide electrophoresis (BN-PAGE) and size exclusion chromatography (SEC). (a) BN-PAGE of CH2 (Lane 1 and 5), CH3 (Lane 2 and 6), CL (Lane 3 and 7), and CH1-CL dimer (Lane 4 and 8) at pH 6.8. BN-PAGE was electrophoresed on 4-20% polyacrylamide gradient gels. The disulfide bonds of samples applied to lanes 1 to 4 were reduced by treating with 5% beta-mercaptoethanol (�-ME) prior to analysis, whereas the cysteines in the samples applied to lanes 5 to 8 were oxidized. SEC of the domains at pH 7, 150 mM NaCl (b), pH 2, 0 mM NaCl (c), and pH 3, 300 mM NaCl (d)

3.4. Discussion

3.4.1. Generalized Phase Diagrams by Merging Conformational and Colloidal Major-States

The conformational and colloidal stabilities of isolated antibody constant domains under a wide range of pH and salt conditions were investigated by systematic biophysical measurements, and the EPDs and PSDs of the domains were obtained (Figure 3-3 and Figure 3-4). Significant similarities were observed between these two matrixes. For example, the distinguishable regions at low pH and high NaCl concentration in the EPDs of CH3 and CH2 were very close to the red regions in the PSDs of CH3 and CH2 (Figure 3-3a-d and Figure 3-4a, b), indicating that the conformational change and oligomer formation resulting from ionic effects are likely associated. We therefore classified the conformational and colloidal sub-states of the domains under 49 solvent conditions into three conformational and colloidal major-states (Figure 3-6). The first major-state is the native state (N state); this is the preferred state near neutral pH and exhibits native conformation and monodispersity. The second major-state is the monomeric monodispersed non-native state (MMNN state), observed preferentially at acidic pH and low NaCl concentration. The third major-state is the polymeric polydispersed non-native state (PPNN state), predominant at acidic pH and higher NaCl concentration. CH3 exhibited all three major-states (Figure 3-6a), whereas CL and CH1-CL dimer exhibited only the N state and MMNN state under the conditions tested (Figure 3-6c, d). In addition to these three major-states, CH2 exhibited another conformational major-state, a partially distorted state (Figure 3-6b).

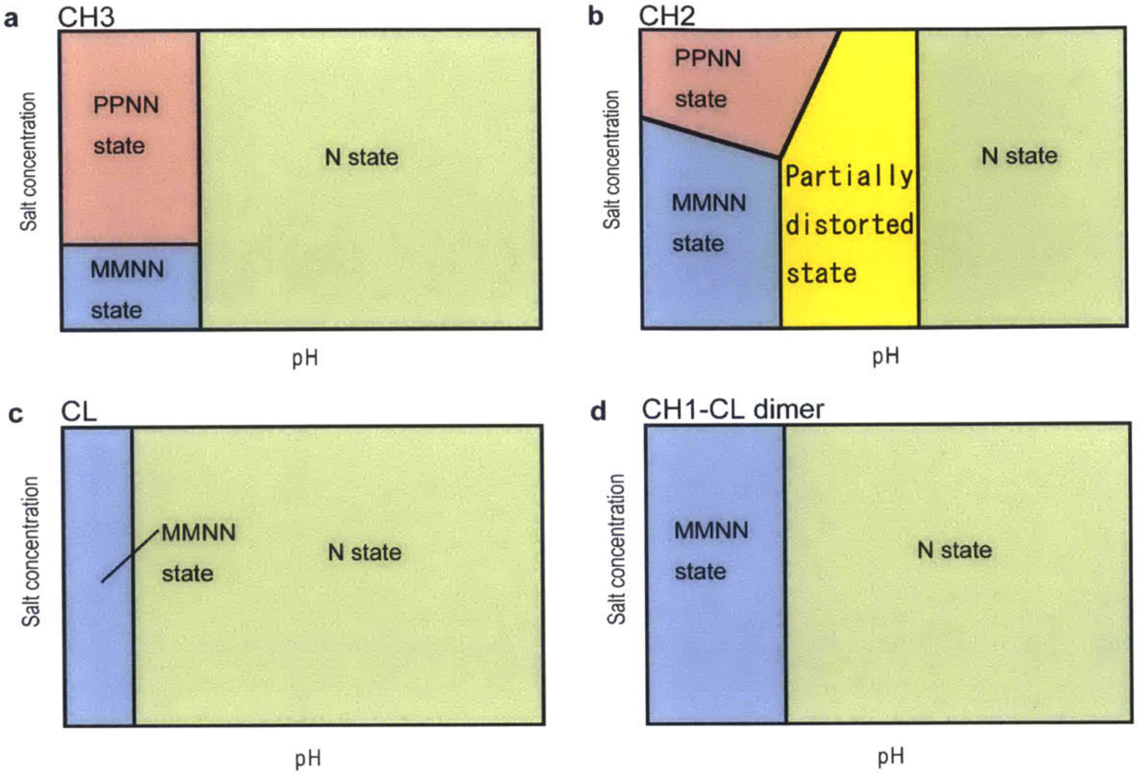

Figure 3-6 - Generalized phase diagrams for the antibody constant domains. The N state, MMNN state, PPNN state, and a partially distorted state in CH3 (a), CH2 (b), CL (c), and CH1-CL dimer (d) are shown, as applicable

3.4.2. Determination of the Order of Conformational Instabilities among Antibody Domains

The conformational transition from N state to non-native state(s) was observed at a different acidic pH for each domain (Figure 3-3 and Figure 3-6). The EPDs show boundaries between pH 3 and pH 4 for CH3, CH2, and CH1-CL dimer, but between pH 2 and pH 3 for CL.

Conformational stability was evaluated quantitatively by performing two- or three-state transition model fittings and calculating the fraction of the three conformational major-states (N state, MMNN state, and PPNN state) against the CD spectra at pH 2 and pH 3 (Figure 3-7). The CD spectra obtained at pH 7, 150 mM NaCl, pH 2, 0 mM NaCl, and pH 3, 300 mM NaCl, were taken as pure spectra of the N state, MMNN state, and PPNN state, respectively (See Materials and Methods). A three-state transition model fit the CH3 and CH2 data well and allowed the fraction of each conformational major-state to be estimated (Figure 3-7a-d). The two-state transition model fit the CL and CH1-CL dimer data well (Figure 3-7e-h). At pH 3, 0 mM NaCl, the fraction of N state was approximately 32% for CH3, 21% for CH2, 92% for CL, and 61% for CH1-CL dimer (Figure 3-7b, d, f, h), which suggests that each domain was partially unfolded to some extent. For CH3, the fraction of PPNN state rapidly increased as the NaCl concentration increased, to over 80% at 100 mM NaCl (Figure 3-7a, b). For CH2, the fraction of the PPNN state gradually increased as the NaCl concentration increased (Figure 3-7c, d). If the sum of the fraction of protein in the PPNN and MMNN state at pH 3, 0 mM NaCl is taken as an indicator of conformational instability, then the order of the conformational instability of the antibody constant domains is: CH3 > CH2 > CH1-CL dimer > CL (Figure 3-8a).

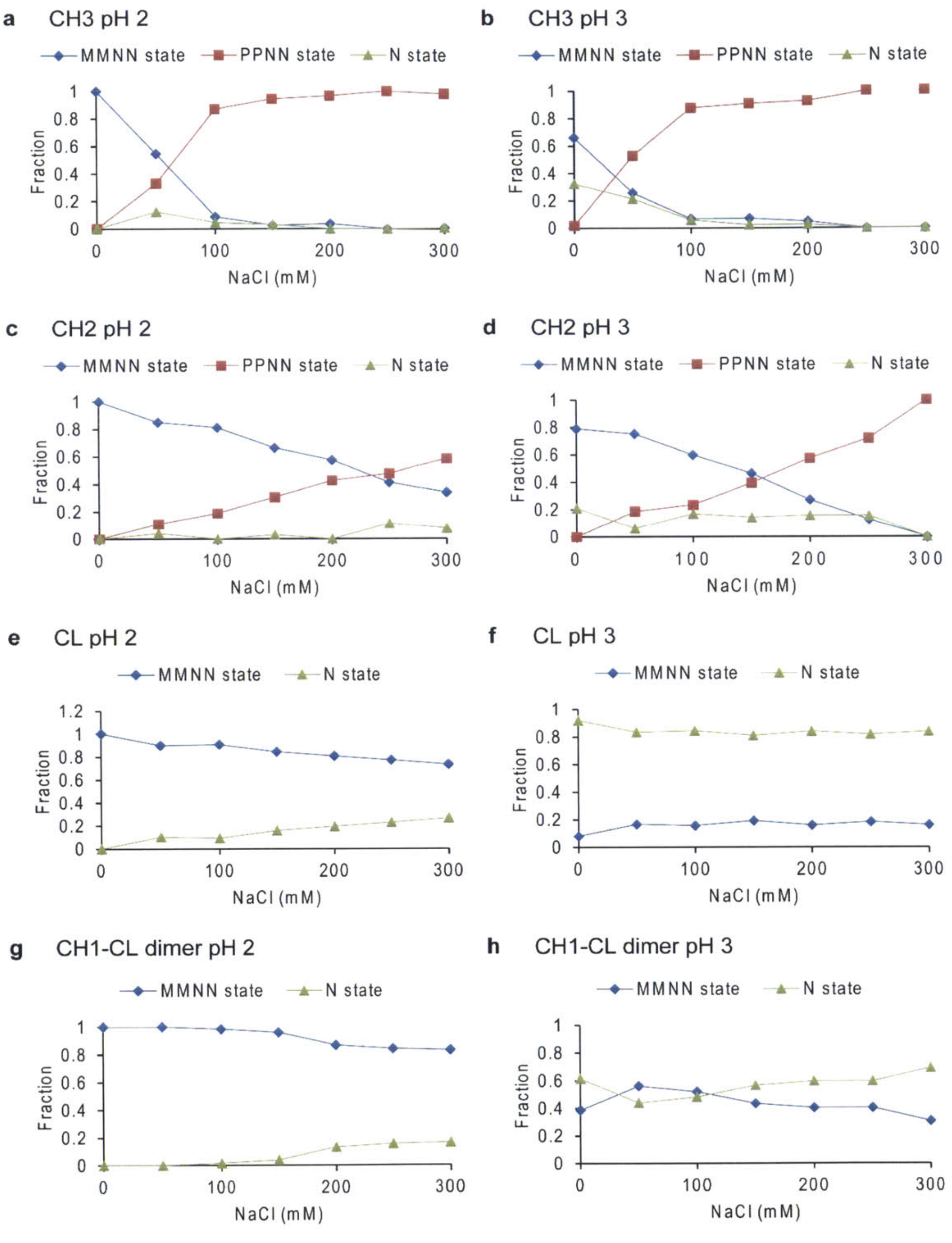

Figure 3-7 - Fractions of N-state, MMNN-state, and PPNN state of the domains. CH3 at pH 2 (a) and pH 3 (b), CH2 at pH 2 (c) and pH 3 (d), CL at pH 2 (e) and pH 3 (f), and CH1-CL dimer at pH 2 (g) and pH 3 (h) at 0-300 mM NaCl are shown. The values for CH3 and CH2 were obtained using a three-state transition model, and those for CL and CH1-CL dimer were obtained using a two-state transition model.

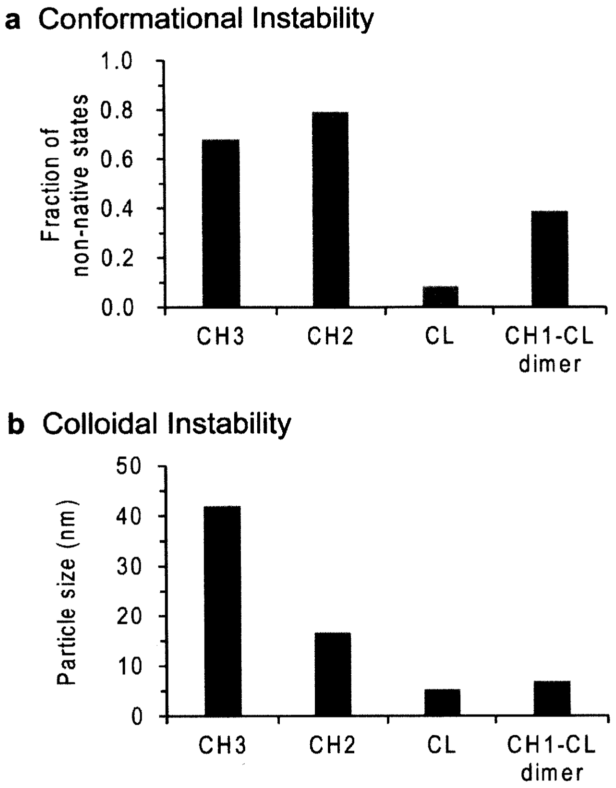

Figure 3-8 - Conformational and colloidal instabilities of the domains. The values on the vertical axis indicating conformational instability (a) were calculated by the sum of the fractions of the MMNN and PPNN states in 0 mM NaCl, pH 3 solution. The values on the vertical axis indicating colloidal instability (b) correspond to the particle size in 300 mM NaCl, pH 3 solution.

3.4.3. Aggregation Propensities of Antibody Domains Estimated from Several Protein Properties

What characteristics contribute to the different aggregation propensities? Although the structures of the four constant domains are almost identical (RMSD of backbone atoms < 1.5 angstrom), their sequence homology is relatively low (sequence identity < 30%). Therefore, we hypothesized that the differences in colloidal stability might depend on the amino acid composition, sequence, or surface charge characteristics.

The amino acid composition of each domain was analyzed first. It was reported that several specific amino acid compositions affect protein solubility, so we calculated these parameters for each domain in an effort to explain the differences in colloidal stability. The results did not suggest any significant relationship between amino acid composition and the observed colloidal instability of the domains.

We next searched the aggregation prone regions (APRs) of the domains using several sequence-based aggregation prediction algorithms: Aggrescan, PASTA, Zyggregator, and Tango. All programs predicted several APRs at approximately similar positions. However, the order of the output scores of three of the programs did not coincide with the experimentally-observed order of colloidal instability (Figure 3-9): only Zyggregator provided an output consistent with our experimental data.

Lastly, we assessed the aggregation propensities of the domains based on structural surface properties using Spatial Aggregation Propensity (SAP) and Developability Index (DI) algorithms (Figure 3-10). SAP searches for surface hydrophobic patches and DI evaluates the propensity for protein aggregation based on the SAP score and surface charge. SAP and DI calculations were performed on the native static structure of each domain. The SAP calculations identified hydrophobic patches concentrated on the interface region of each domain (Figure 3-10a-d) and the DI calculation interestingly estimated the order of aggregation propensity to be: CH3 > CH2 > CL > CH1-CL dimer; this order is very similar to the order of colloidal instability determined in the present study (CH3 > CH2 > CH1-CL dimer ≈ CL; Figs. 11e and 9b).

The aggregation propensities predicted by SAP/DI calculations on the static crystal structure agreed with our experimental results, despite the aggregation-prone conformation of the PPNN state being clearly different from the native structure. This led to speculation that domains in the PPNN state adopt particular conformations whose structural surface properties, such as hydrophobicity and electrostatic potential, are similar to the native structure in the N state. Therefore, the relationship between SAP/DI and Zyggregator with our experimental result suggests that relatively small hydrophobic surface patches and/or short hydrophobic sequence segments primarily participate in aggregate formation.

DI calculations suggested that the CH3 homodimer is less prone to aggregation than the CH3 monomer (Figure 3-10e). This correlates with our experimental results that CH3 remained a stable, dispersed dimer in the absence of acid stress, and that the CH3 dimer dissociated into monomers and oligomerized at pH 3 and below (Figure 3-3, Figure 3-4, and Figure 3-5). Therefore, the dissociation of CH3 dimer by acidic conditions is strongly associated with the higher aggregation propensity of CH3. This viewpoint raises the possibility that CH1 aggregates when the CH1-CL dimer is dissociated by acid conditions. Feige et al. reported that murine CH1 formed oligomers at pH 2, 100 mM NaCl (***54***). However, CH1-CL dimer did not form oligomers in the present experiments. The higher colloidal stability of CH1-CL dimer may arise from the close proximity of CH1 and CL, since the two domains are covalently connected through their C terminal disulfide bond, allowing CL to function as a "solubility tag" and improve the solubility of CH1.

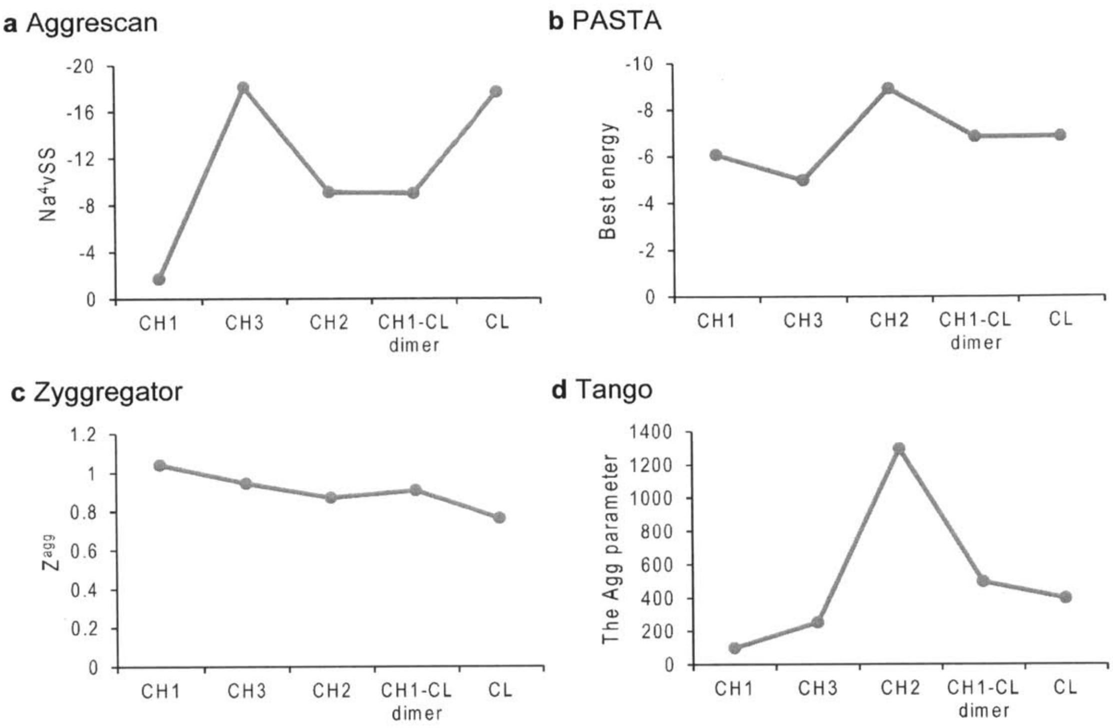

Figure 3-9 - Sequence-based estimation of aggregation propensity. Na4vSS (Aggrescan) (a), best energy (PASTA) (b), Zagg (Zyggregator) (c), and the Agg parameter (Tango) (d) were plotted against the experimental order of colloidal instability of each domain

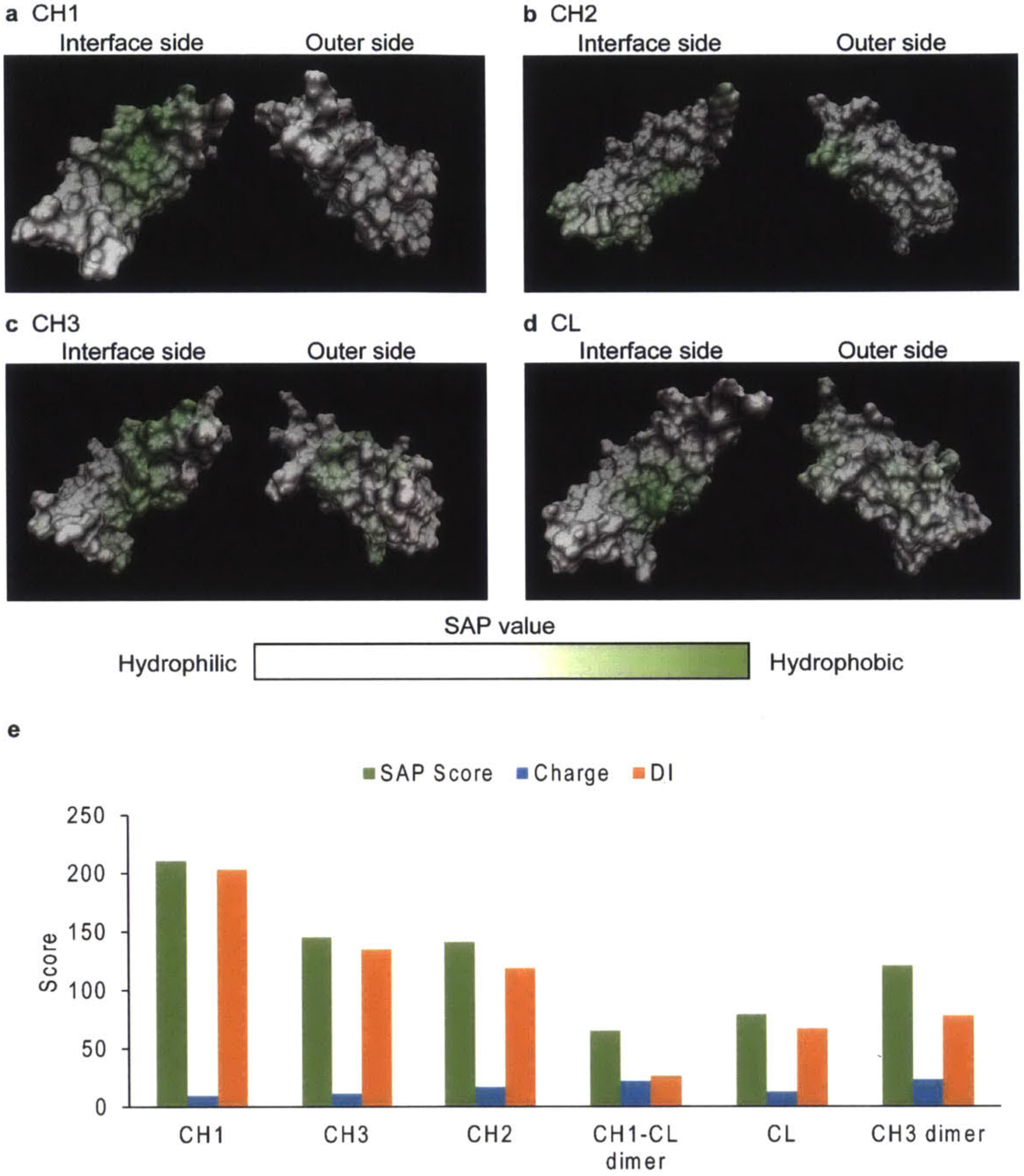

Figure 3-10 -Surface-based estimation of aggregation propensity. SAP values were mapped on the structures of CH1 (a), CH2 (b), CH3 (c), and CL (d). "Interface side" and "Outer side" are illustrated using the native structure of IgG (PDBID: 1n8z and 3d6g). Green regions represent hydrophobic regions and white regions are hydrophilic regions. (e) The calculation results of the SAP score, net charge, and DI of the domains.

3.5. Conclusion

Taken together, the experimental data presented here showing the conformational and colloidal stabilities of the isolated domains, and the good correlation of these data with SAP/DI calculations, suggest not only why the various domains exhibit different stabilities, but also how whole antibody aggregation occurs. We propose that antibody aggregation under acidic conditions involves highly structured domains exhibiting essentially native-like surface properties, rather than random coil conformations, and thus the aggregated protein is structurally distinct from amorphous aggregates characteristic of simple polymers.

4. A SYSTEMATIC STUDY OF THE TIME DEPENDENCE OF RESIDUE AVERAGED SAP AND THE SAP SCORE DURING A MOLECULAR DYNAMICS SIMULATION.

4.1. Introduction

Aggregation is a major degradation pathway for monoclonal antibodies (mAbs). Several attempts have been made to predict the aggregation prone regions and the likelihood that aggregation will present a problem for these proteins (*10*), (*11*), (*12*), (*66*), (*13*). One of these tools is SAP, or spatial aggregation propensity (*13*), (*16*), (*17*), (*18*), (*19*). This was further expanded with the SAP Score of a protein (*62*), which can used with the charge of an antibody to predict its likelihood to aggregate. One of the features that sets SAP apart from most of the other aggregation prediction methods, is that it is based on the 3-D structure of the protein, rather than just the primary sequence of the protein. This allows for direct inclusion of interchain interactions and other tertiary structure impacts that are not present in the protein's sequence. However, a protein can adopt many different conformations in solution. In order to accurately calculate SAP of a protein, a collection of structures must be used. One of the simplest methods to gather a collection of structures is through an all-atom molecular dynamics (MD) simulation.

Several of these works (*13*), (*62*), (*67*) have used an MD simulation to sample a number of conformations, by which an average SAP can be computed, to better estimate an individual residue's role in aggregation. However, they have not determined how long of an MD simulation is needed to adequately sample the possible conformation space. Some works have suggested to rely solely on a crystal or homology structure of the protein, (*62*), as short MD does not appear to improve the average. It is unknown how accurately a single structure can capture the ensemble of protein conformations in this regard. In order to address these concerns, the correlation time of SAP and the SAP Score will be investigated to determine how long a simulation is needed to sample many conformations, and then these simulation averages will be compared to the crystal structure to quantify the error in using only a single structure rather than the ensemble of conformations.

4.2. Methods

4.2.1. SAP Definition and Uses:

The SAP, spatial aggregation propensity, tool was developed to identify regions of hydrophobic residues on the protein surface. It has been previously used to investigate a number of antibodies (*13*), (*16*), (*17*), (*18*), (*19*), and has been used to design mutants of enhanced stability. The SAP score of an atom is defined as:

$$SAP_{atom\ i} = \sum_{\substack{Simulation\\ Average}} \left\{ \sum_{\substack{Residues\ with\ at\\ least\ one\ atom\\ within\ R\ of\ atom\ i}} \left[\frac{SASA\ of\ side\ chain\ atoms\ within\ radius\ R}{SASA\ of\ side\ chain\ atoms\ of\ fully\ exposed\ residue} \times \begin{matrix} Residue\\ Hydrophobicity \end{matrix} \right] \right\}$$

Equation 4-1

where

5) *SASA* is the 'solvent accessible surface area' of side chain atoms contained within radius R from atom i.
6) *SASA of side chain of fully exposed residue* is obtained by calculating the SASA (solvent available surface area) of the side chain of the middle residue in the fully extended conformation of tripeptide "Ala-X-Ala" (where X is the residue of atom i).
7) *Residue Hydrophobicity* is obtained from the hydrophobicity scale of Black and Mould (*28*). The scale is normalized such that glycine has a hydrophobicity of zero, the most hydrophobic residue (PHE) has a value of 0.5, and the least hydrophobic residue (ARG) has a value of -0.5. Hydrophobic residues have residue hydrophobicity values greater than 0. Residue hydrophobicity values less than zero are more hydrophilic.

SAP can be used in two ways, the first is to compute the average SAP value of a residue, an average of the SAP values for each atom. This quantity can be used to select which residues should be mutated to decrease aggregation propensity. The second use option is to compute the SAP Score of the protein, the sum of all hydrophobic atoms. This is a measure that can be used to rank proteins by their aggregation propensity.

$$SAP\ Score\ of\ Protein = \sum_{\substack{All\ atoms\\ SAP\ Value>0}} (SAP\ Value_{atom\ i})$$

Equation 4-2

$$\overline{SAP_{residue}} = \frac{1}{N_{atoms\ in\ Residue}} \sum_{\substack{All\ atoms\\ in\ residue}} (SAP\ Value_{atom\ i})$$

Equation 4-3

4.2.2. Definition of Correlation Time and the Statistical Inefficiency

During the time evolution of a MD simulation, each step is based on its predecessor, this causes the state at any step to be correlated to its predecessor, therefor when looking at average properties over the course of a simulation, some number of steps must be skipped to avoid correlation between adjacent steps. Because the time a property is correlated varies between different properties it must be computed separately for both the SAP Score and the SAP value per residue each residue. In order to determine how long these properties are correlated, the correlation time, the method presented by Allen and Tildesley (***68***) will be used. This method is based on breaking the simulation in blocks of time. During each of these time blocks computing average value of the property of interest will be calculated. This collection of blocked averages will then be used to computing the statistical inefficiency which is defined in Equation 4-4, where $\tau_b$is the number of steps being averaged over, $\sigma^2(\langle \mathcal{A} \rangle_b)$ is the standard deviation of the of the block averages of property $\mathcal{A}$ and $\sigma^2(\mathcal{A})$ is the standard deviation of the property during the entire simulation.

$$s = \lim_{\tau_b \to \infty} \frac{\tau_b \sigma^2(\langle \mathcal{A} \rangle_b)}{\sigma^2(\mathcal{A})}$$

Equation 4-4

Equation 4-4 yields the statistical inefficiency. This is a measure of how many steps are correlated, and thus processing these steps is inefficient and adds no new information to the average. This is directly proportional to the correlation time, and can be converted between using Equation 4-5 by using the physical amount of time passes between frames.

$$\tau_c = \frac{s}{2\left(\frac{time}{step}\right)}$$

Equation 4-5

4.2.3. Homology Modeling and Simulation Detail

In order to investigate correlation time for several cases, the antigen binding fragments for 5 antibodies are simulated here. These five Fabs (labeled Fab1, Fab2, Fab3, Fab4, and Fab5 from (***69***)) have very similar sequences (>95% identity to Fab1). Therefore, the homology model for each Fab was based on the crystal structure for Fab1 (PDBID: 4G6F (***70***)). These structures were then used to carry out classical molecular dynamics simulations using the Gromacs package (***71***). The protein was simulated using the AMBER99SB force field (***72***). The protein was solvated by adding a solvent shell of 10 Å. This water was modeled with the TIP3P force field (***32***). The charges for histidine were assumed to be at pH 7, with a periodic boundary condition was applied in all three dimensions. Sufficient ions were added to neutral the system. First, the system was minimized. The minimized structure was used as a starting point for a 150 ns NPT simulation, using a 2 fs time step with a temperature of 300 K, and a pressure of 1 bar. 30 ns of the simulation was used to equilibrate the protein, while the remaining 120 ns was used to evaluate the property of interest.

4.3. Results

4.3.1. Stability of proteins

The RMSD (root mean square displacement) was used to quantify the stability of each protein during the course of its simulation. Each protein was found to be stable during the simulation. A sample RMSD graph for Fab1 can be found in Figure 4-1. The first 30 ns of simulation time was used to equilibrate the system while the final 120 ns of simulation was for the following calculations.

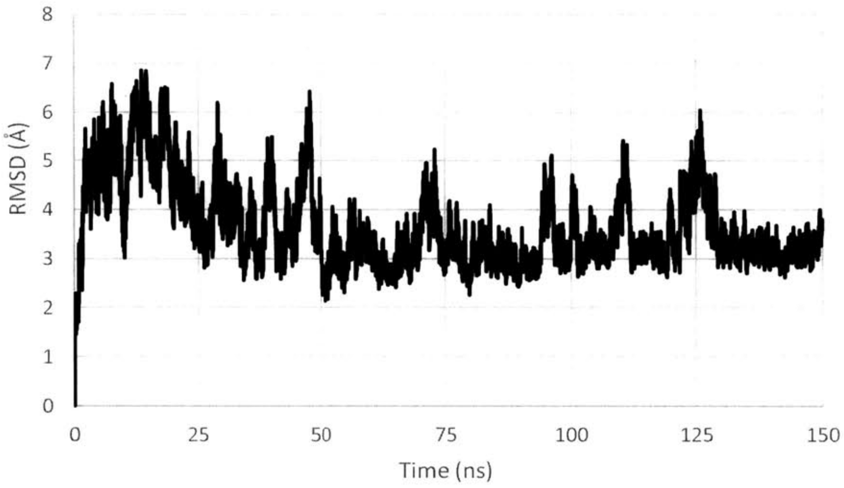

Figure 4-1 - The root mean squared deviation of the structure for Fab1 over the course of the simulation

4.3.2. Correlation time of SAP Score

Correlation time measurements for all five Fabs were carried out. Examples of this can be seen in Figure 4-2 and Figure 4-3. Figure 4-2 shows the range of SAP Scores sampled during the simulation. Fab1 showed a wide variation in SAP Score during the simulation, about 20% of average. This is much larger than the average seen in previous works of approximately 5% (***62***) of the SAP Score over the simulation. This was collection of SAP Scores was used to determine the collation time, of about 3ns for Fab1's SAP Score at 5 Å. A similar process was followed for the remaining Fab segments and a table of the results can be found in Table 4-1. In all cases the statistical inefficiency converges for the SAP Score, normally between 2 and 10 ns, depending on the SAP radius and the fragment being studied.

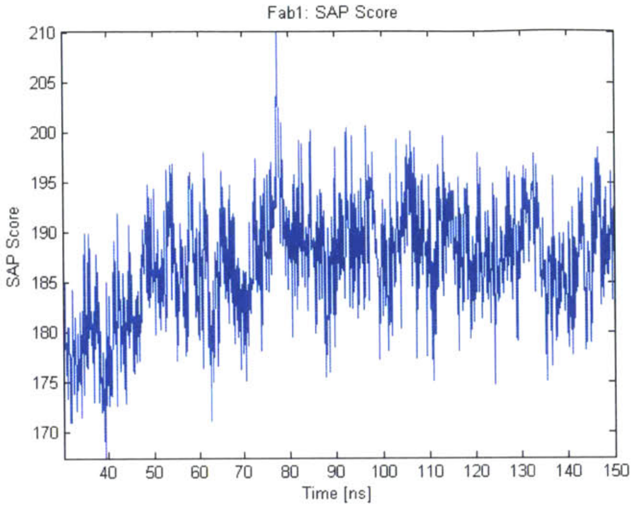

Figure 4-2 - SAP Score at 5 Angstrom Over last 120 ns of simulation for Fab1. SAP Radius 5Å.

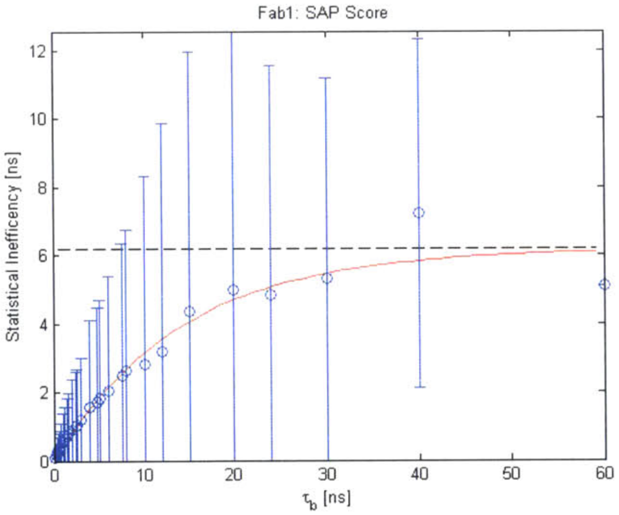

Figure 4-3 - Statistical inefficiency at various time block sizes. Blue dots are measures of the statistical inefficiency for a given time block size, while the blue bars represent plus and minus 1 standard deviation at the given statistical inefficiency. The black line is an estimate of the correlation time, based on the limit of the statistical inefficiency as the size of the blocks goes to infinity. The red line is a present to show the general trend. All data gathered from simulation of Fab1. SAP Radius 5 Å.

Table 4-1 – Comparison of SAP Score for the start of the MD run, the average SAP Score over the final 120ns of simulation, and the standard deviation of the SAP Score over the course during the final 120ns of simulation. All data presented using four different SAP radii 5, 7, 10, and 15Å.

Fab	SAP Score (initial)				$\langle SAP\ Score \rangle$				$\sigma_{SAP\ Score}$			
	5 Å	7 Å	10 Å	15 Å	5 Å	7 Å	10 Å	15 Å	5 Å	7 Å	10 Å	15 Å
Fab1	191	285	337	213	187	281	350	234	5.71	10.5	24.6	45.4
Fab2	175	257	305	188	178	263	315	167	6.12	12.4	25.5	31.6
Fab3	185	287	356	216	178	264	325	184	6.98	13.4	31.9	40.4
Fab4	165	245	298	120	174	261	325	178	5.23	11.1	21.2	26.6
Fab5	177	247	295	166	172	247	284	147	5.97	11.6	18.3	32.7

4.3.3. Correlation Time of Residue's SAP Value

A similar procedure was carried out for each residue during the simulations. For illustration purposes, select figures are shown. The relative frequency of statistical inefficiencies for all residues in Fab1 are shown in Figure 4-4. Closer inspection shows that some of these are not be converged, see Figure 4-5, which shows the statistical inefficiency for residue 147. Therefore the values in Figure 4-4, represent a lower bound on the statistical inefficiencies, and some may have correlation times greater than those estimated here. However, these make a small number of residues, as less than 2% of residues have a correlation time more than 20ns.

A similar analysis was carried out on all residues for all fragments. A table containing the average correlation times can be found in Table 4-2. Figure 4-6 shows the frequency for various correlation times for all fragments. No correlation was found between the correlation time and either the residue averaged SAP value or its standard deviation. The error between the initial structure and the average is plotted in Figure 4-7. The average error was -0.00039 with a standard deviation of 0.018. The error is normally distributed, with a mean that is not statistically different from zero.

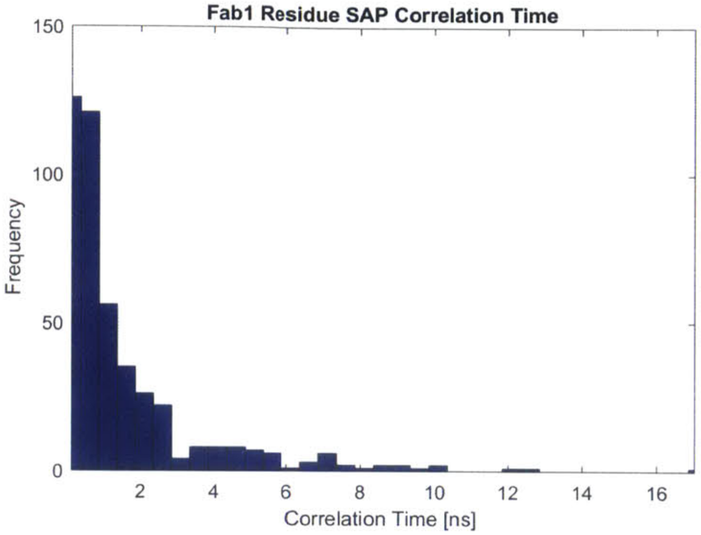

Figure 4-4 - Histogram of correlation times for the residue averaged SAP for all residues in Fab1. SAP Radius 5 Å.

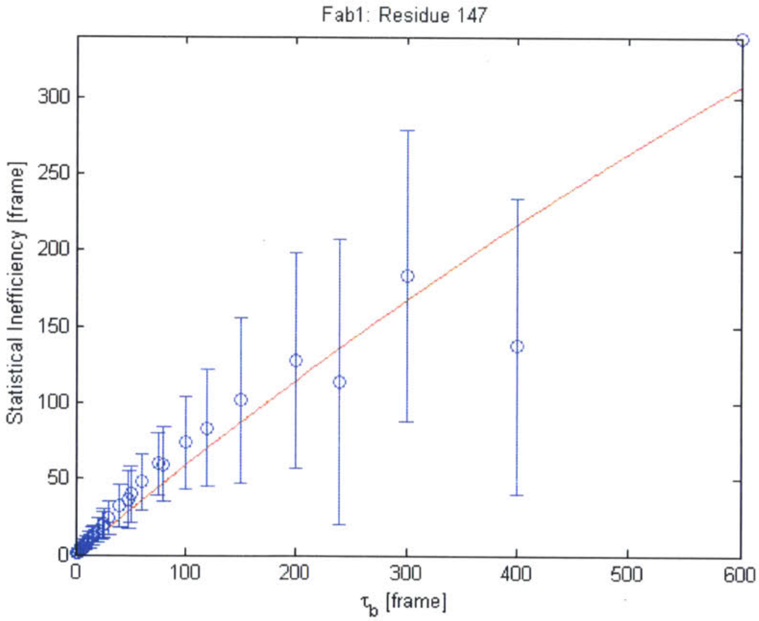

Figure 4-5 - Statistical inefficiency at various time block sizes. Blue dots are measures of the statistical inefficiency for a given time block size, while the blue bars represent plus and minus 1 standard deviation at the given statistical inefficiency. The red line is a present to show the general trend. All data gathered from simulation of Fab1 residue 147 during equilibrated 120 ns of simulation. A frame is 0.1 ns, and SAP Radius 5 Å.

Table 4-2 - Comparison of SAP correlation time for several conditions, the average correlation time of all residues in protein, and the SAP Score.

Fragment	SAP Correlation Time [ns]							
	Average Residue				SAP Score			
	5 Å	7 Å	10 Å	15 Å	5 Å	7 Å	10 Å	15 Å
Fab1	1.55	1.74	1.83	1.96	3.09	4.08	1.67	1.93
Fab2	1.48	1.83	2.25	2.61	3.25	2.46	1.88	1.40
Fab3	1.97	2.25	2.93	3.67	5.08	3.74	3.67	8.42
Fab4	1.39	1.46	1.62	2.00	1.11	1.19	0.84	1.44
Fab5	1.50	1.74	1.83	1.77	1.62	2.28	2.17	3.35
All	1.58	1.80	2.10	2.40	2.83	2.75	2.05	3.31

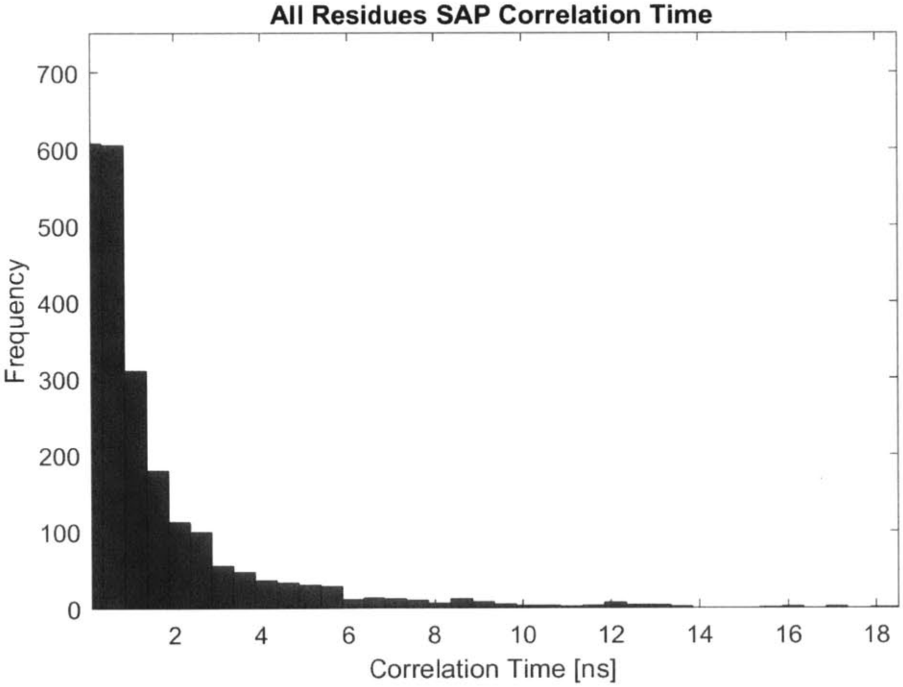

Figure 4-6 – The relative frequency of different correlation times for any residue in all five fragments. SAP Radius 5 Å

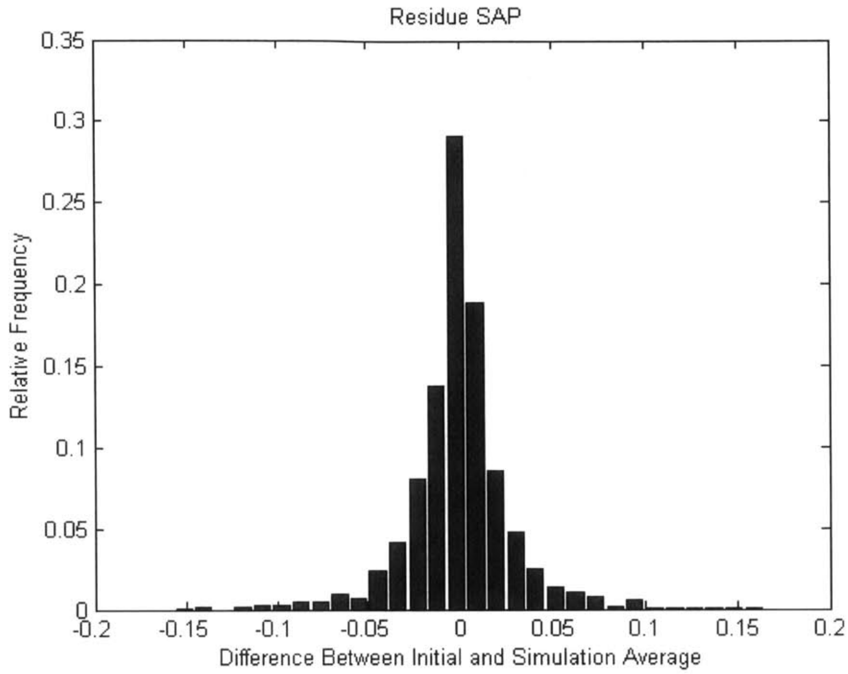

Figure 4-7 – Relative frequency of the difference between the initial and simulation average for the average SAP per residue

4.4. Discussion

When comparing the difference between SAP Score calculated by the initial structure and the average SAP Score, two factors can be compared: the numeric values of the SAP Score, and the resulting ranking of Fabs. The numeric value of the SAP Score, at 5 Å, based on the initial structure differs by up to 8 compared to the simulation average. This difference is approximately equal to the standard deviation of the SAP Score during the simulation. While this is a variation of only 4%, this could greatly impact the classification of the mAb. This difference is about 20% of the original range of values seen in the DI paper (*62*). Meaning that there can be substantial error associated with the use of the initial structure only when evaluating the SAP Score, especially when being used to differentiate highly similar proteins. However, it does give a rough estimate of the ordering.

While the numeric values changed considerably, it only changed the ordering of one fragment, Fab5, when ordering the Fabs based on the SAP Score at 5 Å. Based only on the initial structure the order is: Fab1 > Fab3 > Fab5 > Fab2 > Fab4. When the ordering is based on the simulation average it is: Fab1 > Fab3 > Fab2 > Fab4 > Fab5. The initial structure missed one Fab, Fab5, which the initial structure suggests has average stability but it is actually the most stable fragment. A similar effect can be seen when looking at SAP Scores computed at different radii. In general, the difference between initial value and the simulation value is approximately one standard deviation, calculated from the MD simulation. Additionally, the sign of the difference between the simulation average and the initial value varies. These two factors suggest that the error in using the initial value of the SAP Score is proportional to the observed standard deviation of the SAP Score during the simulation. This also suggests that the crystal structure is representative of the ensemble of solution protein structures. This is most likely partially due to the quality of the available homology model, in this case the crystal structure for Fab1 is known, and all other Fabs are highly similar to Fab1 (similarity > 95% between the different variants). In cases such as this, where all of the Fabs or mAbs are similar, a long MD would be required to accurately rank the proteins, however a rough ranking can be done relatively quickly and cheaply.

This also holds for the residue averaged SAP. In this case, the correlation times sample a much wider range of values compared to those seen in the fragments alone. This is likely due to the increased number of samples; but the trend continues to hold. The difference between the simulation average and the initial structure is most often within one standard deviation calculated over the simulation for that residue. A graph of the difference of the simulation average and the initial structure can be seen in Figure 4-7. The graph appears to be relatively normal. The mean, while not exactly zero, is not statistically different from zero, and the standard deviation of the errors is approximately equal to the standard deviation of the SAP Score measured during the simulation (0.028 and 0.021 respectively). The relatively large magnitude of this error can be an issue when selecting mutation sites. In particular, in cases when most of the values are very similar, and the cost of experiments greatly restricts the number of viable candidates that can be tested. In these cases, extensive MD simulations, possibly measured in the hundreds of nanoseconds, could be required to quantify the difference between two similar appearing residues.

One of the largest differences between the SAP Score and the residue averaged SAP value is the wider range of correlation times seen in the residue averaged SAP. In the case of the residue averaged SAP value, the correlation times range widely, from less than 1 ns to more than 30 ns, see Figure 4-6. While the average is between 2 ns and 4 ns, depending on the averaging radius, most values are small. Approximately 80% of all values are less than 2.5 ns, and only 2% have a correlation time more than 10ns. However, several of these residues, as seen in Figure 4-5, have not converged to the final value, so the presented values are only a lower bound estimate for the correlation times. No correlation was found between either the magnitude of the simulation average of residue averaged SAP, the standard deviation of the residues' averaged SAP, or its correlation time.

4.5. Conclusion

The utility of MD simulations to evaluate SAP is highly dependent on its intended use, and the desired accuracy. If a rough ranking of mAbs is desired, then no simulation may be required, assuming a reasonable quality homology or crystal model is available. However, if the desire is to accurately rank two mAbs with a high sequence identity, then a long MD simulation would be needed, because SAP is correlated for a fairly long period of time, approximately 3ns, and the high value of the standard deviation of the SAP Score during the simulation. Similarly, the error in using the initial structure alone for the selection mutation sites can be problematic, especially if no very residue averaged SAP values are available or if only a couple experiments are possible.

5. RATIONAL DESIGN OF THERAPEUTIC MABS THROUGH INTRODUCTION OF GLYCOSYLIZATION SITES TO DECREASE PROTEIN AGGREGATION

5.1. Introduction

Monoclonal antibodies (mAbs) are the fastest growing area of biotherapeutics with an average yearly market growth rate of 38% in 2012 (***73***). Their use as therapeutic agents has generated unprecedented interest as these molecules can specifically target virtually any molecules implicated in disease. mAbs are often required to be formulated in a very high dosage (often over a hundred mg per mL for subcutaneous injection) and are often manufactured and stored for extended periods of time in a liquid form. At these extreme conditions, proteins stability becomes a greater concern and challenge. Aggregation is one of the most prominent forms of antibody instability and can cause issues from manufacturing failure (***74***) to fatal immunological responses (***75***,***9***) upon drug delivery through loss of efficacy. Protein degradation is usually dealt with through appropriate manufacturing, formulation and storage conditions of the drugs (***45***,***76***). These strategies, though effective, are costly and time consuming and can lead to an inferior product. Another approach is to alter the protein itself, typically by conjugation with a small molecule stabilizer (***77***,***78***) or through substitutions of amino acids in aggregation prone regions (***79***). Molecular-based and computational approaches for the rational design of these proteins, beyond trial and error, permit the determination of protein developability at an early stage (***12***). The incorporation of drug developability within the discovery phase would overall reduce the risk, time, and cost to launch drugs on the market.

Protein aggregation is a complex phenomenon with no single established mechanism. Several states of proteins (folded, partially unfolded, unfolded) can be involved in the aggregation of monomers into small multimers and then into larger oligomeric structures in a reversible or irreversible manner (***75***). Aggregation of macromolecules often involves the dynamic exposure of aggregation prone regions (APRs), hydrophobic patches buried within the folded state of the protein, but it can also occur through interactions of the APRs displayed on the surface of the proteins. This duality renders the prediction of APRs for stable protein engineering that much more complicated and challenging. A handful of computational tools have been validated for the identification of APRs on antibody molecules (***12***). Sequence based computational tools (TANGO and PAGE) allow the identification of several APRs in mAbs including buried ones (***80***), but even though these tools are useful they provide relatively low accuracy and coverage (***12***). The spatial aggregation propensity (SAP) tool (***16***,***13***,***17***,***19***,***18***) has been developed based on molecular dynamic simulations of protein structure and has proven to be efficient in determining potential APRs on the surface of mAbs. Unlike other methods, the SAP tool takes into account protein dynamic fluctuations and the spatial clustering of residues to identify the hydrophobic dynamically exposed residues on the protein surface and define APRs.

With the purpose of improving the stability of mAbs without a loss in efficacy, the SAP tool has been applied to identify the aggregation prone regions on the surface of the fragment antigen binding (Fab) domain of a model IgG1, the therapeutic antibody bevacizumab (Avastin® Genentech). Bevacizumab is an anti-VEGF-A (vascular endothelial growth factor) recombinant humanized monoclonal IgG1 used in the treatment of several cancers (***81***,***82***) as well as against age-related macular degeneration. (***83***) This mAb is an interesting target to rationally design biobetters with enhanced stability as bevacizumab is particularly unstable with respect to aggregation. Bevacizumab is formulated at a low concentration and has previously been shown to be highly aggregation prone (***84***,***85***), leading to not only a substantial loss of activity (***86***) but also to large aggregates, which could potentially be harmful to patients.

Previous work has shown that the degree of glycosylation of biologics affects both their biological and biophysical properties. In particular, it has been shown that N-glycans in antibodies have an impact on the conformational and colloidal stability of mAbs, protecting the protein from both thermal and chemical denaturation. Additionally, glycosylation can stabilize the tertiary and quaternary structures of mAbs, and in the case of the Fc domain, hydrophobic regions, which are aggregation prone, are covered by a glycosylation moiety (***87***,***88***,***89***,***90***). Glycoengineering has proved to be effective in substantially increasing the solubility of biotherapeutics (***91***,***92***) as well as reducing the aggregation propensity of mAbs. It is well accepted that N-linked carbohydrates participated in the stabilization of mAbs against aggregation by covering aggregation prone motifs and through steric hindrance that disrupts intermolecular interactions (***93***,***78***). This strategy was investigated here to stabilize bevacizumab against aggregation. Four glycosylation sites were independently introduced on the surface of the constant region of the Fab domain of bevacizumab by single point mutation in the CH_1 and CL domains, which are far from the binding region.

5.2. Materials and methods

5.2.1. Molecular Simulation

Several Fab MD simulations were performed, one of the wild-type of bevacizumab and one of each of the hyperglycosylated Fab variants (L118N, Q160N, Q160s and E195N). All three simulations were based on the crystal structure of bevacizumab, which was obtained from the RCSB PDB (PDB ID: 1BJ1) (***94***). Hydrogen atoms were added to this structure at pH 7 using the PSFGEN plugin of the VMD (***95***). Topology and structure files for the hyperglycosylated Fab variants were generated using the Glycam website (***96***), assuming a G0 glycosylation pattern (GLYCAM notation: DGlcpNAcb1-2DManpa1-6[DGlcpNAcb1-2DManpa1-3]DManpb1-4DGlcpNAcb1-4DGlcpNAcb1). For all simulations, the AMBER12SB (***97***) and Glycam06 (***98***) force fields were used for the protein and glycosylation, respectively. A molecular-dynamics (MD) simulation was performed on each of these all-atom structures of the Fab domain with an explicit TIP3P water model (***32***). Each Fab domain was solvated in a cubic box with periodic boundary conditions in all three directions. The dimensions of the water box were adjusted such that the surface of the Fab domain is at least 10 Å away from any side of the box. The solvated system was made charge neutral by adding chlorine ions. The system temperature and the pressure were maintained at 300 K and 1 atm, respectively, by the Berendsen coupling scheme (***99***). The GROMACS (***71***) package was used to perform the MD simulations in the NPT ensemble. The systems were initially minimized and then equilibrated for 20 ns. Production runs of 80 ns were then performed and frames were extracted at every 0.1 ns for further analysis. The average spatial aggregation propensity (SAP) values of each residue at 5 Å and 10 Å were computed (***13***) over the 80 ns of the MD production runs. The average effective hydrophobicity (Φ_{eff}) value of each residue (***100***) was also computed over these MD trajectories.

5.2.2. Identification of glycosylation sites to engineer

To identify potential glycosylation sites in the C_H1 and C_L domains of bevacizumab, we identified all of the high SAP residues in these domains as described above. Using the last frame from the MD simulation, we then identified all Serine, Threonine and Asparagine residues that are within 10 Å of these high-SAP residues (the distance is the minimum distance between all atom-pairs of two residues) and belong to the CH1 and CL domains. For all of the selected S/T/N residues, we chose a neighboring residue (in the sequence) for mutation to generate an NXS or NXT glycosylation motif. We rejected all of the variants in which a GLY or PRO needs to be mutated because mutation of these residues may cause a signification perturbation in the structure of the domain. Furthermore, we also rejected all of the mutations that lead to an N-P-S or N-P-T motif as this motif does not undergo N-glycosylation. To be efficiently glycosylated, the side-chain of the ASN residue should be surface-exposed. Therefore, we rejected all of the mutations where the exposed surface area of the side-chain atoms (of the ASN or the residue to be mutated to ASN) is less than 15 Å^2. All residues oriented on a different face relative to the high SAP value residues to be masked were dismissed as well. Lastly, we rejected all of the mutations of high SAP residues. In order to explicitly show the impact of coverage of aggregation prone regions by the glycan, only those variants in which did not mutate high SAP residues were considered. In this work, we generated three such variants: E195N, L118N and Q160N/S to test our hypothesis of whether the introduction of a glycosylation site near a high SAP region leads to an overall reduction in the aggregation propensity of the antibody.

5.2.3. Cloning, Generation of variants, Expression, and Purification of mAbs

The bevacizumab genes, synthesized by Genscript (Piscataway, NJ), were codon-optimized for expression in mammalian cells and subsequently subcloned separately into the vector gWiz (Genlantis, Torreyana San Diego) using the Gibson method, resulting in the vectors gWiz-A-LC and gWiz-A-HC. bevacizumab variants were generated by site directed mutagenesis and confirmed by sequencing. Oligonucleotides (IDT, Coralville, Iowa) were designed to introduce single mutations on bevacizumab LC (Q160S, E195N) or on bevacizumab HC (L118N). WT bevacizumab and variants were expressed by transient transfection of FreeStyle 293-F cells (Life Technologies, Grand Island, NY) grown in GIBCO FreeStyle 293 Expression Medium (Life Technologies). Transfections were performed using 0.5 g of each heavy-chain and light-chain mAb-expressing vector and 2 mg of polyethyleneimine (Polysciences, Inc., Warrington, PA) per liter of 10^6 FreeStyle 293-F cells. After five to six days, the supernatant was collected and filtered (0.22 µm) prior to purification. Expressed mAbs (8 to 25 mg) were purified first by affinity chromatography (protein A sepharose from GE Healthcare, Piscataway, NJ), and then, concentrated on centrifugal filter devices AMICON YM30 (EMD Millipore - division of Merck KGaA, Darmstadt, Germany) for further purification by cation exchange. The pure proteins were buffer exchanged into 10 mM histidine (pH 6.0) and further concentrated to the desired concentration, as determined by measuring the absorbance at 280 nm.

5.2.4. Stability of mAbs

The accelerated aggregation studies of bevacizumab and its variants were performed within three days after formulation. A total of 50 mg/mL of protein in 10 mM histidine (pH 6.0) were incubated at 52°C in a Bio-Rad MyCycler Thermal Cycler (Hercules, CA). Aggregation was stopped at several time points by diluting the sample down to 10 mg/mL in cold 15 mM potassium phosphate buffer (pH 6.5), followed by a 10 min incubation in ice. Part of each sample was further diluted to 1 mg/mL for turbidity measurement by assessing the absorbance at 320 nm. The extent of aggregation was measured at 22°C by size exclusion chromatography (SEC-HPLC) performed on an Agilent 1200 LC (Santa Clara, CA) using a Tosoh TSKgel super SW3000 column (Tokyo, Japan). Samples (10 mg/mL) were spun down for 3 min at 6000 rpm to remove large insoluble aggregates prior to injection (5 μL) in the SEC-HPLC. The proteins were eluted with a mobile phase of 150 mM potassium phosphate buffer (pH 6.5) at a flow rate of 0.5 mL/min. Proteins were quantified by detection at 280 nm. Areas of the peaks were integrated at each time point. The ratio of the aggregates peak to the total peak area at each time provides the amount of soluble aggregates. The mass balance with the initial concentration (t=0) allows the estimation of the amount of insoluble aggregates. Each experiment was reproduced at least in duplicate with different batches of proteins. The standard errors reported herein represent the deviation observed during the all mAb stability assessments (production, purification/formulation, accelerated aggregation study).

A propagation of error was applied to calculate standard deviations reported for the stability increase factor and the aggregation rate reduction. Aggregation rate constants were extracted from the fitting of a second order function to the monomer loss measured over time.

Turbidity was estimated by measuring the absorbance at 320 nm of 50 mg/mL samples previously stressed for 48 h at 52°C and diluted down to 1 mg/mL for measurement.

5.3. Results

5.3.1. Rational design of stabilizing hyperglycosylated variants

An alternate approach to increasing the stability of bevacizumab is to introduce glycosylation sites near the identified aggregation prone regions to mask them. To introduce these new sites, residues near each aforementioned aggregation prone region were considered. Residues close to the antigen binding site were not considered to prevent any chance of altering bevacizumab activity. Therefore, we did not investigate the introduction of a carbohydrate moiety masking either the V5 or F50 high SAP value residues. To reduce the aggregation propensity of bevacizumab, we intended to mask with a glycosylation motif the high SAP value residues V110, L154, L180 and L201. Upon mutation in more hydrophilic and charged residues, these residues have been shown to increase the stability of bevacizumab.

To generate a list of potential N-glycosylation sites, i.e., NXS or NXT (X is any amino acid but P), close enough for the sugar group to mask those residues of interest, all of the serine, threonine and asparagines (residue i) within 10 Å of these high SAP value residues within the constant domains of the Fab region of bevacizumab (CH_1 and C_L domains) were identified.

The neighboring residues (residues i-2 and i+2) were then considered as candidates for mutation to generate a foreseeable glycosylation site. All of the residues that can potentially be mutated to generate glycosylation sites are listed in Table 5-1 in the column "Variants to generate glycosylation site". The residue (i) T197 is located at a distance less than 10 Å from L154 and is highly surface exposed. If the residue (i-2) E195 is mutated into an asparagine, it will create the sequence NXT_{197} (instead of EXT), making it a potential glycosylation site (Table 5-1). The carbohydrate introduced *in vivo* during mAb expression in position N195 is expected to mask the L154 residue (and its surrounding residues) and could potentially reduce the aggregation propensity of bevacizumab. A large set of residues were identified and further selective parameters were applied to ensure that feasible and efficient glycosylation would take place. Figure 5-1 summarizes the different criteria which must be satisfied (details in the Materials and Methods section). After dismissing mutations leading to glycosylation site motifs, which would not undergo glycosylation, as well as mutations potentially affecting protein structure, together with high SAP value residues, eight residues were identified for directed site-mutagenesis to create potential glycosylation sites. Four different possible variants have been identified which will introduce a glycosylation site to mask the V110 high SAP value residue, whereas the four other mutations identified should permit the masking of three residues (L154, L180 and L201). With the aim of making our experiments more efficient and ensuring that we chose pertinent glycosylation sites, we did not introduce any glycosylation sites to mask the high SAP value residue V110. The V110 residue has been shown to be involved in aggregation, based on the observation that its mutation into lysine resulted in a 2.8-fold stabilization of bevacizumab against aggregation. Nevertheless, this reduction in aggregation is similar to that observed for variants L154D and L201K, and V110 has one of the lowest SAP scores among the above listed high SAP value residues, making its masking by a carbohydrate moiety less attractive.

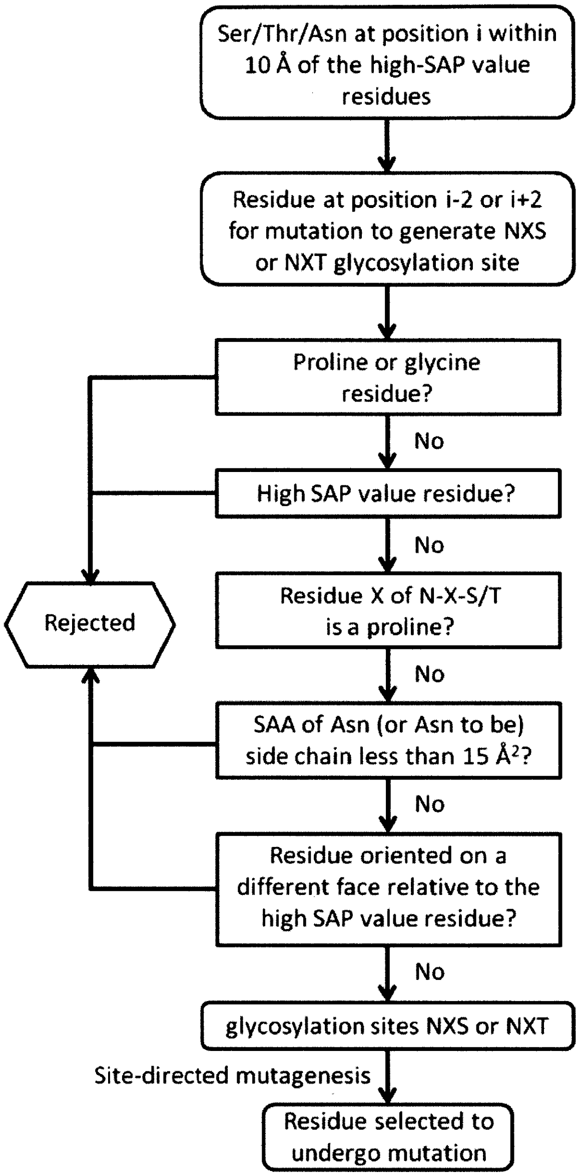
Ser/Thr/Asn at position i within 10 Å of the high-SAP value residues
Residue at position i-2 or i+2 for mutation to generate NXS or NXT glycosylation site
Proline or glycine residue?
No
High SAP value residue?
No
Residue X of N-X-S/T is a proline?
No
SAA of Asn (or Asn to be) side chain less than 15 Å²?
No
Residue oriented on a different face relative to the high SAP value residue?
No
glycosylation sites NXS or NXT
Site-directed mutagenesis
Residue selected to undergo mutation
Rejected

Figure 5-1 - Rational selection of residues for mutation to introduce glycosylation sites on the bevacizumab Fab domain

As a proof of principle, hyperglycosylated variants were generated to cover high SAP residues L154, L180 and L201. Interestingly, there is an overlap of the potential glycosylation sites (Table 5-1). L154 could be masked by a glycosylation motif introduced by the mutations Q160S and E195N, which generate in bevacizumab LC the glycosylation sites NSS_{160} and NVT_{197}, respectively. These two mutations generate a site with glycosylation moieties that should mask not only residue L154 but also residue L201. The high SAP value residue L180 could be masked by a glycosylation motif introduced by either the substitution of the Q160 residue in asparagine generating here the glycosylation motif NES_{162} or by the mutation of the residue L118 in asparagine introducing the glycosylation site **NVT$_{120}$** on the heavy chain of bevacizumab.

Table 5-1 - List of residues for glycosylation site engineering. Identification of variants which are likely to be glycosylated in the vicinity of high SAP regions. Residues in grey were not selected for the reasons described in the footnotes.

High SAP Residues masked	Ser/Thr/Asn within 10 Å of the high-SAP value residues	Variants to generate glycosylation site	SAA of side-chain atoms of Asn or residue to be mutated to Asn (Å^2)
L:L154 (Ø$_{eff}$ = 0.26)	L:N152	L:L154S[2]	106.8
	L:S156	L:L154N[2]	84.2
	L:N158	**L:Q160S**	53.22
	L:S159	L:G157N[1, 4]	12.5
	L:S177	L:L175N[2, 4]	7.6
	L:T197	**L:E195N**	33.9
	L:T206	L:P204N[1]	56.9
L:V110 (Ø$_{eff}$ = 0.18)	L:T109	L:K107N	107.7
	L:N138	L:Y140S	67.8
	L:S171	L:K169N	160.1
	L:T172	L:D170N	57.3
	L:S202	L:G199N[1, 4]	12.3
L:L201 (Ø$_{eff}$ = 0.27)	L:S114	L:A112N[3]	39.2
	L:N137	L:F139S[4]	8.4
	L:T197	**L:E195N**	33.9
	L:S202	L:G199N[1, 4]	12.3
	L:S203	L:L201N[2]	85.2
	L:N158	**L:Q160S**	53.2
	L:S159	L:G157N[1, 4]	12.5
H:L180 (Ø$_{eff}$ = 0.28)	L:S162	**L:Q160N**	28.5
	L:T178	L:S176N[4]	1.56
	L:T180	L:T178N[4]	6.6
	H:T120	**H:L118N**	28.2
	H:S122	H:T120N[4]	10.3
	H:S182	H:L180N[2]	87.7
	H:S183	H:Q181N[4]	12.0
	H:S187	H:L185N[5]	34.8

Reasons for rejections:

1. Residue to be mutated is either a glycine or a proline.
2. Residue to be mutated is a high SAP value residue. Only the effect of masking the aggregation prone region was to be investigated, therefore, potential aggregation prone residues must not be mutated.
3. Residue X of N-X-S/T is a proline. NPS/T are not substrates for glycosylases.
4. SAA of the side chain of asparagine or of the residue to be mutated in asparagines is less than 15 Å^2 and not accessible to glycosylases.
5. NXS/T motif to be generated is oriented on a different face relative to the high SAP value residue.

5.3.2. Glycoengineered proteins for increased stability against aggregation

The four variants L118N, E195N and Q160N/S (like the WT and reduced SAP variants) were produced in HEK293 human embryonic kidney cells, which are able to carry out the original post-translational modifications and should produce mAbs with glycosylation at the engineered sites. To test whether the introduction of an N-glycosylation site near a high SAP region led to an overall reduction in the aggregation propensity of bevacizumab, the hyperglycosylated variants were expressed, purified and characterized by DSC, turbidity evaluation and SEC-HPLC. The incorporation of N-glycan on the Fab domain of our bevacizumab variants was verified by reducing SDS-PAGE. The HC of L118N and the LC of E195N and Q160N/S have clearly higher molecular weights than those of the WT, corresponding to the addition of an N-glycan (data not shown). All variants were tested by DSC for their thermal stability. Three transition temperatures were extracted from the obtained thermograms (Table 5-2). The introduction of the N-glycan motif on the Fab domain of bevacizumab does not affect the melting transition of the CH3 domain (Tm_3), which varied by less than 0.7°C. The Fab domain presents a higher thermal stability when hyperglycosylated at one of the four positions tested in this work (Tm_2 increased by 1.6°C to 2.3°C). Surprisingly, the melting temperature Tm_1 attributed to the CH2 domain[45] is also affected. Glycosylation of the sites N195, N160 and N158 increased the CH2 domain stability by 1.6°C to 2.1°C, whereas glycosylation of N118 reduced the CH2 domain thermal stability by 2.3°C.

Table 5-2 – Bevacizumab stabilization by hyperglycosylation of the Fab domain. The amount of monomer and soluble aggregates detectable by SEC-HPLC was measured at various time points up to 48 h of incubation at 52°C of 50 mg/mL (His 10 mM, pH 6.0) mAbs. The monomer percentage of the WT and variants in the soluble fraction after 48 h heat is reported in the table. Data are the mean ± SD (n=3 experiments with three different protein batches, *n=2 experiments with two different protein batches). The kinetic data were fitted to a simple equation to extract a second order rate constant ("Aggregation rate"). The monomer percentage at 48 h for each variant was compared to the WT through the "Fold increase stability". For each hyperglycosylated variant, a MD simulation was performed and a SAP score (the sum of all positive SAP values) at R=5 Å was computed in addition to the standard errors. The melting transition temperatures in degrees Celsius

for the WT and each hyperglycosylated variant were obtained by fitting three Gaussians to each thermogram.

Variants	% monomer at 48 h	Aggregation rate *10^$^{-2}$ (mol^{-1}.L.min^{-1})	Fold increase stability	SAP Score	Melting temperature (°C)			K$_D$ (nM)
					Tm$_1$	Tm$_2$	Tm$_3$	
WT	68 ± 2	31.3 ± 5.6	1.0 ± 0.1	164 ± 1.04	70.4	71.8	82.4	0.85 ± 0.28
L118N	85 ± 3	9.4 ± 5.5	2.2 ± 0.5	149 ± 0.41	68.1	74.1	83.1	0.45 ± 0.06
Q160N	79 ± 4	14.8 ± 8.7	1.5 ± 0.3	156 ± 1.21	72.5	73.4	82.9	1.56 ± 0.13
Q160S	**90** ± 3	4.2 ± 0.9	**3.2** ± 1.1	159 ± 1.05	72.3	73.4	83.0	1.72 ± 0.15
E195N	89 ± 4	5.5 ± 2.9	2.9 ± 1.2	155 ± 0.66	72.0	73.4	82.5	0.59 ± 0.04

Hyperglycosylated variants were tested for their stability through accelerated aggregation studies at an elevated temperature. As with the reduced SAP variants, 52°C was chosen as the temperature to induce aggregation, and monomer concentrations were monitored over a 48-h period by SEC-HPLC (Figure 5-2). Our SEC-HPLC data are the measure of the monomer concentration at various time points, and Figure 5-2 represents the average of three independent experiments (i.e., three different protein production batches) unless stated otherwise. We observed the presence of soluble aggregates in all samples from the beginning of the experiment, which was certainly due to the high concentration formulation (50 mg/mL) that was close to the solubility limit of bevacizumab. Our wild-type bevacizumab sample contained 8.5% soluble aggregates at t=0, whereas the variants Q160N, Q160S and E195N contained less than 4% soluble aggregates at t=0. After 48 h incubation at 52°C, we observed 32% soluble aggregates for the WT, whereas our best variant displayed less than 10% aggregates, representing over a 3-fold increase in stabilization. The four hyperglycosylated variants are all more stable than the WT bevacizumab against heat-induced aggregation, having a 1.5 to 3.2-fold increase in stability (Table 5-2). A second order rate constant was extracted from the fitting of the monomer loss measured over time for each variant. L118N, Q160S and E195N are the three variants with the most stabilizing effect, and their aggregation rate was reduced by a factor 3.3 to 7.4 (Table 5-2 and Figure 5-2).

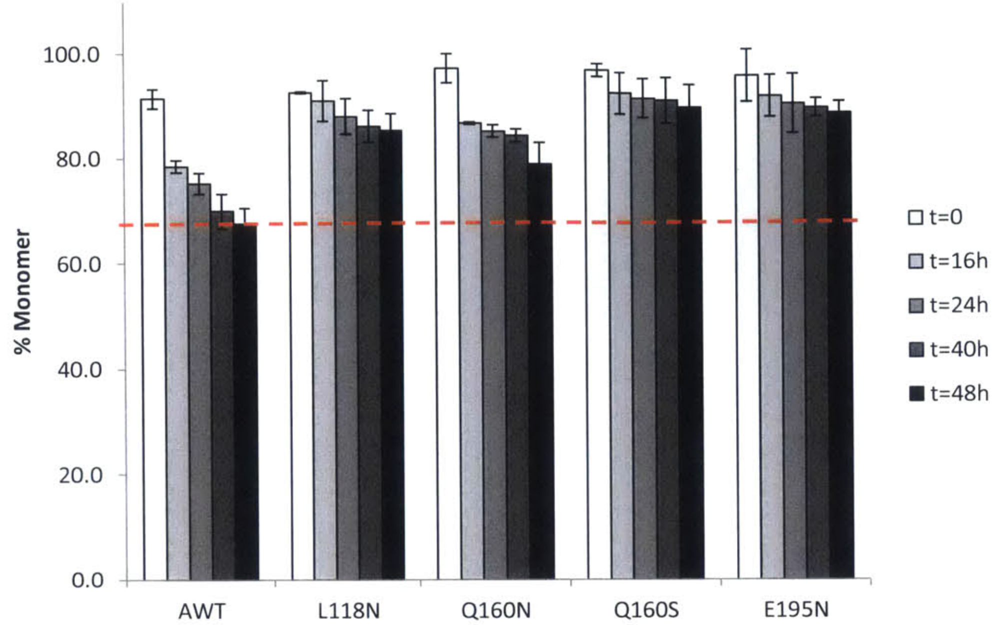

Figure 5-2 - Stability comparison of WT bevacizumab and variants by SEC-HPLC. Monomer loss for WT and hyperglycosylated variants (50 mg/mL in histidine buffer, pH 6.0) was measured at various time points upon heat stress at 52°C for 48 h. Data are the mean ± SD. (n=3 experiments with three different protein batches * n=2 experiments with two different protein batches).

Using the SEC-HPLC data and by comparing the amount of monomer measured at each time point (t=16 h, t=24 h, t=48 h) to the amount of monomer at t=0, one can estimate the amount of insoluble aggregates which did not enter the separation column. Our best hyperglycosylated variants, Q160N/S and E195N exhibited reduced insoluble aggregates (3-5%) compared to the WT (~8%), whereas L118N exhibited nearly double the amount of insoluble aggregates (15%) clearly making the variant L118N the least effective hyperglycosylated variant overall (1.3-fold stabilization). This result was confirmed by turbidity measurements: Abs_{320nm}=0.014 to 0.052 for our three best variants, Abs_{320nm}=0.216 for L118N versus Abs_{320nm}=0.091 for WT.

As described above, the engineered mAbs must have at least the same affinity to the target as the WT. The efficacy of our hyperglycosylated engineered bevacizumab was estimated *in vitro* via a competitive ELISA developed in-house and allowing measurement of the affinity of the WT and variants for the antigen, VEGF. Figure 2B shows the binding curve of WT bevacizumab and the hyperglycosylated variants to VEGF-A. Our mAbs were preincubated with VEGF, which were then incubated with anti-VEGF IgG1 for further detection by ELISA. A low absorbance at 450 nm indicates a low amount of VEGF bound to the ELISA plate and a high amount of our mAb bound to VEGF. The corresponding dissociation constants (K_D) are reported in Table 5-2. Our competitive assay results show that all of the mutations introduced in the Fab fragment, far from the CDR, produce hyperglycosylated variants which bind to VEGF with the same affinity as the commercial drug, our WT bevacizumab. When glycosylation site locations are chosen carefully, N-glycans can stabilize mAbs without compromising their activity *in vitro*.

To better understand the effect of glycosylating the Fab domain of bevacizumab on aggregation, we performed molecular dynamics simulations of the Fab domains of the WT and the four hyperglycosylated variants. Although the nature of the glycan structures (32 different glycoforms) is of high importance, the heterogeneity and control of glycosylation pattern are complex topics (***101***) and the identification of the N-glycosylation modification was not investigated in this study for our hyperglycosylated variants. The glycosylation of mAbs is highly dependent on the culture conditions (***102***); therefore, we assumed here the same glycosylation pattern observed previously in our laboratory for mAbs produced in the same conditions as our hyperglycosylated variants (***93***). Molecular dynamic simulations were performed based on the crystal structure of the bevacizumab Fab domain (1BJ1) and assuming a G0 glycosylation pattern. Figure 5-3 shows snap shots of typical simulation results obtained for our four variants. In these simulations, Q160N bears a glycan moiety covering a large surface of the Fab domain, resulting in a biobetter with a reduced SAP score (Table 5-2). However, L118N, Q160S and E195N exhibit a carbohydrate moiety pointing away from the surface of the Fab rather than covering it. The carbohydrate of these three variants are present in various orientations (Figure 5-3), some unexpected, but all lead to a substantial reduction in the SAP score of the Fab domain (Table 5-2).

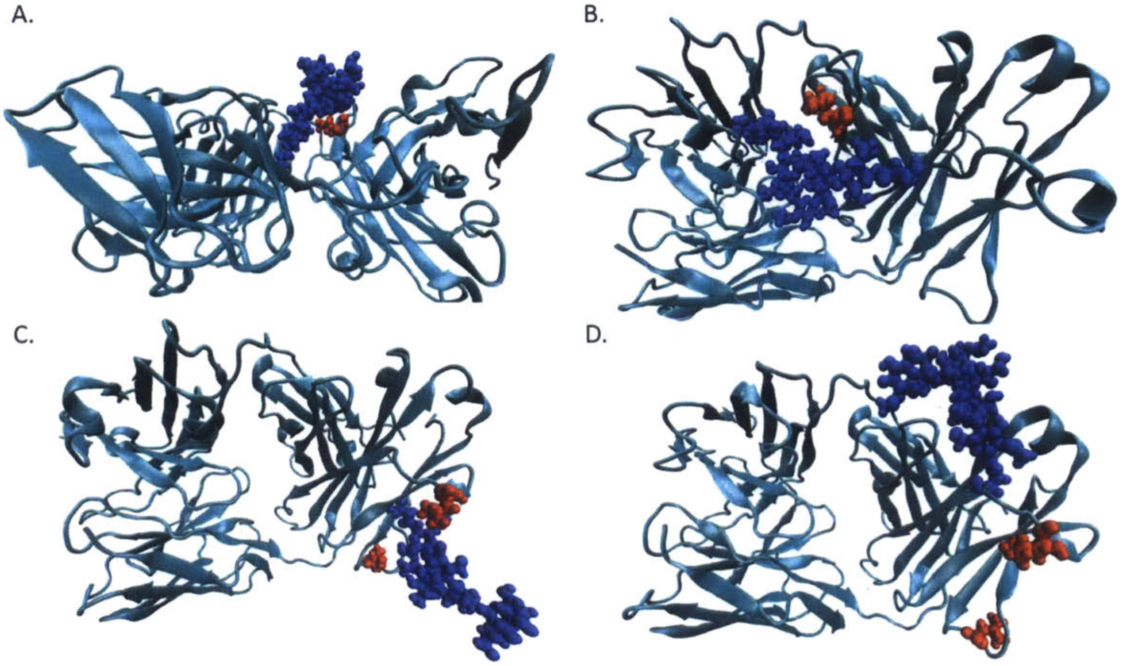

Figure 5-3 - Representative structures from the MD for the four MD simulations of the hyperglycosylated variants. In all cases, the left of the image is the CDR and the right side is the hinge region. The protein structure is shown in teal, the glycosylation moiety is shown in blue, and the residues to be masked by the glycosylation motif are shown in red. (A) L118N hyperglycosylated variant designed to cover L180. (B) Q160N hyperglycosylated variant designed to cover L180. (C) E195N hyperglycosylated variant designed to cover L154 and L201. (D) Q160S hyperglycosylated variant designed to cover L154 and L201.

5.4. Discussion

In the work presented here, we successfully engineered biobetters for the therapeutic monoclonal antibody bevacizumab and designed mAbs with substantially reduced aggregation propensity while maintaining high affinity to the target antigen. This was achieved by masking the high SAP value residues with a carbohydrate motif to prevent protein-protein interactions. The nature and degree of glycosylation of most biologics play an important role during the design and development of biologics. It has been well established that oligosaccharide moieties play an important role in therapeutic biological activity, effector functions such as antibody-dependent cellular cytotoxicity and complement-dependent cytotoxicity, immunogenicity, serum half-life, and clearance (***103***). In addition to those biological attributes, carbohydrate groups participate in the integrity of the drug by reducing aggregation propensity, increasing solubility, stabilizing the native conformation, protecting against various degradation pathways (hydrolysis, oxidation) and overall stabilizing the macromolecules (***91***,***103***). This functionality is particularly relevant for antibodies as all IgGs are naturally N-glycosylated at position N297 in each of the CH2 chains of the Fc domain. It has been demonstrated on several occasions that this N-glycan participates in the stabilization of mAbs against aggregation caused by various stresses (***78,87,89,90,104,105***). Natural IgGs present in human serum are all glycosylated in the Fc domain whereas only less than a third are glycosylated in the Fab domain (***103***). N-glycans are found attached to the variable region of the LC, the HC or both. It is estimated that approximately 20% of the variable region of mAbs bear an N-glycosylation site motif (***106***). These Fab oligosaccharides are believed to affect mAb functions in a different way than the sugar moiety at N297 in the Fc domain and could potentially have an impact on antigen binding depending on the location of the glycosylation motif on the surface of the variable domain (***101***,***107***). The natural occurrence of Fab glycosylation supports the idea that a glycoengineered antibody might be viable as a biotherapeutic as well (***108***). For example, cetuximab, a chimeric therapeutic mAb bears oligosaccharides in its VH region (***109***), and Fab domain glycoengineering has already proved to be successful for solubility improvement (***106***) and aggregation prevention (***92***).

In this study, we took advantage of the capacity of N-glycans to mask APRs to reduce the aggregation of our model therapeutic mAb bevacizumab. The N-glycosylation sites NXS/T were carefully chosen to fit a series of criteria, in particular to be introduced by a single point mutation. In addition, the sites were chosen so that the N-glycans would not interfere with bevacizumab antigen binding while masking high SAP residues shown to be involved in aggregation. Four glycosylation sites were independently engineered: L118N, Q160N, Q160S and E195N. The introduction of an N-glycan at position N160 in the Fab domain, which masked the high SAP residue L180, appears to increase the solubility and to slow down the aggregation propensity (2.1-fold reduction in aggregation rate). The L180 high SAP value residue can also potentially be masked by another carbohydrate motif, the L118N mutation. Introducing an N-glycan at this position actually reduces the aggregation rate by 3.3-fold.

Two leucine residues of high SAP value, L154 and L201, are on the same face of the Fab domain of bevacizumab, toward the hinge region. Interestingly, two residues have been identified as candidates for mutation to generate glycosylation substrate sites which could mask both L154 and L201: Q160S, generating the motif NSS_{160} and E195N introducing the NVT_{197} glycosylation site (Table 5-1). Both of the hyperglycosylated variants present the highest degree of stabilization (3-fold) and the slowest rate of aggregation (5 to 7-fold decrease), with only approximately 10% soluble aggregates detected after 48 h of incubation at 52°C versus 32% for the WT. The correlation observed between the stability for the hyperglycosylated variants and the reduced SAP mutants that were predicted to be covered suggests that the N-glycans introduced on the Fab surface actually did mask the high SAP residue identified and attributed to be masked. Nevertheless, steric hindrance preventing protein-protein interactions cannot be dismissed and may be part of the protective effect of the N-glycans.

The N-glycans could be shielding the aggregation prone residues (L154, L180, L201) impacting initial aggregate formation by disrupting hydrophobic interactions. Furthermore, one should consider that the N-glycans could also increase the colloidal stability of bevacizumab by crowding or steric hindrance (***106***). The added glycan is a large moiety that could prevent another molecule from interacting with the nearby residues, thus reducing their role in aggregation. The simulation of the hyperglycosylated Fab domain shows a difference in the behavior of the added carbohydrate moieties, whereas the structure of the Fab domain is only minimally perturbed by the addition of the glycans, as the root mean squared displacements (RMSD) of the proteins to the wild-type are comparable to that of the wild-type during its simulation. The masking of hydrophobic residues was the original goal for the addition of glycosylation sites, and it is present in all of the simulations. While masking is present in all simulations, it is present to varying degrees, in particular when considering the masking of the targeted residues. Typically, L118N and Q160N were engineered to introduce a carbohydrate that masked L180, and both variants have a comparable effect on bevacizumab stability. Interestingly, the carbohydrates behave very differently. On one hand, the Q160N carbohydrate consistently covers a wide region of residues through its interaction with the Fab surface, but at the cost of masking L180, with a negligible effect on its SAP value. On the other hand, the L118N carbohydrate completely masked L180, whose SAP value dropped dramatically (decreased to a fifth of its original value) and made one of the most hydrophobic regions nearly hydrophilic with a decrease in Fab SAP score of approximately 15 units. Interestingly, the L118N carbohydrate was forced to point away from the surface of the Fab domain and does not reliably cover many other residues other than L180. In fact, many of the terminal sugars spent a significant amount of time being fully solvated, suggesting that the L118N carbohydrate might acts as an exclude solute or stabilize bevacizumab through steric hindrance (Figure 5-3A). The same orientation of the carbohydrate was observed for the variant E195N, which does not interact with the Fab surface. The glycan spends the vast majority of the simulation in solution. While it does not mask the high SAP residues L154 and L201, it could prevent large molecules from interacting with these APRs. While there was a reduction in the SAP score of the E195N variant in comparison to the wild-type, it was not a large enough reduction to account for the drastic reduction in aggregation propensity. Its rate constant is a third of the rate of Q160N, which has a comparable SAP score but is more aggregation prone. The E195N variant is also more stable than L118N, which presented the lowest SAP score. These results point to the possibility that crowding plays a major role in the stability of the E195N mAb. The variant Q160S bears a carbohydrate interacting partially with the Fab domain surface and pointing away from it as well. In this case, the carbohydrate does not cover the neighboring high SAP residues L154 and L201, thereby reducing the SAP score to a smaller extent. The reduced SAP score of the hyperglycosylated variants compared to the WT correlates well with the change in the melting temperature of the Fab domain. This outcome supports the idea that reducing the hydrophobicity of the Fab surface increases its conformational stability, which can be predicted through the SAP Score. However, this stabilization is not the only factor affecting the protein aggregation propensity. The best biobetters against aggregation do not have the lowest SAP score, as seen for Q160S and E195N variants. This finding supports the idea that masking the high SAP residues also participates in the stabilization of the mAb, but the colloidal stabilization through a steric effect may be a major contributor. The hyperglycosylated variants subject to a slower aggregation process are conformationally more stable via the masking of hydrophobic patches and present increased colloidal stability as well due to the steric hindrance caused by the large solvated glycosylation motif.

While the glycosylation sites were selected such that high SAP residues were not mutated, to help distinguish the effect of high SAP mutations and hyperglycosylated variants, any mutation will impact the SAP profile of the protein. In some cases, such as Q160N, this change is minimal, as glutamine and asparagine are similar residues. However, in the case of L118N, the change in hydrophobicity is more substantial and the mutation accounts for half of the change of SAP score compare to that of the wild-type. Therefore, when possible, choosing a glycosylation site of high SAP value would greatly aid in reducing the SAP score of the all Fab domain and should be added to the guideline for selecting glycosylation sites. Furthermore, the type of the carbohydrate could also play a role in destabilizing initial aggregate formation by modifying the net charge of the Fab domain and impacting the overall SAP score of the mAb domain.

Nonetheless, it is important to note that the nature of the glycoforms, both in the Fc and Fab domains, was not investigated here. The nature of the glycoforms would need to be controlled to ensure molecular homogeneity and limit undesirable immunological responses (***110***). In particular, it has been reported that the mAb Fab domains are subject to a higher degree of glycosylation than the Fc domain and with an extremely diverse composition of glycoforms (***108***). This finding is probably due to a higher accessibility of the glycosylation sites exposed on the Fab surface compared to the N297 residue buried at the interface of the two CH2 domains. Combining methods to control the nature and homogeneity of glycoforms with a single point mutation to insert an N-Glycan on the Fab domain is an efficient general approach to generate biotherapeutics and biobetters to enhance the biophysical properties as well as chemical stability. The carbohydrates potentially protect the shield domain from degradation pathways such as oxidation and hydrolysis. Glycosylation sites must be cautiously chosen for each individual mAb, and clinical efficacy and immunogenicity need to be carefully investigated upon mAb modification.

5.5. Conclusion

Overall, the masking of APRs raised the stability of bevacizumab to the same level of the formulated drug. It is, therefore, likely that an appropriate formulation would further increase the stability of our biobetters. The masking and crowding of APRs by glycosylation motifs could potentially present further advantages over the removal of aggregation patches. The added carbohydrate moiety could also stabilize the mAbs against other degradation routes such as hydrolysis, oxidation or deamidation. Even though experimental high throughput screening methods are being developed to identify developable mAbs, computational predictive approaches remain attractive and competitive due to their low cost and no requirement for materials. There is a clear incentive for the development of new in silico platforms for high throughput screening of the developability and aggregation propensity of proteins with high accuracy, such as the sequence-based statistical model used in Lonza's aggregation prediction tool. Their implementation early on during the discovery phase allows the reduction of costs, easier manufacturing, and the formulation of higher concentrations, opening the door for new delivery routes or reduced dose administration, all to the advantage of the patients and practitioners, with the benefits of a potentially safer drug and lower treatment costs.

6. MOLECULAR INVESTIGATION INTO THE MECHANISM OF NON-ENZYMATIC HYDROLYSIS OF PROTEINS AND PREDICTIVE ALGORITHM FOR SUSCEPTIBILITY

6.1. Introduction

Under normal storage conditions, the rate of non-enzymatic hydrolysis of a typical amide bond in biopharmaceutical products is extremely slow; the half-life of the amide bond between two amino acids can be more than a hundred years (***111***). However, there are instances in which the amide bond reacts much faster. In some cases the half-life can be as low as eleven minutes (***112***). This presents a challenge for the pharmaceutical industry, as hydrolysis can greatly reduce the drug's activity and therein efficacy, and it can increase the potential of immunogenic effects (***5***,***6***,***113***). As is the case with many degradation routes, the loss in efficacy can be through both direct and indirect routes (***5***,***6***). In the case of IgG1 monoclonal antibodies (mAbs), hydrolysis cleaves the amide bonds between residues within the hinge region directly leading to the creation of undesirable protein fragments (***114***). In addition to the direct creation of said byproduct, the hydrolysis of a amide bond can also accelerate many of the other degradation routes, as was the case for an IgG2 antibody where hydrolysis of a single site in the protein drastically increased the rate of aggregation (***115***). In order to understand under what circumstances hydrolysis may be an issue and to develop ways to stabilize biotherapeutics against hydrolysis, a better understanding of both the mechanism for non-enzymatic hydrolysis and the factors controlling it, is needed.

Several studies have investigated non-enzymatic hydrolysis experimentally (***112***,***116***,***114***,***117***,***118***,***119***). These studies have covered a wide range of proteins, from short peptides (***112***) to large antibodies, (***116***). They have found that hydrolysis may occur at many types of residues and structures. One of the common findings is the role of primary sequence. Specifically, when an aspartic acid residue precedes or follows a amide bond, that bond hydrolyzes at a rate at least ten times that of other amide bonds (***112***). Not only does an aspartic acid increase the rate of hydrolysis, but the other adjacent residues may also impact the rate of hydrolysis; ifor proline following aspartic acid, the rate of hydroysis can be increased by an order of magnitude over the case in which another amino acid follows aspartic acid. However, the primary sequence by itself cannot account for the wide range of reactivity seen in proteins, as factors other than the primary sequence affect the rate of hydrolysis. One example of this can be found in the hydrolysis of the hinge region of mAbs, where it was shown that if the sequence of the hinge region was expressed as a short peptide the rate of fragementation is five times greater than that of the full protein (***116***).

Several studies have also reported investigation of the hydrolysis pathway with various computational tools (***120,121,122,123,124,125***). However, many these studies have focused on small model compounds that do not necessarily contain the complex interactions that are present in proteins. In one study (***120***), the hydrolysis of formamide, the smallest molecule that contains a amide bond, was studied. From this study, the reaction seems to proceed through four steps; first, a proton adds to the amide bond's oxygen atom. This is followed by the addition of a water molecule to the carbon, forming a diol and a proton. This proton forms a hydronium ion in solution. Then, a proton from the hydronium ion binds to the nitrogen atom, and, finally, the amide bond breaks and forms the two final products. While this mechanism explains the energetics of formamide hydrolysis, it does not explain the effect of sequence or secondary structure on the rate of hydrolysis in proteins.

In order to develop a better understanding of the non-enzymatic hydrolysis of protein bonds and therein pave the way for a predictive algorithm, two topics are investigated: first, identification of the most energetically favorable mechanism for the reaction and second, the environmental factors which significantly influence this mechanism. To investigate these topics, two different strategies are used. First, quantum chemistry methods are used to investigate several potential non-enzymatic, acid-catalyzed hydrolysis mechanisms. Second, molecular mechanics simulations are used to investigate the differences in amide bond environments for reactive and non-reactive amide bonds. These two aspects are combined to produce a predictive algorithm.

Many potential pathways connect the unreacted peptide to the cleaved amide bond states. Several of these pathways for the non-enzymatic, acid-catalyzed hydrolysis of peptides were studied in this work. Three said pathways are based on the mechanism for formamide, as it is chemically similar to peptides, with the exception of other neighboring chemical moieties. Because the rate-determining step of formamide hydrolysis is the disassociation of water, this step was the focus of this study. One of the most significant differences between formamide and a peptide is the presence of additional moieties, which could act as proton acceptors and aid in this disassociation step or hinder steps of the process.

Four possible pathways are studied here (outlined in Figure 6-1). In the first pathway (labeled as A in Figure 6-1), a backbone carbonyl site (other than the hydrolysis site) aids in the disassociation of the water molecule, by acting as a proton acceptor for the proton produced during the diol formation step. In the next pathway, labeled as B in Figure 6-1, the process is identical to the formamide case, where the surrounding waters act as a proton acceptor. In the third pathway, labeled as C in Figure 6-1, the carboxylic acid on the aspartic acid acts as a proton acceptor. In the fourth pathway, labeled as D in Figure 6-1 and similar to the third pathway, rather than acting as a proton acceptor, the sidechain itself adds directly to the backbone carbonyl carbon, forming a furan ring. Then the nitrogen is protonated, and then the amide bond cleaved. The ring is then broken by the addition of a water molecule.

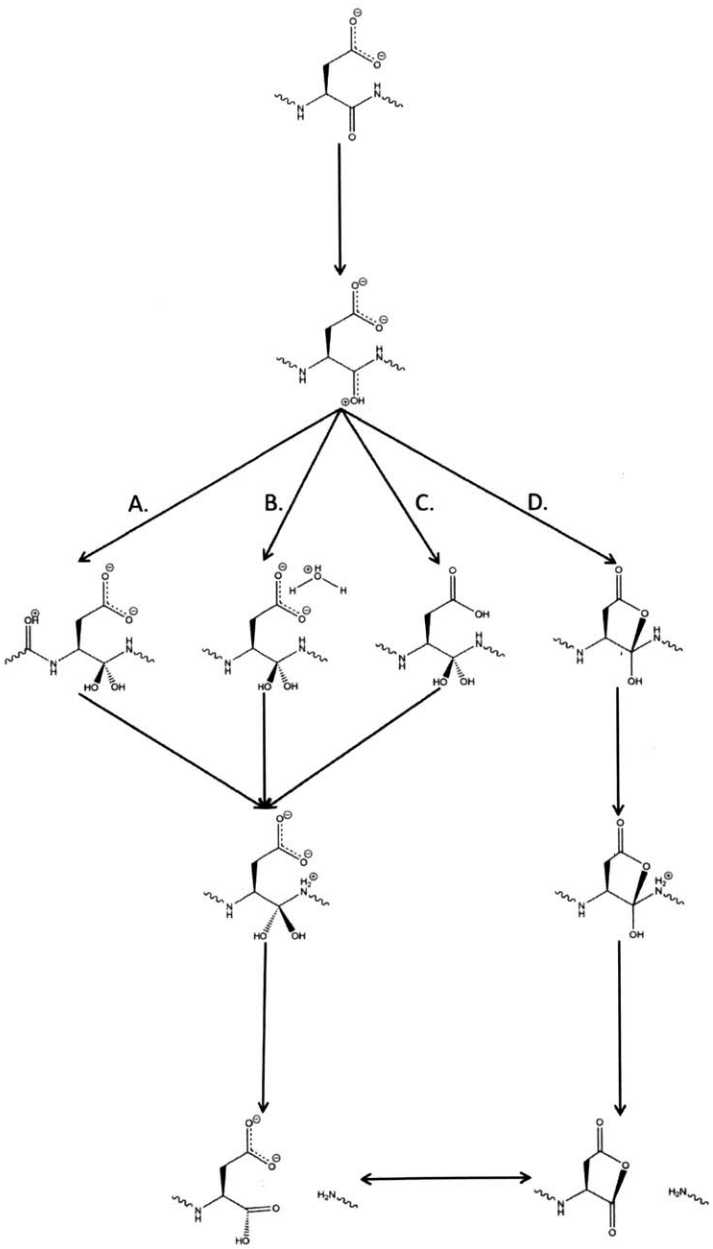

Figure 6-1 – The tested pathways for hydrolysis. All pathways start with an unprotonated peptide and hydronium ion. The next step for all pathways, the proton transfers to a backbone carbonyl unit from the surrounding water. During the next step, the four pathways, labeled A through D, diverge. Pathways A-C are similar and involve the formation of a diol intermediate and the

disassociation of a water molecule. The difference in pathways A-C is what chemical moiety acts as proton acceptor, in A it is a backbone carbonyl, in B the surrounding water, and in C the sidechain. The final pathway is D, where the sidechain adds to the backbone, forming a ring. The next step for all pathways is the addition of a proton to the amide bond's nitrogen. Then the amide bond breaks; this is the final step of pathways A-C, in pathway D, another step is required where a water adds and opens the ring.

In order to investigate how aspartic acid and proline affect the mechanism of hydrolysis, three peptide systems are studied. The first of these is proline – alanine – alanine (PAA) tripeptide, with the hydrolysis occurring between the second and third residues, the two alanine residues. The second peptide is selected to investigate the role of aspartic acid, proline – aspartic acid –alanine (PDA), with hydrolysis occurring between the aspartic acid and the alanine residues. The final peptide is the alanine – aspartic acid – proline tripeptide (ADP), which was included to determine what role proline has in the hydrolysis reaction. To prevent any long range charge-charge interactions, an acetyl group (-$COCH_3$) is attached to the N-terminus, and a nitro-methyl group (-$NHCH_3$) is attached to the C-terminus.

6.2. Methods

6.2.1. Computational Methods

6.2.1.1. Ab Initio Quantum Chemistry Methods

Standard *ab initio* molecular orbital theory (***126***) and density functional theory (***127***) are used. Gaussian 03 (***128***) is used for all structure optimizations, scans, and frequency analyses, while QChem 4.1 (***129***) is used for all single-point energy evaluations and implicit solvent evaluations. All structures are optimized using the Becke three parameter functional, B3LYP (***130***), with a 6-31+G** basis set. In addition to some explicit waters, all optimizations are done in an implicit solvent using the C-PCM (***131***) model. Frequency calculations are performed on all optimized structures to assure that they have either zero (for stable species) or one (for transition structures) imaginary frequency. Intrinsic reaction coordinate calculations (***132,133***) and analysis of the imaginary frequencies are carried out to determine if the transition states connected the appropriate minima on the potential energy surface. All scans are performed using the same parameters as optimizations. Initial guesses for stable species are obtained from a genetic algorithm (see below for more details). All density functional theory calculations are carried out using the ultrafine integration grid (in the case of Gaussian) or a grid involving 120 angular and 190 radial points (in case of QChem). Energy evaluations are done using ωB97X (***134***) functional with the 6-311++G(3df,3pd) basis set. Solvation free energies were calculated using the sm12 (***135***) implicit solvent.

In order to determine the impact of function and basis set choice on the energy and structures, several functionals and basis sets are tested on the formamide system. In terms of relative energy (ΔE_O), the entire ωB97 family of functionals performs well; all functions result in energies (ΔE_O) within 1 kcal/mol of those computed with CCSD(T)/6-311+G**. For this reason, the energy of these system is evaluated using the ωB97X/6-311++G(3df,3pd) which differs from the CCSD(T)/6-31+G** by less than 0.05 kcal/mol on average. Similarly, the impact of basis set on the optimized structure was investigated. Structure optimization using B3LYP/6-31+G** reproduces the structures of B3LYP/6-311++G(3df,3pd) well; with errors in relative energies of less than 1 kcal/mol and comparable imaginary frequencies and associated motions. More details can be found in the supplemental information.

In order to minimize the variability between pathways, a genetic algorithm (GA) is used to find the global minima structure for the complex of each peptide intermediate, surrounding explicit waters, and hydronium. The GA is applied separately for all species in all pathways, with either three or five solvating waters. The dihedral angles along the peptide backbone, the relative orientation, and the relative locations of all species are all included and explicitly optimized by the GA. The optimization is carried out using the GA developed by Wood, Santiso, and Trout to find the global minimum of water clusters (***136***). In addition to the standard mutation and crossover steps of the GA evolution, a short molecular dynamics (MD) simulation, of 2 ps in length, is also performed every fifty generations. This increases the chance that the system will leave the current local minima and explore neighboring minima. To assure that all generated structures are physically meaningful, the forces for all structures were evaluated at each step. Configurations with extremely high forces are removed, i.e. those with a force greater than 300,000 kJ mol^{-1} nm^{-1}. These systems are artifacts of the optimization, and arise when two atoms overlapped. Due to the number of energy evaluations that are required to converge the GA and the computational cost, the structures are evaluated using molecular mechanics. Therefore Gromacs (***71***) and Generalized AMBER force field (***137***) are used to evaluate the energies of the individuals.

6.2.1.2. Protein Simulation and Stability

In order to determine the role of secondary structure, MD simulations are used to determine the local environment for each aspartic acid and glutamic acid. All simulations are carried out using Gromacs 4.5 (***71***). Homology models are produced for each protein, with sufficient ions to neutralize the system. These structures and are initially minimized and then used to carry out NPT molecular dynamics simulations. The temperature and pressure are controlled using the Berendsen weak coupling method (***99***), to maintain a temperature of 300K and a pressure of 1bar. A 15.0 Å cut-off for nonbonded interactions is used in combination with the Particle Mesh Ewald procedure for electrostatics (***138***). Periodicity is enforced in all three directions throughout the simulation.

For prot-X, the initial structure is based on three crystal structures 1XIW (***139***), 1IKQ (***140***), and 2QHR (***141***). These are used to model the light chain, the toxin, and the heavy chain respectively. A 500 ns NPT simulation is run using AMBER99SB (***72***) and the TIP3P forcefield (***32***). A 90 ns simulation is run for each of the two mAbs, prot-Y and prot-Z. The initial structure for prot-Y is generated using three crystal structures 1HZH (***24***), 2XQB (***142***), and 3QEG (***143***), where 1HZH is used to model the CH2, CH3, 2XQB is used to model the CH1 and heavy variable region, and 3QEG is used to model the light chain. The initial structure for prot-Z is generated using three crystal structures, 1HZH, 3EOA (***144***), 4G5Z (***145***) , where 1HZH is used to model the CH2 and CH3, 3EOA is used to model the CH1 and heavy variable region, and 4G5Z is used to model the light chain. Both mAbs are glycosylated by GLYCAM (***96***) with a G0f glycosylation pattern (DGlcpNAcb1-2DManpa1-6[DGlcpNAcb1-2DManpa1-3]DManpb1-4DGlcpNAcb1-4[LFucpa1-6]DGlcpNAcb1-OME) added to the CH2 to produce stable proteins in simulation. The simulations for prot-Y and prot-Z are performed with the AMBER12 (***97***), GLYCAM (***98***), and the TIP3P force fields. The first 50 ns is used to equilibrate the system, with the remaining simulation time being used for property measurements.

All averaging of properties is done using MATLAB ® (***146***). Averages for angular quantities are done using the Circular Statistics Toolbox (***147***) to account for periodicity. Simulation averages are made based on structures stored every 20 ps. In order to account for the correlation of the properties over the time of the simulation and to evaluate the time needed to evaluate accurately these properties, the correlation times for each property were calculated using the block averaging method presented by Allen and Tildesley (***68***). These properties and the experimental data are then used to determine the difference between reactive and non-reactive residues in prot-X. This data is then used to develop an algorithm to separate the reactive and non-reactive residues. This model is then validated using both experimental cleavage site data and simulation derived properties for prot-Y and prot-Z.

6.2.2. Experimental Methods

To aid in the development of the predictive algorithm, and to test the developed algorithm, the stability of three proteins are studied experimentally to determine cleavage sites.

The three proteins studied are one immunotoxin, prot-X, and two IgG1 antibodies, prot-Y and prot-Z. These proteins are stored at elevated temperatures for a period of time; the resulting fragments are separated and studied. The formulation differed between proteins. Prot-X, the immunotoxin, is stored at 1 mg/mL with 25 mM sodium phosphate, 4% sucrose, 8% glycine, 0.02% polysorbate 80, at pH 7.4 and a temperature of 25 °C. While both mAbs are stored at 40 °C and a pH of 6.0 with no excipients. The fragments are separated using reverse-phase liquid chromatography. The mass of the separated fragments is then evaluated using mass spectrometry to determine the location of fragmentation sites. While these methods do produce quantitative data about the relative hydrolysis rates of different sites, due to the small amount of the various fragments, we were only able to categorize each site as either "reactive" or "unreactive".

6.3. Results

6.3.1. Investigations of Potential Hydrolysis Pathways

6.3.1.1. Gibbs Free Energies of Intermediate Species in Hydrolysis Pathways

The Gibbs free energies of all stable species are presented in Table 6-1. Due to the strength of the carboxylic sidechain as a proton acceptor, several species have not been found to be minima on the potential energy surface. Free energies vary depending on the number of explicit, solvating, quantum waters, the primary sequence, and the pathway of interest. However, several trends appear.

Solvation effects can be seen by comparing the Gibbs free energy of systems containing three and five solvating, quantum waters. In general, the systems containing five explicit, quantum waters are lower in Gibbs free energy than those evaluated with only three waters. In some cases, such as most of the A pathways of PDA, this difference is large and can be as much as 10 kcal/mol. However, this is not always the case, as many of the differences were substantially smaller. In the case of species originating from the peptide PAA, the differences with the number of waters is often 2 kcal/mol or less. While the larger differences between five and three waters are likely significant and due to insufficient solvation, the small differences seen in PAA are likely due to the inherit error associated with the energy evaluation method. More details about these errors can be found in supplemental information section 11 and 11.2.

In addition to the larger solvation effects found in the aspartic acid containing peptide, other variations can be seen. In general, for a given pathway, the reaction Gibbs free energy is on average 6 kcal/mol higher when an aspartic acid was present compared to when the aspartic acid is absent. However, this difference between aspartic and non-aspartic peptides is not constant across pathways, and varies from 3 kcal/mol to more than 15 kcal/mol. There are several potential causes for this difference including insufficient solvation, and the impact of the addition of the negative charge of the carboxylic acid. In addition to the increase in free energy due to the presence of aspartic acid, there is another effect with proline. When a proline is located at the after an amide bond, the reaction free energy is often slightly higher, by 3 kcal/mol in comparison to amide bonds followed by an alanine residue.

Additionally, there is also a wide variation between pathways. In general, pathway D, furane ring formation, has the lowest reaction Gibbs free energy of any pathway. This is followed by the energetically similar pathways B and C, in which water acts as proton acceptor or in which the sidechain acts as a proton acceptor, respectively. On average, they both have a reaction Gibbs free energy of 19 kcal/mol, which is nearly 8 kcal/mol greater than that of pathway D. The highest reaction Gibbs free energy pathway is pathway A, where the other backbone carbonyl groups act as the proton acceptor. Due to the three different backbone carbonyl groups present in the peptides, pathway A has a wide range of energies. The trend suggests that the further from the hydrolysis site, in these peptides it was always between the second and third residues, the greater the reaction Gibbs free energy. The difference between carbonyl sites is also very large in some cases. Comparisons of the reaction Gibbs free energies of the carbonyl group nearest the hydrolysis site and the one furthest from the hydrolysis site differ by approximately 10kcal/mol. The lowest of the reaction Gibbs free energies of pathway A are comparable to the Gibbs free energies of pathway B.

Table 6-1 - The change in reaction Gibbs free energy for each species in the four hydrolysis pathways. The first column is the pathway from Figure 6-1. A is the pathway where backbone carbon acts as the proton acceptor, B is where the solvating waters acts as the proton acceptor, C is where the sidechain acts as the proton acceptor, and D is where the sidechain cyclizes forming a ring. The relative Gibbs Free Energies for the product state are listed by originating peptide sequence and number of waters. All structures are optimized with B3LYP/6-31+G(d,p). The Gibbs free energy was

evaluating ωB97X/6-311++G(3df,3pd), using the harmonic approximation and sm12 for solvation free energy. All energies are reported in kcal/mol relative to the starting species for each pathway.

		Reaction Gibbs Free Energy					
Pathway	**Description**	**PAA Peptide**		**PDA Peptide**		**ADP Peptide**	
		3 H_2O	5 H_2O	3 H_2O	5 H_2O	3 H_2O	5 H_2O
A	**Backbone**	34	35.6	23.2	*	41.3	31.7
A	**Backbone**	25.6	20.5	41.8	23.1	31.8	36.1
A	**Backbone**	14.6	13.5	28	22.5	22.4	26.9
B	**Hydronium**	16.1	14	21.1	18.5	24.8	19.6
C	**Sidechain**	N/A	N/A	17.8	16.9	23.2	19.8
D	**Cyclic**	N/A	N/A	17.5	*	16.3	2.5

* no minima on the potential energy surface could be found
N/A species does not exist

6.3.1.2. Comparison of Geometries of Transition States

Most transition states that have been found tend to have similar geometries, Gibbs free energies, and imaginary frequencies, in particular for the first three pathways (pathways A, B, and C) where assorted chemical moieties act as proton acceptor. For instance, when considering only the first two pathways A and B, those where water (pathway B) or the backbone carbonyl group (pathway A) act as the proton acceptor, the distance between the backbone carbonyl group's carbon and the added water is approximately 1.6 Å in length. Similarly, for these pathways, the distance between the abstracted hydrogen and oxygen is approximately 1.2 Å. The Gibbs free energies of these transition states tend to be similar, between 30 and 40 kcal/mol higher than starting species. Most often the Gibbs free energy barrier was around 35 kcal/mol, especially when five explicit solvating water molecules are present. Lastly, the imaginary frequency associated with these transition states are also comparable between pathways, being between 550 and 750 Hz.

While there were a few outliers, such as the $(H^+)AD(OH)_2P$ of pathway A, the majority of the found transition states are geometrically and energetically similar. Transition states from pathway C, where the sidechain acts as the proton acceptors, have comparable energies to those from pathways A and B, but transition states from pathway C have slightly different interatomic distances. The geometries are fairly similar, except the interatomic distance between the carbon and the added water is slightly longer (by 0.2 Å) when compared to the previous two cases. The interatomic distance between the abstracted hydrogen and the oxygen atom that accepts it are comparable to those of the previous two pathways. The imaginary frequencies, average of -543 Hz, and Gibbs free energies of the transition state, average of 38 kcal/mol, are also comparable to the range of values for the previous two pathways (A and B).

The final pathway for which a transition state has been found is pathway D, where the sidechain bonds directly to the peptide backbone and forms a furane ring. This pathway is markedly different from the other pathways. This pathway has the lowest Gibbs free energy barrier of any pathway, and is more than 10 kcal/mol lower than the next lowest pathway. This pathway has similar carbon-oxygen interatomic distances to other pathways, however there were no comparable oxygen-hydrogen distances as there is no major proton motion in this step. The imaginary frequency associated with this transition state is also much lower than the other pathways, as is the motion associated with this frequency.

Examples of the transition states for each pathway can be seen in Figure 6-2, Figure 6-3, Figure 6-4, and Figure 6-5 (pathways A, B, C, and D respectively). Due to the size of the system and the inclusion of explicit solvating waters, we found it clearer to include only a skeletal representation of the transition states. The transition states for pathways A, B, and C are similar; in all three cases the atom with the greatest motion associated with the imaginary frequency is that of the abstracted hydrogen. This motion is in the expected direction, which shows the proton moving from the water molecule's oxygen to the proton acceptor's oxygen. Similarly, the motion of the amide bond's carbon and the water molecule's oxygen are significant. This motion brings the two atoms together to form the bond needed to produce the diol intermediate. In addition to the previous two motions, which are associated with the reaction, most other significant motions are needed to maintain the hydrogen bonds. This often affected the motion of the non-abstracted hydrogen of the added water, which becomes part of the diol. This motion maintains the hydrogen bond to its surroundings, frequently one of the solvating waters. For the most part, these motions do not appear in the imaginary frequency for pathway D. The motion of the imaginary frequency associated with the transition state of pathway D can be seen in Figure 6-5. There are two main motions associated with this transition state; the first is the motion between the amide bond's carbon and the sidechain's oxygen. This is similar to the other cases, and is associated with the formation of the bond between the carbon and the oxygen. This remaining motion is related to the net motion of the sidechain to form the ring. An example of this can be seen with the other oxygen of the sidechain, the one that does not form a bond to the backbone, which moves in conjunction with the rest of the sidechain to be closer to the amide bond.

Table 6-2 - Comparison of the transition states of hydrolysis pathway. The following properties for the found pathways are listed per transition state for each pathway (labeled by pathway from Figure 6-1 and the intermediate product formed by the transition state), the number of explicit waters, the Gibbs free energy of the transition state relative to the reactant (evaluated using ωB97X/6-311++G(3df,3pd)//B3LYP/6-31+G, harmonic approximation, sm12), the imaginary frequency, the distance between the carbonyl group's carbon to added water's oxygen, and the distance from the added water's oxygen and the abstracted hydrogen.**

Pathway and Expected Product		Explicit Waters [#]	Free Energy [kcal/mol]	Frequency [Hz]	C-O [Å]	O-H [Å]
A	$(H^{+})AD(OH)_{2}P$	3	41.1	-390	1.55	1.33
A	$(H^{+})AD(OH)_{2}P$	5	34.1	-679	1.57	1.27
A	$A(H^{+})D(OH)_{2}P$	3	34.6	-636	1.67	1.16
A	$A(H^{+})D(OH)_{2}P$	5	34.8	-762	1.53	1.19
A	$PD(OH)_{2}A(H^{+})$	3	35.6	-519	1.67	1.13
B	$PA(OH)_{2}A\ [H_{3}O^{+}]$	3	24.9	-569	1.57	1.18
B	$PA(OH)_{2}A\ [H_{3}O^{+}]$	5	25.4	-555	1.57	1.18
B	$AD(OH)_{2}P\ [H_{3}O^{+}]$	5	42.2	-746	1.75	1.19
B	$PD(OH)_{2}A[H_{3}O^{+}]$	3	45.0	-638	1.62	1.20
C	$PD(OH)_{2}(COOH)A$	3	40.2	-719	1.80	1.21
C	$PD(OH)_{2}(COOH)A$	5	36.4	-367	1.88	1.10
D	$PD(OH)_{2}(Cyc)A$	5	13.7	-107	1.68	

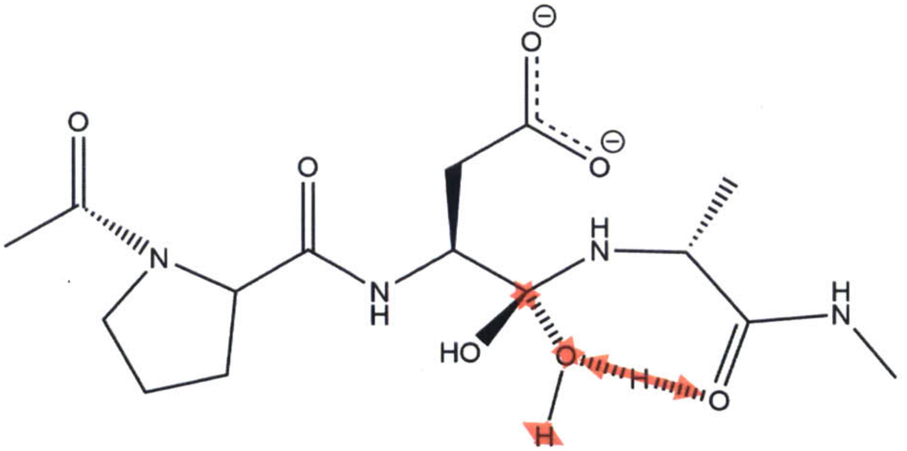

Figure 6-2 – Skeletal formula representation of transition state of pathway A, where one of the backbone carbonyl groups acts as proton acceptor, for peptide PDA, pathway produces product $PD(OH)_{2}A(H^{+})$. Dashed lines represent formed or broken bonds during the transition states. Red arrows indicate displacement associated with the imaginary frequency of the transition state. The length and direction of the arrow indicates the relative displacement and direction of the imaginary frequency of the transition state.

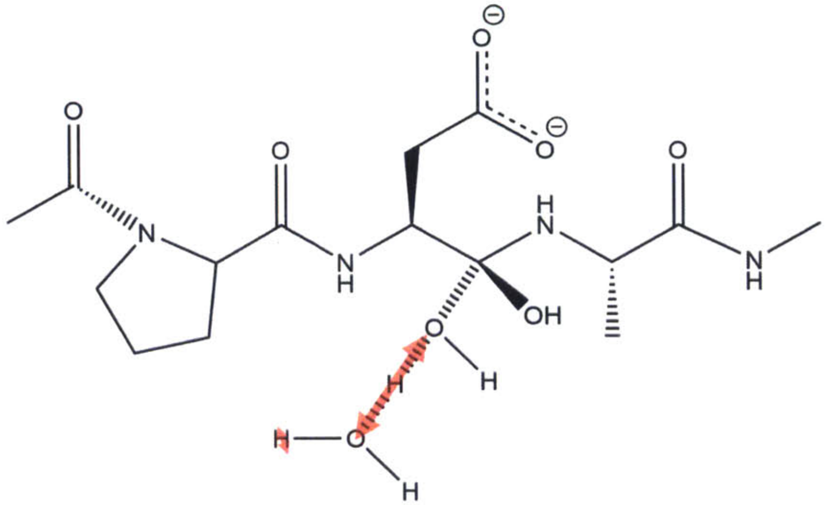

Figure 6-3 - Skeletal formula representation of transition state of pathway B, where the surrounding water acts as the proton acceptor, for peptide PDA, pathway produces product $PD(OH)_2A[H_3O^+]$. Dashed lines represent formed or broken bonds during the transition states. Red arrows indicate displacement associated with the imaginary frequency of the transition state. The length and direction of the arrow indicates the relative displacement and direction of the imaginary frequency of the transition state.

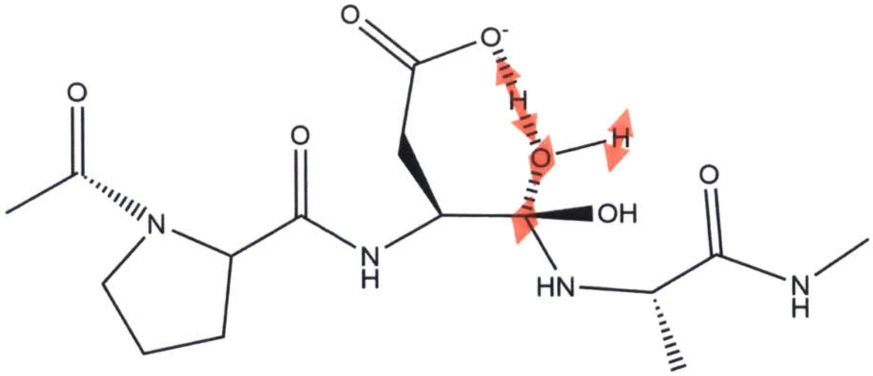

Figure 6-4 - Skeletal formula representation of transition state of pathway C, where the sidechain acting as the proton acceptor, for peptide PDA, pathway produces product $PD(OH)_2(COOH)A$. Dashed lines represent formed or broken bonds during the transition states. Red arrows indicate displacement associated with the imaginary frequency of the transition state. The length and direction of the arrow indicates the relative displacement and direction of the imaginary frequency of the transition state.

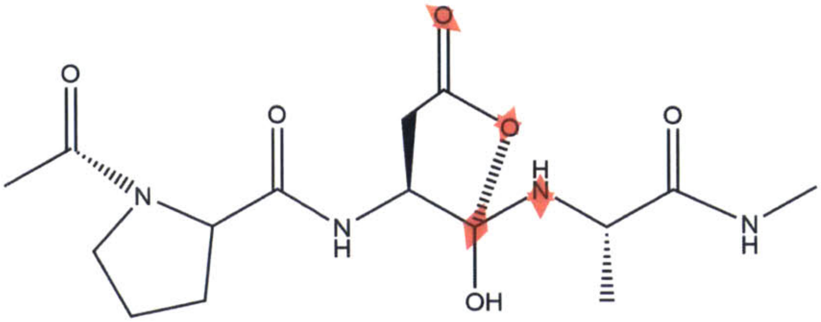

Figure 6-5 - Skeletal formula representation of transition state of pathway D, where the sidechain bonds to the peptide backbone, for peptide PDA, pathway produces product PD(OH)$_2$(Cyc)A. Dashed lines represent formed or broken bonds during the transition states. Red arrows indicate displacement associated with the imaginary frequency of the transition state. The length and direction of the arrow indicates the relative displacement and direction of the imaginary frequency of the transition state.

6.3.2. Investigations of the Impact of Bond Environment

6.3.2.1. Experimentally Identified Cleavage Sites

During the accelerated stability of prot-X, six cleavage sites have been identified with either an aspartic or glutamic acid residue preceding the amide bond. There are two aspartic acid-proline sites, H:D437-P438 and H:D446-P447. One is an aspartic acid – alanine site, H:D250-A251, one aspartic acid – serine, H:D247-S248, one aspartic acid – glycine, H:D266-G267, and one aspartic acid – valine, H:D269-V270. All six reactive residues are on the heavy chain, and their relative locations can be seen in Figure 6-6. Cleavage at the remaining aspartic acid residues, of which there are twenty-three, and all thirty-five glutamic acid residues have not reacted sufficiently to be detected. In the case of the two monoclonal antibodies, only two cleavage sites have been detected. These are the amide bond following the glutamic acid residue near the end of the light chain (prot-Y L:E214-C215 and prot-Z L:E217-C218) and the amide bond following the aspartic acid in the hinge region (prot-Y H:D220-K221 and prot-Z H:D225-K226). All other amide bonds with a following an aspartic or glutamic acid residue did not react in either prot-Y or prot-Z. Figure 6-6, Figure 6-7, and Figure 6-8 show the locations of the reactive amide bonds for prot-X, prot-Y, and prot-Z respectively.

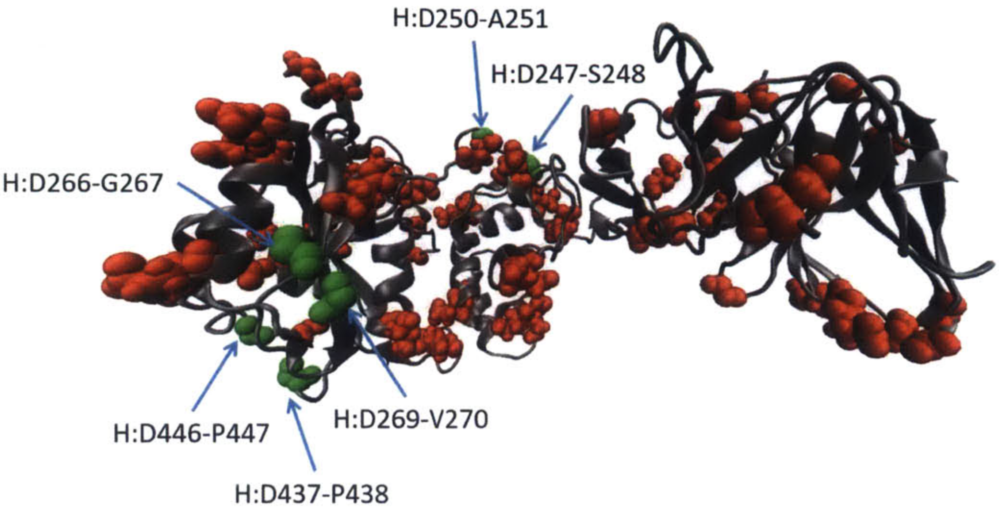

Figure 6-6 – Location of aspartic and glutamic acid residue in prot-X. Residues preceding cleaved amide bonds are shown in green and are labeled by chain and residue. Aspartic and glutamic acid residues that precede amide bonds but do not react are shown in red. Protein backbone for remaining residues is shown in grey with shape indicating the secondary structure.

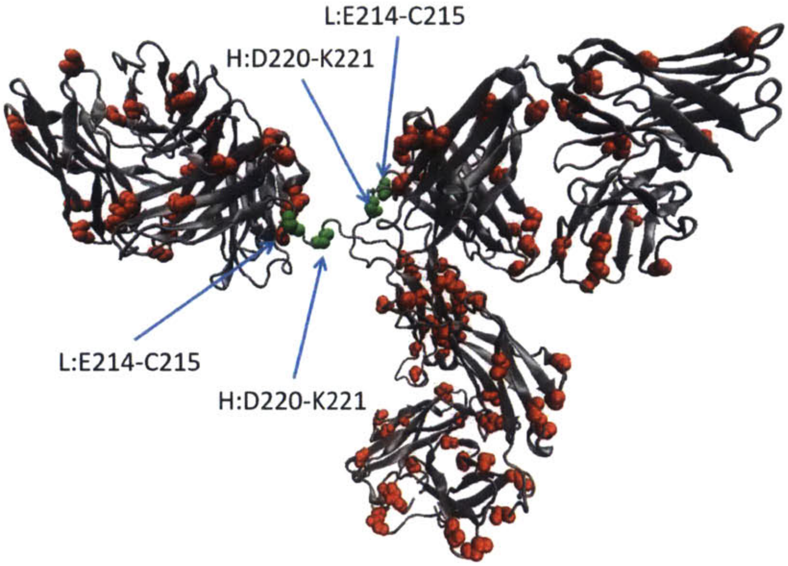

Figure 6-7 - Location of aspartic and glutamic acid residue in prot-Y. Residues preceding cleaved amide bonds are shown in green and are labeled by chain and residue. Aspartic and glutamic acid residues that precede amide bonds but do not react are shown in red. Protein backbone for remaining residues is shown in grey with shape indicating the secondary structure.

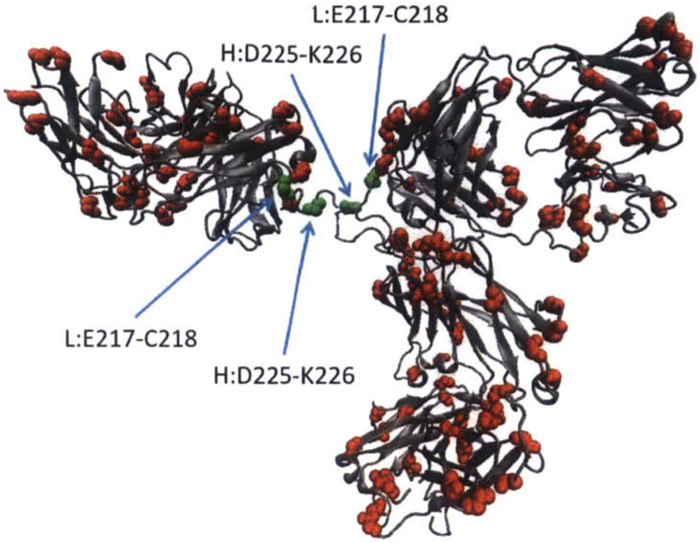

Figure 6-8 - Location of aspartic and glutamic acid residue in prot-Z. Residues preceding cleaved amide bonds are shown in green and are labeled by chain and residue. Aspartic and glutamic acid residues that precede amide bonds but do not react are shown in red. Protein backbone for remaining residues is shown in grey with shape indicating the secondary structure.

6.3.2.2. Properties that Correlate with the Reactivity of Amide bonds in Prot-X

Many environmental factors may affect the four explored pathways. These vary from properties based on the carboxylic acid's conformation, such as the distance between the sidechain and the amide bond, and interactions between the amide bond and surrounding chemical moieties, such as the number of water molecules within a given distance. A detailed listing of the measured properties can be located in supplemental information. A select few of these properties correlate well with the reactivity of the amide bonds. These are grouped into four basic properties: the secondary structure, the surface exposure of the amide bond, the relative orientation of the sidechain, and the availability of the sidechain to participate in the reaction.

6.3.2.2.1. Secondary Structure

VMD (***95***) and STRIDE (***148***) are used to evaluate the secondary structure, it is found that for all cases when a residue reacts, it spends most of its time, more than 90% of simulation time, as a turn or a random coil (see Figure 6-9). Additionally, the reactive residues spends virtually no time as either an extended conformation, an isolated bridge, an α helix, a 3_{10} helix, or a π helix (see Figure 6-10 and Figure 6-11). Combined, the reactive residues spend less than 10% of the simulation time in any of these conformations, and spends the entire remaining time either as a random coil or as turn. However, the relative amount of time it spends in these two conformation varied greatly; in one case the residue spends the entire time as a turn while in another case it spends nearly the entire time in a random coil. No reactive residue spends significant time in any conformation other than a turn or random coil. The non-reactive residues are located in a wider range of secondary structures; most spends significant time (more than 9% of simulation time) in a type of helix; twenty-two of the non-reactive residues spends more than 9% of simulation time as an α helix; eleven of the non-reactive residues spends sufficient time as a 3_{10} helix; nine of the non-reactive residues are in the extended conformation (β-sheets); and only one residue spends more than 9% of the time as an isolated bridge. No residue spends significant time in a π helix. This leaves twenty non-reactive residues (about 35% of the unreactive residues) that are unexplained.

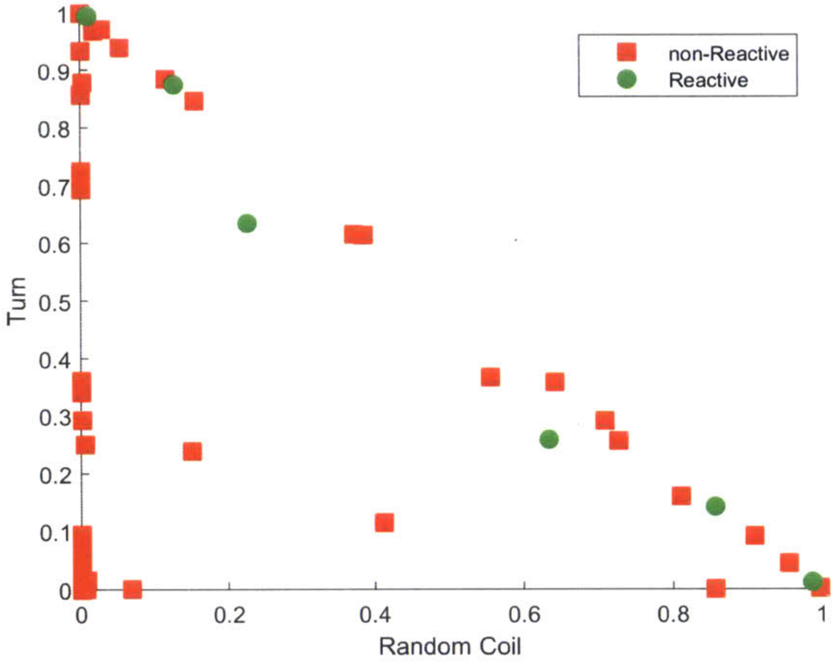

Figure 6-9 – The fraction of simulation time spent as a turn or random coil for each aspartic or glutamic acid residue in prot-X. The x-axis is the fraction of simulation time spent in a random coil. The y-axis is the fraction of simulation time spent as a turn. Red squares are values for residues that are followed by amide bonds that do not react. Green circles are values for residues that are followed by amide bonds that do react.

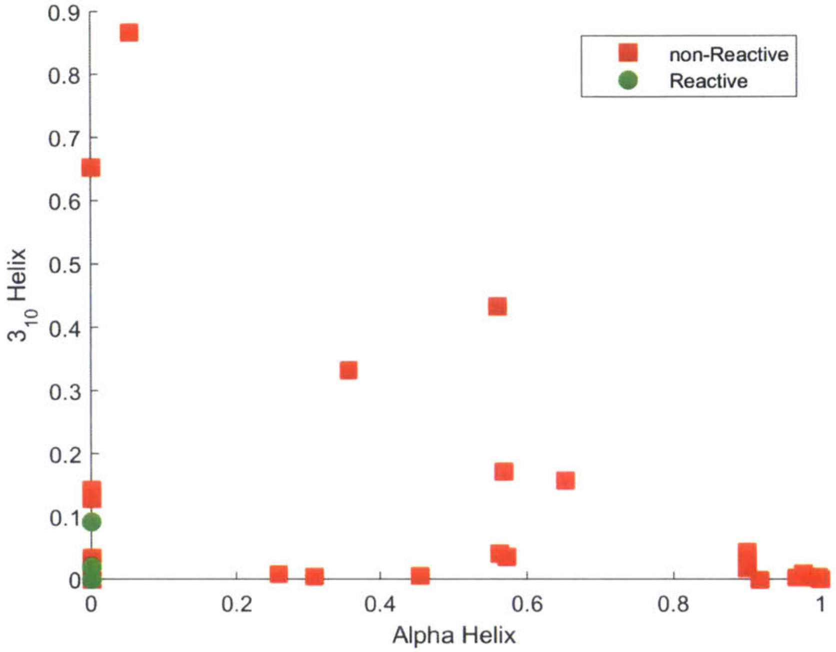

Figure 6-10 - The fraction of simulation time spent as an α helix or 3_{10} helix for each aspartic or glutamic acid residue in prot-X. The x-axis is the fraction of simulation time spent in an α helix. The y-axis is the fraction of simulation time spent as a 3_{10} helix. Red squares are values for residues that are followed by amide bonds that do not react. Green circles are values for residues that are followed by amide bonds that do react.

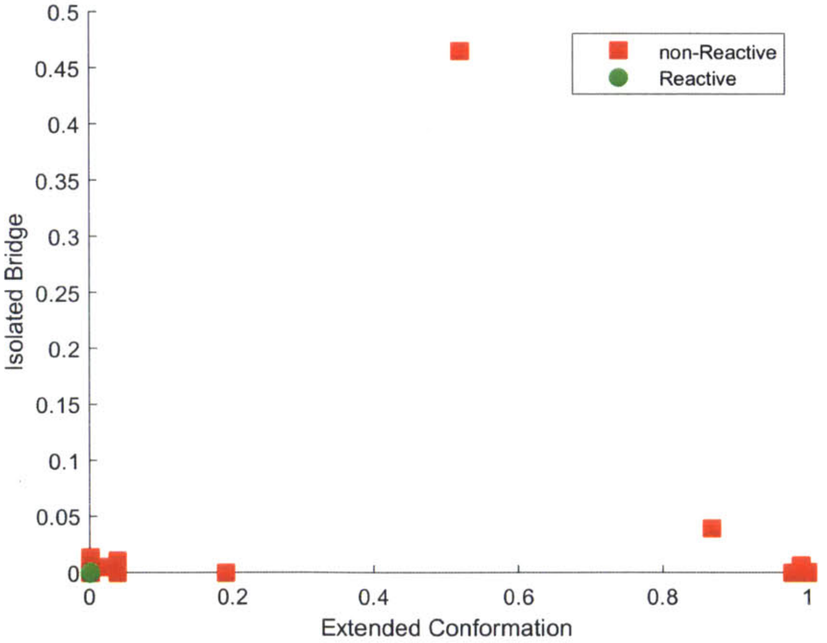

Figure 6-11 - The fraction of simulation time spent in an extended conformation or isolated bridge for each aspartic or glutamic acid residue in prot-X. The x-axis is the fraction of simulation time spent in an extended conformation. The y-axis is the fraction of simulation time spent as an isolated

bridge. Red squares are values for residues that are followed by amide bonds that do not react. Green circles are values for residues that are followed by peptides that do react.

6.3.2.2.2. Solvent Exposure

There are many other environmental factors that could also contribute to the reactivity or non-reactivity of an amide bond. One of these factors is the surface exposure of the amide bond. This may be a significant factor because a proton must migrate to the amide bond in order to react. The role of surface exposure can be seen by comparing the three aspartic acid – proline amide bonds. These three bonds include one that did not react, H:D337, and two that did react, H:D437 and H:D446. Using the radial distribution function (RDF) of either the amide bond's oxygen or carbon, it shows that the amide bond of H:D337 is partially buried, see Figure 6-12. In the case of reactive residues, there are significant non-zero values for radial distribution function approximately 2.75 Å from backbone oxygen (or 3.00 Å in the case of the backbone carbon). In the case of the non-reactive residues, the first substantial values for the radial distribution function are approximately 2 Å further than the reactive case. This spike between 2.75 and 3.00 Å in the RDF is present in all reactive residues (see Figure 6-13).

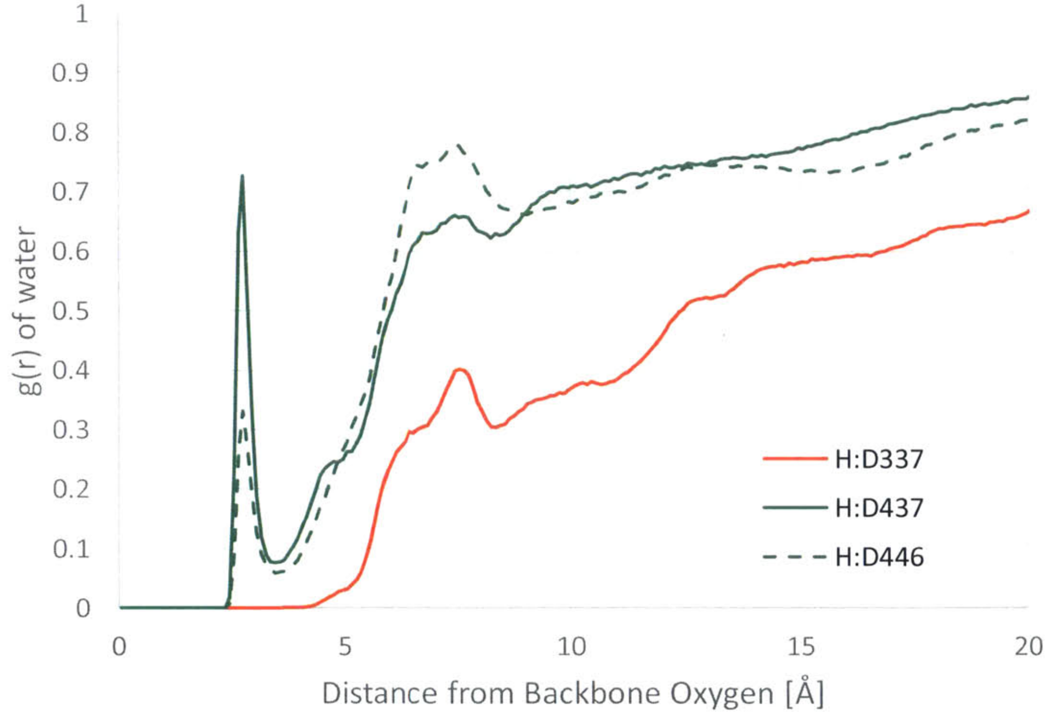

Figure 6-12 - The radial distribution function of water from the carbonyl oxygen for the three aspartic acid proline bonds in the immunotoxin. Reactive residues are in green, and the red line is for a non-reactive residue. This is the average value during the later 450ns of the 500ns simulation.

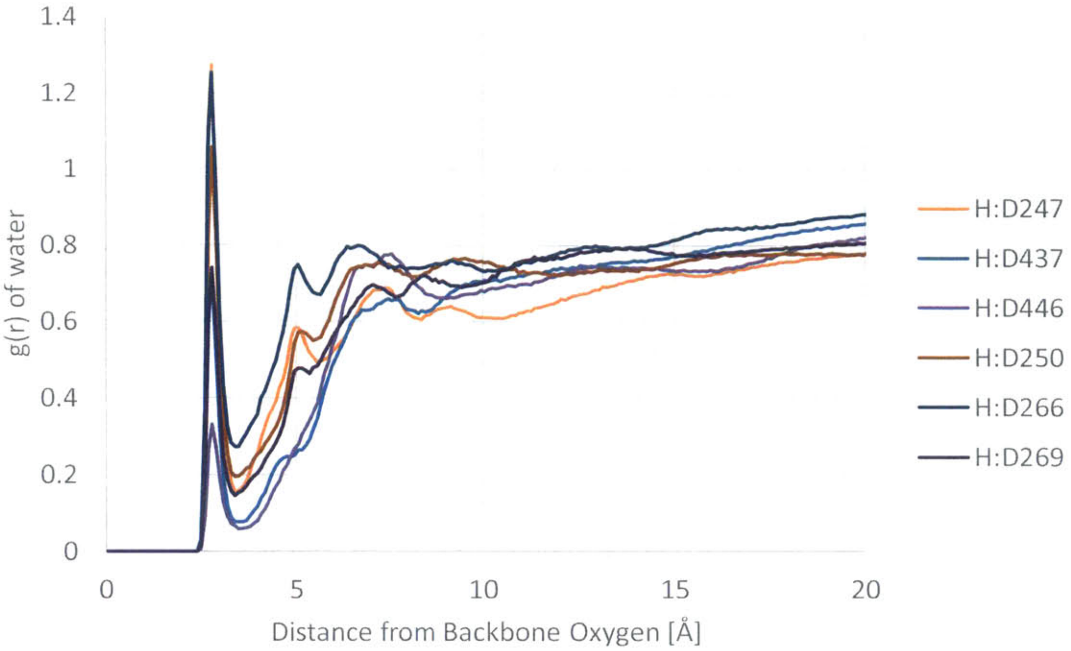

Figure 6-13 – The radial distribution function of water from the carbonyl oxygen for all experimentally identified reactive residues in the immunotoxin. This is the average value during the later 450ns of the 500ns simulation.

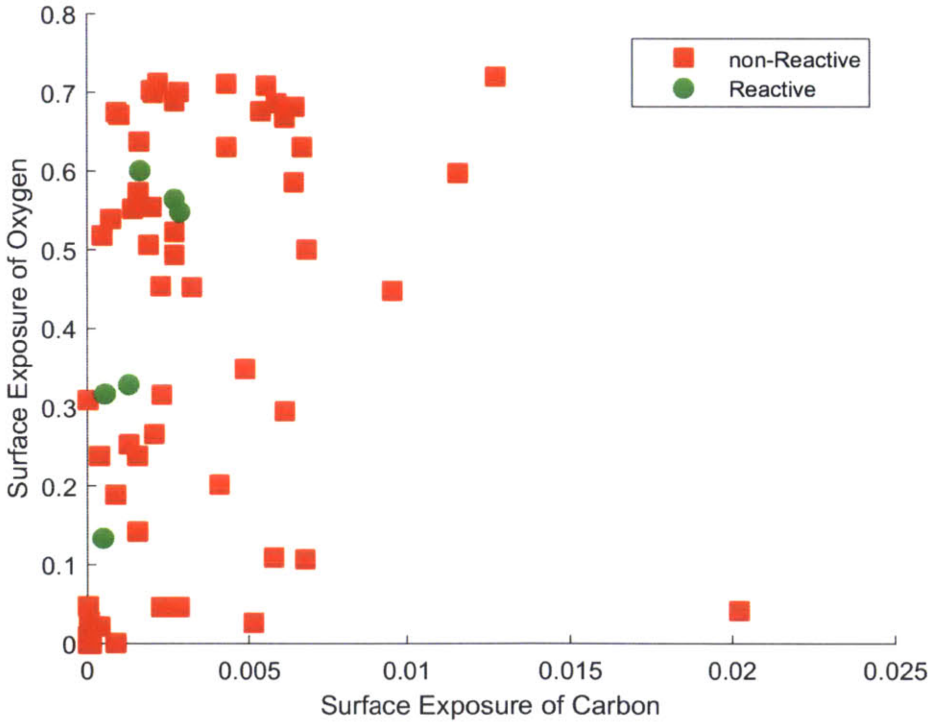

Figure 6-14 – The average number of waters within 3.00 Å of the residue's carbon and 2.75 Å of the residue's oxygen for each aspartic or glutamic acid residue in prot-X. The x-axis is the average

number of waters within 3.00 Å of the residue's carbon. The y-axis is average number of waters within 2.75 Å of the residue's oxygen. Red squares are values for residues that are followed by amide bonds that do not react. Green circles are values for residues that are followed by amide bonds that do react.

This trend of unreactive residues being buried and non-reactive residue not being buried can be seen in other measures of the surface exposure of the amide bond. While the different surface exposure measurement methods result in slightly different relative burial levels of the amide bonds, the separation based on these measures are very similar classifications. For instance, solvent available surface area (SASA) of the backbone oxygen atom classifies fifteen of the residues as buried, while the count of the number of waters within 2.75 Å of the backbone oxygen classifies sixteen residues as buried. While the various methods of surface measurement result in similar separations, many of the buried residues can also be excluded based on their secondary structure. Overall, two thirds of the residues are classified as "buried" also spent more than 9% of the simulation time in a helical or extended conformation, and could react.

6.3.2.2.3. Relative Orientation of Sidechain

There are many internal degrees of freedom that could also influence the reaction in various ways, in particular if the sidechain's orientation prevents interaction with the amide bond. Several dihedral angles, interior angles, and bond distances have been investigated to determine if the sidechain is in an orientation that permits the reaction to proceed, or if it is in a conformation that would require an additional conformation change before the reaction could proceed. While several dihedral angles have been tested, two correlate well with the reactive/non-reactive residue separation. They are the dihedral angle formed by the backbone carbon, the alpha carbon, the beta carbon, and the sidechain oxygen closest to the carbon atom (C-C_a-C_b-O_s) and the dihedral formed by the alpha carbon, the beta carbon, the gamma carbon, and the sidechain oxygen closest to backbone carbon (C_a-C_b-C_g-O_s). Both of these dihedrals separate the reactive vs. non-reactive very well, but they overlap. The dihedrals mark five and four residues as non-reactive respectively, with three of the residues overlapping. The limits on the reactive vs. non-reactive for these dihedrals (more than 36° and less than 45° respectively) place the sidechain's oxygen atoms over the backbone carbonyl. The spread of these two dihedrals is shown in Figure 6-15.

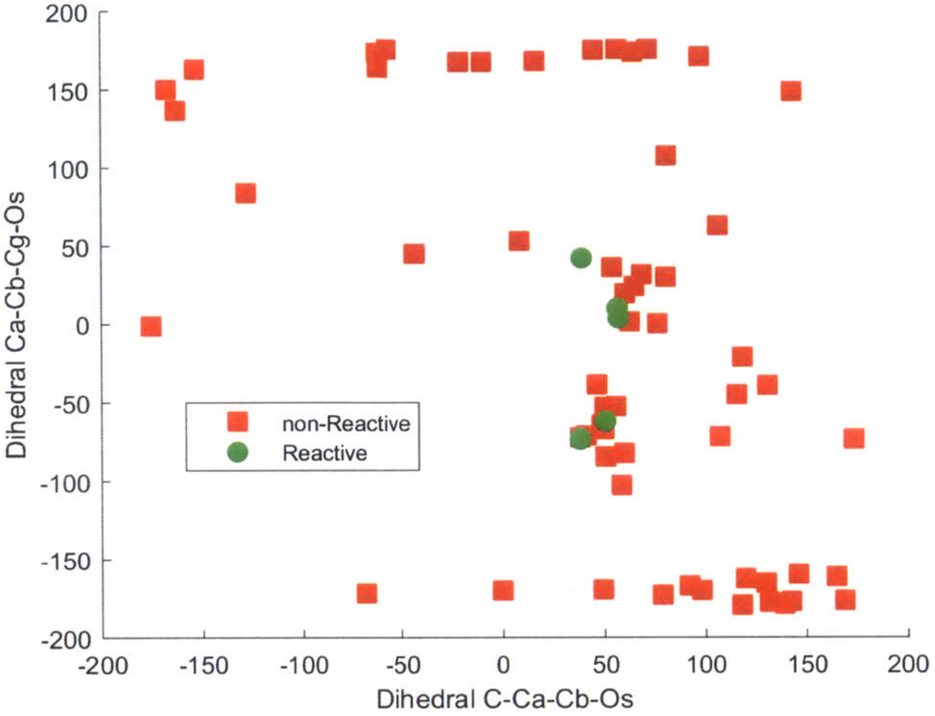

Figure 6-15 – The average dihedral angle of the C-Ca-Cb-Os and Ca-Cb-Cg-Os for each aspartic or glutamic acid residue in prot-X. The x-axis is the average dihedral angle of the C-Ca-Cb-Os. The y-axis is average dihedral angle of the Ca-Cb-Cg-Os. Red squares are values for residues that are followed by amide bonds that do not react. Green circles are values for residues that are followed by amide bonds that do react.

6.3.2.2.4. Availability of the Sidechain

Another property that could be of importance is the hydrogen bonding of the sidechain to other portions of the protein. This would impact the availability of the sidechain to participate in the reaction, as it would require the breaking of an addition bond before the sidechain could react. This is illustrated by H:D351-A352 and H:D250-A251. Both are aspartic acid – alanine amide bonds, and appear to be reactive based on their properties. One major difference is the presence of a nearby arginine residue (it is two residues after the H:D351). Because of its nearby location, the arginine sidechain forms a strong hydrogen bond with the aspartic acid sidechain. In this case, the arginine is ideally situated to form hydrogen bonds with the aspartic acid sidechain, and stays bound for approximately 80% of the simulation. Because the side-chain is hydrogen bound to another residue, it either could not be involved in the reaction, or it raises the energy barrier as it decreases the energy of reactant state.

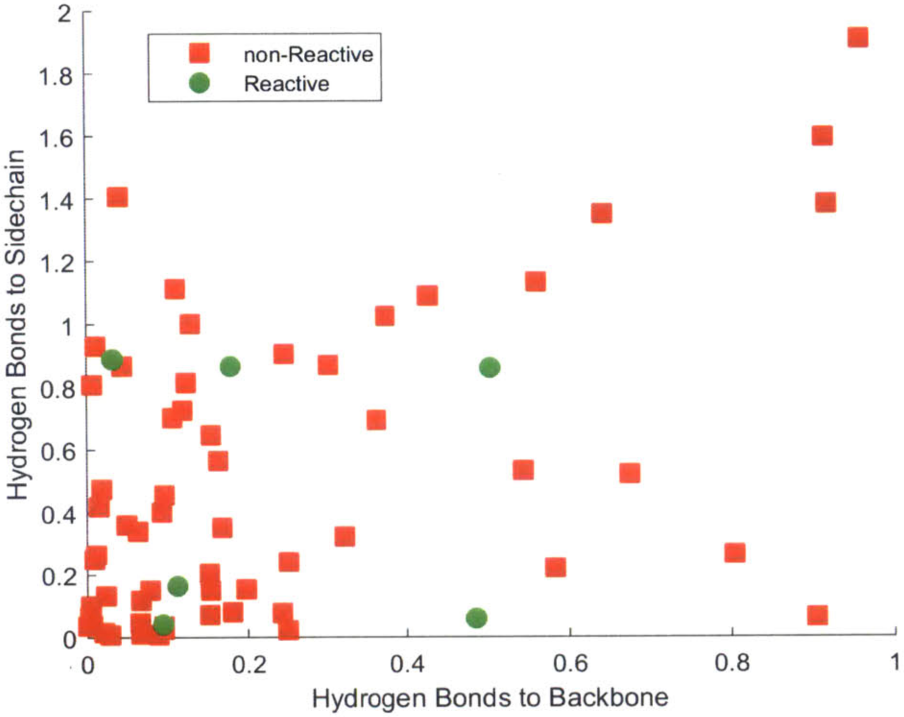

Figure 6-16 - The average number of hydrogen bonds between the backbone and the rest of the protein and the average number of hydrogen bonds between the sidechain and the rest of the protein for each aspartic or glutamic acid residue in prot-X. The x-axis is the average number of hydrogen bonds between the backbone and the rest of the protein. The y-axis is average number of hydrogen bonds between the sidechain and the rest of the protein. Red squares are values for residues that are followed by amide bonds that do not react. Green circles are values for residues that are followed by amide bonds that do react.

6.3.2.3. Algorithm Generation

The algorithm to predict degradation sites is fitted with the data for prot-X. Cutoffs for the various quantities are based on the extrema of the range of values observed by reactive residues during the prot-X simulation plus or minus 10% to account for variation during simulation. The first step of this algorithm determines if the secondary structure slows down the reaction or not. The criteria for this is if the residue spends more than 9.5% of the simulation time in either: α helix, a 3_{10} helix, a π helix, an extended conformation, or an isolated bridge. The next step determines the average number of waters within 2.75 Å of the oxygen atom of the backbone and the average number of waters within 3.00 Å of the backbone carbon atom. If either the number of waters within 2.75 Å of the backbone oxygen is less than 0.13 or 0.00046 waters are within 3.00 Å of the backbone carbon, the residue is said to be buried, and will not react. The last criteria measures the availability of the sidechain to participate in the reaction. This step was comprised of three measures; two intra-residue dihedral angles and the average number of hydrogen bonds between the sidechain and the protein. The first dihedral is between the backbone carbon, the alpha carbon, the beta carbon, and the sidechain oxygen closest to the backbone carbon. If this dihedral is greater than 36°, it may react, if it is less it does not. The second dihedral is between the alpha carbon, the beta carbon, the gamma carbon, and the sidechain oxygen closest to the peptide backbone. If this dihedral was less than 44° it may react. The final parameter is the number of hydrogen bonds between the carboxylic sidechain and the rest of the protein. If there is on average less than 0.9 hydrogen bonds, the residue may react, if it is more it would not react. The outcome of this algorithm applied to the aspartic acids in prot-X is listed in Table 6-3.

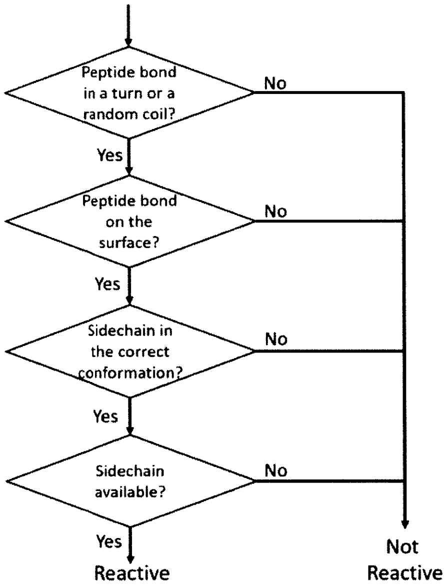
Peptide bond in a turn or a random coil?
No
Yes
Peptide bond on the surface?
No
Yes
Sidechain in the correct conformation?
No
Yes
Sidechain available?
No
Yes
Reactive
Not Reactive

Figure 6-17 – An outline for the algorithm to predict the reactivity of an amide bond following either an aspartic or glutamic acid. First step involves determining if the carboxylic acid (aspartic or glutamic acid) containing residue is inside an α helix, 3_{10} helix, π helix, extended conformation, or an isolated bridge for more than 9.5% of the simulation. Step two is determining if the carbon and oxygen atoms of the amide bond are on the surface as defined by the number of water molecules within 3.00 Å and 2.75 Å. The third step is to determine if the dihedral angles place the sidechain over the carbonyl group. The last step is determining if the sidechain is available to interact with the backbone peptide based on the number of hydrogen bonds between the sidechain and the rest of the protein.

Table 6-3 – Results of the Algorithm applied to prot-X's Aspartic Acids. Listed properties include the residue after the aspartic acid, the experimental result as reactive or non-reactive, whether or not the model predicts it will react or not react, and the reason it does not react if it is not reactive.

	FOLLOWING	EXPERIMENTAL	MODEL	REASON
	RESIDUE	RESULT	RESULT	
H:D188	GLN	Non-Reactive	Non-Reactive	In α helix
H:D203	LEU	Non-Reactive	Non-Reactive	In α helix
H:D237	GLU	Non-Reactive	Reactive	
H:D247	SER	Reactive	Reactive	
H:D250	ALA	Reactive	Reactive	
H:D266	GLY	Reactive	Reactive	
H:D269	VAL	Reactive	Reactive	
H:D324	LEU	Non-Reactive	Reactive	
H:D326	ALA	Non-Reactive	Non-Reactive	Buried oxygen
H:D337	PRO	Non-Reactive	Non-Reactive	Buried carbon
H:D34	MET	Non-Reactive	Non-Reactive	In extended conformation
H:D347	GLN	Non-Reactive	Non-Reactive	In extended conformation
H:D351	ALA	Non-Reactive	Non-Reactive	Too many Hydrogen Bonds
H:D403	ALA	Non-Reactive	Non-Reactive	Buried carbon
H:D437	PRO	Reactive	Reactive	
H:D444	LEU	Non-Reactive	Reactive	
H:D446	PRO	Reactive	Reactive	
H:D452	LYS	Non-Reactive	Non-Reactive	In α helix
H:D462	TYR	Non-Reactive	Reactive	
H:D474	LEU	Non-Reactive	Non-Reactive	In α helix
H:D63	THR	Non-Reactive	Non-Reactive	In 3_{10} Helix
H:D74	ASN	Non-Reactive	Non-Reactive	In extended conformation
H:D91	THR	Non-Reactive	Non-Reactive	In 3_{10} Helix
L:D18	ARG	Non-Reactive	Non-Reactive	In extended conformation
L:D2	ILE	Non-Reactive	Non-Reactive	Low C-C$_a$-C$_b$-O$_s$
L:D29	ILE	Non-Reactive	Reactive	
L:D42	GLY	Non-Reactive	Non-Reactive	Low C-C$_a$-C$_b$-O$_s$
L:D71	TYR	Non-Reactive	Non-Reactive	In extended conformation
L:D83	PHE	Non-Reactive	Non-Reactive	In 3_{10} Helix

6.3.2.4. Algorithm Based Predictions on Prot-Y and Prot-Z

The algorithm described in Section 6.3.2.3 is applied to the two mAb proteins. As there are two copies for each chain, each of the mAb simulation has two potential peptide cleavage sites per residue in the primary sequence, and the two bonds are different due to the asymmetry of the starting structure. Because of this, prot-Y has 114 potential hydrolyzing amide bonds. The model correctly predicts 105 of the residues as non-reactive, predicts 9 reactive residues, while only 3 of them are reactive. It missed 1 of the 4 reactive bonds (H:D220-K221), as it incorrectly classifies the bond as buried, as the amide bond's carbon is completely buried throughout the simulation. When applied to prot-Z it performs slightly worse, getting most the non-reactive residues correct, with 107 correct classifications as non-reactive, 3 correct classifications as reactive, 1 incorrect classification as non-reactive (L:E217-C218), and 20 incorrect classifications as reactive. Overall, the model classified 75% of all reactive residues correctly and 89% of the non-reactive residues are correctly classified.

Table 6-4 – A count of the number of residues classified as non-reactive based on the algorithm in section 6.3.2.3. The first column is the name of the measure used to classify residues as either reactive or non-reactive. The second, third, and fourth columns are the number of residues which are classified as non-reactive for prot-X, prot-Y, and prot-Z respectively. Extended Conformation, Isolated Bridge, α helix, 3_{10} helix, and π helix refers to the number of residues that spend more than 9.5% of simulation time in each protein. Buried Carbon and Buried Oxygen are the number of residues marked as non-reactive based on a buried carbon or buried oxygen atoms. $C\text{-}C_a\text{-}C_b\text{-}O_s$ is the number of residues marked as non-reactive due to the dihedral angle between the backbone carbon, the alpha carbon, the beta carbon, and the oxygen closet to the amide bond. $C_a\text{-}C_b\text{-}C_g\text{-}O_s$ is the number of residues marked as non-reactive due to the dihedral angle between the alpha carbon, the beta carbon, the gamma carbon, and the oxygen closet to the amide bond.

	PROT-X	PROT-Y	PROT-Z
EXTENDED CONFORMATION	9	45	39
ISOLATED BRIDGE	0	3	2
A HELIX	22	10	13
3_{10} HELIX	7	8	15
Π HELIX	0	0	0
BURIED CARBON	4	12	16
BURIED OXYGEN	1	4	1
$C\text{-}C_A\text{-}C_B\text{-}O_S$ DIHEDRAL	5	16	14
$C_A\text{-}C_B\text{-}C_G\text{-}O_S$ DIHEDRAL	1	4	4
HYDROGEN BONDS TO SIDECHAIN	1	3	3

Table 6-5 – The overall performance of the algorithm for each protein. False Non-Reactive Residues is the number of residues incorrectly classified as non-reactive. True Non-Reactive Residues is the number of residues that are correctly classified as non-reactive. False Reactive Residues is the

number of residues incorrectly classified as reactive. True Reactive Residues is the number of residues that are correctly classified as reactive. The results are separated by protein.

	PROT-X	PROT-Y	PROT-Z
FALSE NON-REACTIVE Residues	0	1	1
TRUE NON-REACTIVE Residues	40	105	107
FALSE REACTIVE Residues	8	6	19
TRUE REACTIVE Residues	6	3	3

6.4. Discussion

6.4.1. Mechanism of Non-Enzymatic Hydrolysis

From the Gibbs free energies of the stable species and the transition states, two main conclusions can be drawn: when an aspartic acid is present, the lowest Gibbs free energy pathway is pathway D from Figure 6-1, where the sidechain and peptide backbone form a ring, and proper solvation plays an important role in determining the energetics of these system.

Of the four tested pathways, all are impacted, to varying extent, by the presence of an aspartic acid. In two of the cases, pathway A and B where the backbone carbonyl and the surrounding waters act as proton acceptors, the addition of the aspartic acid increased the free energy of both the products and the transition states. In the other cases, pathways C and D, where the sidechain acts as the proton acceptor or the sidechain adds directly to the amide bond, the pathways are not possible if the carboxylic acid sidechain is absent. While the role of aspartic acid may vary between pathways, most of the transition states are similar in terms of energetics, geometry, and motion due to the imaginary frequency. This indicates that either the surrounding chemical moieties can do little to aid in the disassociation of water by acting as proton acceptors, or the differences are too small to be detected using the energy evaluation methods here. While pathways A and B may be the likely mechanism when an aspartic acid is not present, as the Gibbs free energy barriers of pathway B for the peptide PAA are comparable to the experimentally found activation energy of 23.5 kcal/mol for Glycine – Glycine bonds (*12*). Pathway D is the more likely when an aspartic acid is present. However, it is markedly different from the other three pathways.

Pathway D is the lowest energy pathway of the four pathways for the hydrolysis of a protein bond which follows an aspartic acid residue. Not only are the reaction Gibbs free energies slightly lower than any other pathway, but so is the Gibbs free energy barrier. Compared to the next lowest pathway, pathway B, the Gibbs free energy barrier of pathway D is nearly half of pathway B, more than 10 kcal/mol lower in energy. The cause for this drastic reduction in energy is likely due to the delay in the disassociation of the water molecule until after the amide bond cleavage. This is significant because pathways A, B, and C showed little change in the Gibbs free energy barriers despite various chemical moieties aiding in the disassociation; however, with pathway D this occurs after the measured event, the cleavage of the amide bond.

Outside of the presence of aspartic acid and the pathway, the next most significant factor in the determination of the reaction Gibbs free energies is the solvation of the product species. The importance of solvation can be seen in multiple ways: first in the energy difference when three explicit solvating waters are used and the same species with five solvating waters. While there are other sources for the inconsistent energies other than solvation, such as the change in conformation, most of these should be small and are unlikely to account for the large differences in energy, in some cases these differences are as much as 10 kcal/mol. The impact of solvation can also be seen in the comparing the energies of species and transition states for pathway A. Because each peptide has multiple carbonyl groups, and each of these groups is chemically different because it follows or proceeds a different residue, they were tested individually. The energies for different carbonyls varied by as much as 10 kcal/mol within a given sequence. While this could be due to the impact of primary sequence, no single pattern of preceding or following residue could account for this variability. A more likely cause is the uneven distribution of explicit waters along the peptide. Most of the explicit quantum waters solvate the diol, and the amide bond that will be later cleaved. Because of this, the carbonyls nearest this site were better solvated than those further from the cleavage site. This leads to the first instance of pathway A for each peptide, the cases where the proton accepting carbonyl group is furthest from the diol site, having the highest energy, regardless of the variations in sequence. These large changes in energy, in excess of 10 kcal/mol, are most likely due to insufficient solvation of the species. Because of the importance of solvation in determining the reaction Gibbs free energy of the system, it will likely be a significant factor in determining the reactivity of a residue.

There are two significant factors that affect the energies of species. These are the pathway to which the species is a part and the solvation of that species. Other factors also affect the energetics of the species, such as the impact of a proline following an aspartic acid, however these were close to the accuracy of the methods used for energy evaluation and may not be significant.

6.4.2. Factors Effecting Hydrolysis and Their Impact on The Hydrolysis Mechanism

Using this information about the mechanism of hydrolysis- that the most energetically favorable pathway is pathway D in Figure 6-1 and the importance of solvation- a predictive algorithm has been developed. Knowledge of the mechanism is critical in identifying key factors and in interpreting why those factors are significant. Four factors have been found that influence the reactivity of an amide bond. The first is the secondary structure of the amide bond. The second is the surface exposure of the bond. The third is the orientation of the sidechain relative to amide bond. The fourth factor is the availability of the sidechain to participate in the reaction. The first two factors, the secondary structure and the surface exposure of the amide bond, directly impact the ability of the amide bond to accept the proton. The last two factors,, the orientation and availability of the sidechain, directly impact the ability of the sidechain to form a ring. Of these four factors, the factor that can account for the most non-reactive classifications is the secondary structure. It determines the non-reactivity of nearly 65% of all residues. However, the secondary structure is a complex property; it is determined by several other properties, such as dihedral angles along the backbone. Therefore, the impact the secondary structure has on hydrolysis is unclear; in order to understand if it is the secondary structure, or one of its determining factors that controls hydrolysis, both were investigated separately. In STRIDE, the algorithm used to determine the secondary structure, a number of properties are used to determine which class of secondary structure best describes a residue. These include three dihedral angles, the φ, ψ, and ω and the hydrogen bonding of the backbone. Overall, the three dihedral angles do not correlate with reactivity. While the data for prot-X suggests that it may be possible that the φ and ψ dihedrals are correlated, see Figure 6-18 and Figure 6-19, this is due to the relatively small data set of reactive residues in prot-X. A wider range of angles are explored in prot-Y and prot-Z, which can help rule out these angles as a determining factor in the reactivity of the amide bond. The next factor that determines the secondary structure is the hydrogen bonding of the backbone. As can be seen in Figure 6-16, amide bonds that are hydrogen bound to other parts of the protein for more than half of the simulation can still react, making the mere presence of a single hydrogen bond an unlikely factor in determining the reactivity of an amide bond.

Because none of the factors which determine the secondary structure directly correlate with the reactivity of amide bonds, the question becomes how does the type of secondary structure effect the hydrolysis reaction? One possible route by which the secondary structure could impact the hydrolysis mechanism is through raising the barrier to the first step of the reaction, the protonation of the amide bond. Fundamentally, this step turns the carbonyl group which had been a hydrogen bond acceptor into a hydrogen bond donor. Because secondary structures are stabilized by patterns of hydrogen bonding, this could drastically impact the stability of the pronated state. In the case of a random coil or a turn, it would only require the breaking of a single hydrogen bond for the backbone carbonyl to be present on the protein surface and able to accept a proton. If the amide bond was in another type of secondary structure, such as an α helix, multiple hydrogen bonds would need to be broken to open up the helix so that amide bond could be protonated. This would require a more significant change in structure and would further increase an already high energy barrier, and would greatly reduce the probability of the reaction occurring.

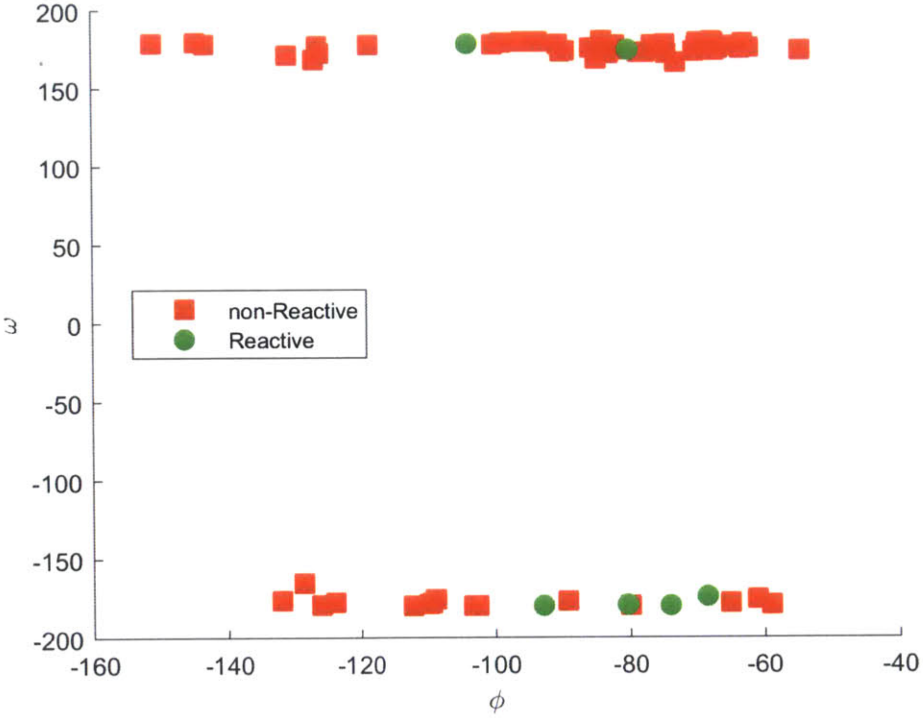

Figure 6-18 - The average dihedral angle of the ϕ and ω for each aspartic or glutamic acid residue in prot-X. The x-axis is the average dihedral angle of the ϕ. The y-axis is average dihedral angle of the ω. Red squares are values for residues that are followed by amide bonds that do not react. Green circles are values for residues that are followed by peptides that do react.

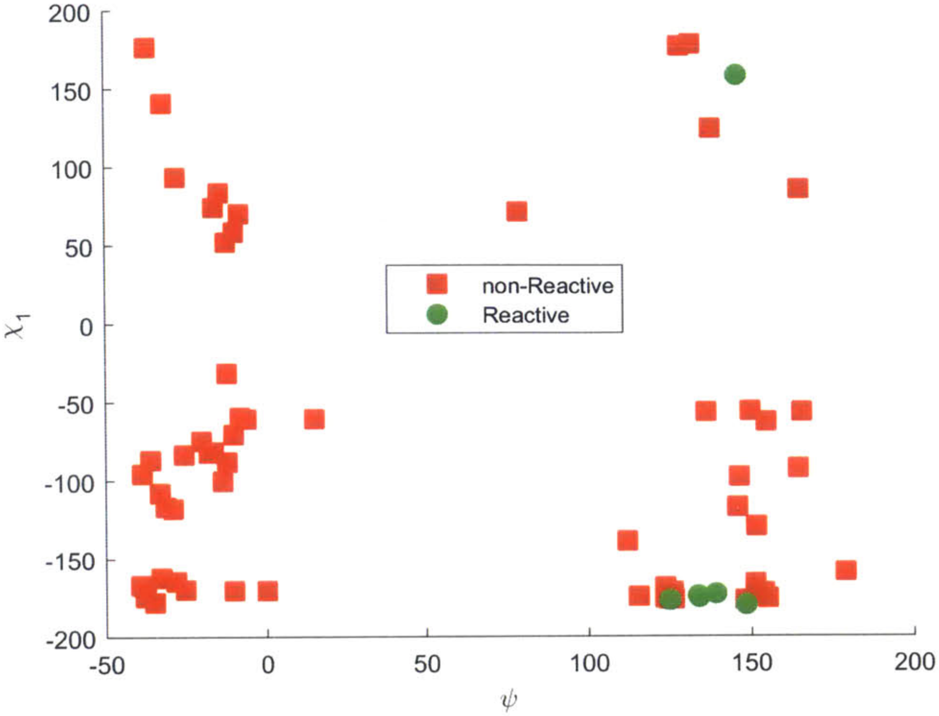

Figure 6-19 - The average dihedral angle of the ψ and χ_1 for each aspartic or glutamic acid residue in prot-X. The x-axis is the average dihedral angle of the ψ. The y-axis is average dihedral angle of the χ_1. Red squares are values for residues that are followed by amide bonds that do not react. Green circles are values for residues that are followed by peptides that do react.

Another factor that also captures this accessibility of the amide bond to the proton is the surface exposure of the amide bond. In a similar way to the secondary structure, the surface exposure of the amide bond controls if the proton can get to the amide bond. If the surface exposure of the bond is low, meaning the amide bond is buried, a conformation shift would be required to expose the amide bond. This conformation change would further increase the barrier of the reaction, and slow the hydrolysis of that amide bond. Unlike the secondary structure, this factor also measures how well solvated the bond is, and as previously mentioned, the energies of the both the transition states and the products can vary greatly with differences in solvation. A poorly solvated amide bond not only will have difficulty in transporting a proton to the site for the reaction to begin, but will also have an overall slower reaction due lack of solvent molecules to stabilize the transition state. This would result in an increase of the energetic barrier, and would further slowdown the reaction. For these reasons, the surface exposure has a significant impact on the reactivity of the amide bonds.

The third factor which correlates with reactivity is the orientation of the sidechain relative to the peptide backbone. This is measured through the two dihedral angles; these place the sidechain on the same side of the amide bond as the oxygen of the amide bond. These angles are significant because this orientation places the sidechain's oxygen in a similar location as can be found in the product of pathway D, where the sidechain cyclizes. The carbon of the amide bond forms an sp3 carbon after the reaction; and this would require the oxygen of the sidechain to be on the same side of the amide bond as the amide bond's oxygen. If the sidechain spends the majority of its simulation time in a conformation similar to the conformation in the product of pathway D, the reaction can proceed at any time. In other cases, where the sidechain is in another conformation, a conformation change would be required before the cyclization step could begin. This would add another energetic barrier that would further slowdown the reaction.

The final factor is the availability of the sidechain to participate in the reaction. It is measured through the average number of hydrogen bonds between the sidechain and the rest of the peptide. If a sidechain is strongly hydrogen bound to another part of the protein, then these bonds would need to be broken for the sidechain to participate in the furane formation step, regardless of the its current conformation. Like the conformation of the sidechain, the availability of the sidechain will impact the second step of pathway D, and would further increase the reaction barrier to the whole process.

These four factors, the secondary structure, the surface exposure, the orientation of the sidechain, and the availability of the sidechain, each impact the cyclization mechanism. Two factors, the secondary structure and the surface exposure, determine if protonation of the amide bond, the first step of the reaction, is possible. While the other two factors, the orientation and availability of the sidechain, determine if sidechain can interact with the peptide and participate in the reaction. These were then combined into an algorithm to predict if a bond will hydrolyze or not.

6.4.3. Performance of Algorithm and Sources of Error

Overall, these descriptors did well in predicting which bonds would react. More than 89% of bonds are predicted correctly as reactive or non-reactive for the two mAbs. Most bonds could be excluded based on their secondary structure; approximately 65% of all amide bonds are excluded based on spending too much time in one of the non-reactive secondary structure classes. The burial of amide bonds, and the relative orientation of the sidechain account for roughly equal amounts of non-reactive amide bonds, 14% and 16% of amide bonds respectively. The hydrogen bonding of the sidechain to other parts of the protein accounts for few non-reactive classifications, only 3% of the overall amide bonds. Of those that are incorrectly classified as reactive or non-reactive, no single parameter can explain the thirty-three incorrectly classified residues. However, there are several potential causes that could explain the misclassification. These include: the role of the residue following the aspartic or glutamic acid, the binary representation of the data, the small training set size, insufficient measurements of the properties of the various amide bonds, and the complexity of the mechanism.

As has been shown both here (see section 6.3.1.1) and experimentally elsewhere (*2*) that the primary sequence does affect the rate of hydrolysis, specifically the type of residue following the amide bond can impact both the energetics or the rate of the reaction. However, no term for the primary sequence was explicitly included. No term has been included because how the residue after the aspartic or glutamic acid residue impacts hydrolysis is unknown. This presents two issues; the first is how to quantify the difference between residues. There are many differences between the common residues such as hydrophobicity, charge, dipole, aromaticity, etc., any of which could impact the reaction in slight or significant ways. But, the experimental and computational data does not cover a sufficient range of residues to determine which measure or measures would impact the mechanism enough to warrant inclusion in this algorithm. The second issue is that there is the possibility that the residue after the aspartic or glutamic acid could fundamentally alter the mechanism. While that was not seen here, only two residues were investigated. Further investigation of this could improve the algorithm; if the hydrophobicity of the following residue were included, then the number of incorrect classifications could be cut by a third. However, there is no statistically significant correlation between the limited experimental rate data and the hydrophobicity of the following residue.

Another potential explanation for the misclassifications is in the binary representation of the data, specifically that an amide bond will either react or not react. This does not match reality, as all amide bonds undergo non-enzymatic hydrolysis. The relative rates of hydrolysis will differ between amide bonds, but all are susceptible to some degree. As mentioned in section 6.2.2, the use of only the reactivity of the amide bonds was necessary due to the small amount of fragments produced during the accelerated stability study. Because of the relatively small amounts, often less than 1%, the uncertainty associated with the quantified data is high. Due to this binary nature of the data, two different amide bonds may react very similarly, and have similar rates, if that rate is close to the detection limit, then one may be found to be reactive while the other is non-reactive. If this is the case, then some of the incorrectly classified as reactive, may in fact be reactive, only they react slowly enough not to be detected.

Along with the limits on the reactivity of the different amide bonds, the algorithm was trained on relatively small data set, only sixty-four amide bonds, and of these, only six were reactive. There is the risk that when the cutoff between reactive and non-reactive residues was selected it may not perfectly match the actual separation. These cutoffs may be better selected using any of a number of machine learning algorithms to select the optimal cutoff (***149***). However this would require a much larger training set, as when these methods were attempted here (see section 11.5), they produced a model with minimal predictive ability. The same imperfection in cutoffs can be also be seen in section 6.3.2.2.2, were several methods for the classification of buried vs. non-buried residues were tested. Ideally, all of these methods will result in the same classifications, but due to the limits of sampling the classifications are slightly different. With the inclusion of data for more proteins, better reactive/non-reactive cutoffs may be determined.

Another potential cause for misclassification is insufficient sampling of the properties of interest. This is always a possibility when performing MD simulations. The simulation may not be long enough to adequately sample the conformational space. There is some evidence of this in the predictions made for the two proteins, prot-Y and prot-Z; because prot-Y and prot-Z are both mAbs, there are two instances of both the heavy and the light chain, and because the proteins are symmetric, there is no physical reason why one instance of the heavy chain should react and not the other. Overall, twenty-five amide bonds are misclassified; however, in fifteen of these cases only one of the two instances of that amide bond is misclassified. This includes the two false non-reactive residues; where one instance of each is classified as buried because of its surface exposure and the other instance is not. This suggests that it may be possible for some of the misclassified bonds to be correctly classified by the algorithm if a longer simulation was used. In order to determine how long of a simulation would be required; the correlation times for each parameter were measured.

The correlation times for these properties varied greatly. By both the average and maximum values, the longest correlated properties were the secondary structure, which had a correlation time on average of 3.5 ns, suggesting that a very long simulation might give better results. However, the correlation time varies greatly with the residue. The longest of these correlation times was more than 35 ns, nearly as long as the simulation time used to evaluate properties for both mAbs, which would indicate that only a few independent evaluations were possible in some cases. Long correlation times affect not only the secondary structure classifications, but a number of properties used to separate reactive and non-reactive bonds have correlation times on the nanosecond time scale, including the number of water within 2.75

Å (average time 0.8 ns) and the average number of hydrogen bonds between the sidechain and the rest of the peptide (on average 2 ns). These correlation times indicates that a long MD would be required to adequately sample these degrees of freedom, probably in excess of 100 ns of measureable MD time.

Lastly, hydrolysis is a complicated mechanism and this only considers a single mechanism. Many factors could impact the rate of hydrolysis, while many of them were investigated here, one or more factors may have been missed that could have been included. Also, there are several pathways that could lead to amide bond cleavage and produce products similar to hydrolysis. This includes cases for which a carboxylic acid containing residue, aspartic or glutamic acid, is not present. In the case of prot-X, the only remaining cleavage site is Asn243-Gly. Asparagine is chemically similar to aspartic acid, so it could react in a similar method as aspartic acids, however deamidation is a major degradation route for asparagine residues. Prior work has been done to understand deamidation, (***150***), this includes the development of a predictive method. When this predictive method was applied to prot-X, Asn243 was found to be the most likely site for deamidation, and has a half-life a tenth that of the next most reactive residue. There are several ways that deamidation of Asn243 could cause cleavage of the amide bond. One of these is through conversion of the asparagine into an aspartic acid and then the aspartic acid reacting in the manner previously described for carboxylic acid containing residues. However, a side reaction to deamidation has also been found that results in bond cleavage (***150***). There are several other potential causes for peptide cleavage beyond hydrolysis that could be investigated and lead to a more advanced predicative method that is applicable to a wider range of residues.

6.5. Conclusion

The mechanism, and factors that affect it, of the non-enzymatic, acid-catalyzed hydrolysis of amide bonds which follows either an aspartic or glutamic acid residue was investigated. The lowest energy pathway was the protonation of the amide bond, followed by the formation of a furane intermediate, which is later opened, after the addition of a water molecule. Several environmental factors can influence this mechanism; these include those which hinder the protonation step, such as the secondary structure and the solvation of the peptide backbone, and those which affect the furane formation step, such as the conformation of the sidechain and the availability of the sidechain to participate in the reaction. This information was used to generate an algorithm to predict if a residue would or would not react. This algorithm predicted nearly 90% of residues correctly, as reactive or non-reactive, in two different test cases. The algorithm is tuned so that incorrect categorization of non-reactive is minimized. (Only two reactive residues were incorrectly predicted as non-reactive.) This algorithm can thus be incorporated into the discovery phase or early stage development to identify potential problems with hydrolysis early on.

7. CONCLUSION

The development of a new drug is a substantial investment for a pharmaceutical company. Not only is the development process expensive, some estimates place the cost at nearly a billion dollars per drug, and time consuming, taking a decade to go from drug candidate to drug on the market, but it is also risky, with less than one drug candidate in ten ever reaching the market. In order to decrease the risk, and thereby the cost, of biotherapeutics, a number of procedures have been suggested, including using *in silico* tools to predict the properties of the biotherapeutic. In general, *in silico* tools have a number of advantages over experiment based tools, including: no need for material, high throughput, and rapid results. Additionally, these tools can be designed to address virtually any protein property from immunogenic response to the viscosity of the protein in solution. Through the use of *in silico* tools, potential issues can be identified early in the development process; allowing the developers to mitigate these issues. However, relatively few of these tools have been developed that have been validated for the larger proteins of interest to the pharmaceutical industry, such as antibodies. In this thesis, two *in silico* tools are presented that tackle two very different degradation routes; the first predicts the aggregation propensity of a protein and the second predicts the peptide bonds that are susceptible to non-enzymatic, acid-catalyzed hydrolysis.

The first of the *in silico* tools is the Developability Index. This tool was developed to address the aggregation of proteins in solution. It classifies proteins into one of three classes; highly-aggregating, moderately aggregating, and slowly aggregating proteins. The classification is based on two calculations of the protein properties: the surface hydrophobicity of the protein (the SAP Score) and the net charge of the protein. The first factor, the SAP Score, is a measure of tendency for the protein to stick to other proteins, including itself. It is based on the spatial aggregation propensity (SAP) of the protein and is a measure of the severity and size of hydrophobic regions on the protein surface. The second measurement is the net charge of the protein, and accounts for the electrostatic repulsion between the two similarly charged proteins in solution. When combined, these two factors can rank proteins. While, this tool was originally developed for IgG1 antibodies; it has since been successfully applied to a wider range of proteins including: IgG1, IgG2, and IgG4 antibodies and globular proteins derived from antibody fragments.

The second of the *in silico* tools is an algorithm for the prediction of non-enzymatic hydrolysis of peptide bonds for bonds following a carboxylic acid containing residue. Due to the number of potential routes by which this reaction could occur, the mechanism was also investigated. When an aspartic acid residue, the residue experimentally found to hydrolyze the most quickly, is present on the N-terminal side of a peptide bond, the most energetically favorable path starts with the protonation of the peptide bond's oxygen atom. Then, a ring is formed by the sidechain bonding to the peptide bond. Later, the peptide bond breaks, and the ring opens by addition of a water molecule. This knowledge was then used to develop an algorithm that predicts if a particular peptide bond, with a carboxylic acid at the N-terminus, will react. The algorithm takes four factors into account, which directly impact the first two steps of the hydrolysis reaction, the proton addition and the cyclization of the sidechain. The first two factors controlling this reaction are the secondary structure and the surface exposure of the peptide bond; these directly impact the ability of the proton to diffuse to and add to the peptide bond. The other two factors determine the likelihood that the sidechain can participate in the reaction; these other factors are the orientation of the sidechain and the hydrogen bonding of the sidechain to its surroundings. The algorithm was trained on cleavage site data for an immunotoxin, and then applied with great accuracy, nearly 90%, for two IgG1 antibodies.

As can be seen in these two cases, *in silico* tools can be invaluable in understanding and predicting the properties of proteins in solution. These tools can be made to predict a wide range of protein properties, from the physical degradation route of aggregation to the chemical degradation route of hydrolysis. Neither of these tools require any material to make a prediction; and these tools can predict the properties of a protein that take several months to measure. The use of tools like these can increase the knowledge about a drug candidate and thereby decrease the risk of the investment in that drug.

8. ACKNOWLEDGMENTS

First, I would like to thank my friends and family whom offered encouragement, advice, and support throughout my graduate career. I would especially like to thank the past and current members of the Trout group from whom I have learned quite a bit. In particular, I would like to thank Dr. Neeraj Agrawal and Dr. Geoffrey Wood, both of whom helped not only start the aggregation and hydrolysis projects, but also taught many of the techniques I used to investigate these problems. I would also like to thank them for offered countless hours of debate and discussion on the results or lack of results throughout these projects. Additionally, I would like to thank Dr. Elise Champion and Dr. Fabienne Courtois, whom helped to experimentally investigate aggregation, and offered many insightful comments on the meaning of the experimental results.

I would like to our collaborators whom helped drive the aggregation and hydrolysis projects forward. These include Dr. Bernhard Helk from Novartis, Dr. David Farkas, Dr. Hasige Sathish, and Dr. Hardeep Samra from MedImmune whom not only helped form the project, but gathered vital results for validation and training of the various methods. I would also like to thank Novartis and MedImmune for funding these projects.

Finally, I would like to thank Professor Bernhardt Trout, my advisor, for numerous hours of insightful comments and help in focusing on the goal of the projects. I would also like to thank the members of my thesis committee, Professor Bruce Tidor and Professor K. Dane Wittrup, for the numerous insightful comments that helped advanced these projects.

9. LITERATURE CITATION

1. DiMasi, J. A.; Hansen, R. W.; Grabowski, H. G. The price of innovation: new estimates of drug development costs. *Journal of Health Economics* **2003,** *22,* 151-185.
2. DiMasi, A. J.; Grabowski, H. G. The Cost of Biopharmaceutical R&D: Is Biotech Different? *Manage. Decis. Econ.* **2007,** *28,* 469-479.
3. Adams, C. P.; Brantner, V. V. Estimating The Cost of New Drug Development: Is It Really $802 Million? *Health Affairs* **2006,** *25* (2), 420-428.
4. Zurdo, J. Developability assesment as an early de-risking tool for biopharmaceutical development. *Pharm. Bioprocess.* **2013,** *1* (1), 29-50.
5. Manning, M. C.; Patel, K.; Borchardt, R. T. Stability of Protein Pharmaceuticals. *Pharmaceutical Research* **1989,** *6* (11), 903-918.
6. Manning, M. C.; Chou, D. K.; Murphy, B. M.; Payne, R. W.; Katayama, D. S. Stability of Protein Pharmaceuticals: An Update. *Pharmaceutical Research* **2010,** *27* (4), 544-575.
7. Aggarwal, S. What's fueling the biotech engine—2007. *Nature Biotechnology* **2008,** *26* (11), 1227-1233.
8. Correia, I. R. Stability of IgG isotypes in serum. *mAbs* **2010,** *2* (3), 221-232.
9. Rosenberg, A. Effects of protein aggregates: an immunologic perspective. *AAPS J.* **2006,** *8* (3), E501-E507.
10. Fernandez-Escamilla, A.-M.; Rousseau, F.; Schymkowitz, J.; Serrano, L. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nature Biotechnology* **2004,** *22* (10), 1302-1306.
11. Tartaglia, G. G.; Vendruscolo, M. The Zyggregator method for predicting protein aggregation propensities. *Chem. Soc. Rev.* **2008,** *37,* 1395–1401.
12. Agrawal, N. J.; Kumar, S.; Wang, X.; Helk, B.; Singh, S. K.; Trout, B. L. Aggregation in Protein-Based Biotherapeutics: Computational Studies and Tools to identify aggregation prone regions. *J. Pharm Sci* **2011,** *100* (12), 5081-5095.
13. Chennamsetty, N.; Voynov, V.; Kayser, V.; Helk, B.; Trout, B. T. Design of therapeutic proteins with enhanced stability. *PNAS* **2009,** *106* (29), 11937-11942.

14. Wang, X.; Singh, S. K.; Kumar, S. Potential Aggregation-Prone Regions in Complementarity-Determining Regions of Antibodies and Their Contribution Towards Antigen Recognition: A Computational Analysis. *Pharm Res.* **2010,** *27,* 1512–1529.

15. Bowerman, C. J.; Ryan, D. M.; Nissan, D. A.; Nilsson, B. L. The effect of increasing hydrophobicity on the self-assembly of amphipathic b-sheet peptideswz. *Molecular BioSystems* **2009,** *5,* 1058–1069.

16. Chennamsetty, N.; Helk, B.; Voynov, V.; Kayser, V.; Trout, B. L. Aggregation-Prone Motifs in Human Immunoglobulin G. *J. Mol. Biol.* **2009,** *391,* 404-413.

17. Chennamsetty, N.; Voynov, V.; Kayser, V.; Helk, B.; Trout, B. T. Prediction of Aggregation Prone Regions of Therapeutic Proteins. *J. Phys. Chem.* **2010,** *114,* 6614–6624.

18. Voynov, V.; Chennamsetty, N.; Kayser, V.; Helk, B.; Trout, B. L. Predictive tools for stabilization of therapeutic proteins. *mAbs* **2009,** *1* (6), 580-582.

19. Voynov, V.; Chennamsetty, N.; Kayser, V.; Wallny, H.-J.; Helk, B.; Trout, B. L. Design and Application of Antibody Cysteine Variants. *Bioconjugate Chem.* **2010,** *21,* 385-392.

20. Whitelegg, N. R. J.; Rees, A. R. WAM: an improved algorithm for modelling antibodies on the WEB. *Protein Engineering* **2000,** *13* (12), 819-824.

21. Marcatili, P.; Rosi, A.; Tramontano, A. PIGS: automatic prediction of antibody structures. *Bioinformatics* **2008,** *24* (17), 1953–1954.

22. Sivasubramanian, A.; Sircar, A.; Chaudhury, S.; Gray, J. J. Toward high-resolution homology modeling of antibody Fv regions and application to antibody–antigen docking. *Proteins* **2009,** *74,* 497–514.

23. Roberts, C. J. Kinetics of Irreversible Protein Aggregation: Analysis of Extended Lumry-Eyring Models and Implications for Predicting Protein Shelf Life. *J. Phys. Chem.* **2003,** *107,* 1194-1207.

24. Saphire, E. O.; Parren, P. W. H. I.; Pantophlet, R.; Zwick, M. B.; Morris, G. M.; Rudd, P. M.; Dwek, R. A.; Stanfiel, R. L.; Burton, D. R.; Wilson, I. A. Crystal Structure of a Neutralizing Human IgG Against HIV-1: A Template for Vaccine Design. *Science* **2001,** *293,* 1155-1159.

25. Sun, Z.; Almogren, A.; Furtado, P. B.; Chowdhury, B.; Kerr, M. A.; Perkins, S. J. Semi-extended Solution Structure of Human Myeloma Immunoglobulin D Determined by Constrained X-ray Scattering. *Journal of Molecular Biology* **2005,** No. 353, 155-173.

26. Corper, A. L.; Sohi, M. K.; Bonagura, V. R.; Steinitz, M.; Jefferis, R.; Feinstein, A.; Beale, D.; Taussig, M. J.; Sutton, B. J. Structure of human IgM rheumatoid factor Fab bound to its autoantigen IgG Fc reveals a novel topology of antibody-antigen interaction. *Nature Structural Biology* **1997,** *4* (5), 374-381.

27. Deisenhofer, J. Crystallographic Refinement and Atomic Models of a Human Fc Fragment and Its Complex with Fragment B of Protein A from Staphylococcus aureus at 2.9- and 2.8-A Resolution. *Biochemistry* **1981,** *20* (9), 2361-2370.

28. Black, S. D.; Mould, D. R. Development of Hydrophobicity Parameters to Analyze Proteins Which Bear Post-or Cotranslational Modifications. *Analytical Biochemistry* **1991,** *193,* 72-82.

29. Brooks, B.; et al. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Comput Chem* **1983,** *4,* 187-217.

30. Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. Scalable Molecular Dynamics with NAMD. *J Comput Chem* **2005,** *26,* 1781-1802.

31. MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem.* **1998,** *102,* 3586-3616.

32. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chern. Phys.* **1983,** *79* (2), 926-935.

33. Bas, D. C.; Rogers, D. M.; Jensen, J. H. Very fast prediction and rationalization of pKa values for protein-ligand complexes. *Proteins* **2008,** *73,* 765-783.

34. Li, H.; Robertson, A. D.; Jensen , J. H. Very Fast Empirical Prediction and Rationalization of Protein pKa Values. *Proteins* **2005,** *61,* 704-721.

35. Chari, R.; Jerath, K.; Badkar, A. V.; Kalonia, D. S. Long- and Short-Range Electrostatic Interactions Affect the Rheology of Highly Concentrated Antibody Solutions. *Pharmaceutical Research* **2009,** *26* (12), 2607-2618.

36. Lehermayr, C.; Mahler, H. C.; Maeder, K. Assessment of Net Charge and Protein-Protein Interactions of Different Monoclonal Antibodies. *Journal of Pharmaceutical Sciences* **2011,** *100,* 2551-2562.

37. Ishikawa, T.; Ito, T.; Endo, R.; Nakagawa, K.; Sawa, E.; Wakamatsu, K. Influence of pH on Heat-Induced Aggregation and Degradation of Therapeutic Monoclonal Antibodies. *Biol. Pharm. Bull.* **2010,** *33* (8), 1413-1417.

38. Franey, H.; Brych, S. R.; Kolvenbach, C. G.; Rajan, R. S. Increased aggregation propensity of IgG2 subclass over IgG1: Role of conformational changes and covalent character in isolated aggregates. *Protein Science* **2010,** *19,* 1601-1615.

39. Hari, S. B.; Lau, H.; Razinkov, V. I.; Chen, S.; Latypov, R. F. Acid-Induced Aggregation of Human Monoclonal IgG1 and IgG2: Molecular Mechanism and the Effect of Solution Composition. *Biochemistry 49,* 9328-9338.

40. Wang, W.; Singh, S.; Zeng, D.; King, K.; Nema, S. Antibody structure, instability, and formulation. *J Pharm Sci 440* (1), 83-98.

41. ICH Guideline. *Q6B: Specifications: Test Procedures and Acceptance Criteria for Biotechnological/Biological Products;*, 1999.

42. Wang, W.; Singh, S.; Li, N.; Toler, M.; King, K.; Nema, S. Immunogenicity of protein aggregates--concerns and realities. *Int J Pharm* **2012,** *431* (1-2), 1-11.

43. Chi, E.; Krishnan, S.; Randolph, T.; Carpenter, J. Physical stability of proteins in aqueous solution: mechanism and driving forces in nonnative protein aggregation. *Pharm Res* **2003,** *20* (9), 1325-1336.

44. Saito, S.; Hasegawa, J.; Kobayashi, N.; Tomitsuka, T.; Uchiyama, S.; Fukui, K. Effects of ionic strength and sugars on the aggregation propensity of monoclonal antibodies: influence of colloidal and conformational stabilities. *Pharm Res* **2013,** *30* (5), 1263-1280.

45. Vázquez-Rey, M.; Lang, D. Aggregates in monoclonal antibody manufacturing processes. *Biotechnol Bioeng* **2011,** *108* (7), 1494-1508.

46. Kameoka, D.; Masuzaki, E.; Ueda, T.; Imoto, T. Effect of buffer species on the unfolding and the aggregation of humanized IgG. *J Biochem* **2007,** *142* (3), 383-391.

47. Feng, Y.; Ooishi, A.; Honda, S. Aggregation factor analysis for protein formulation by a systematic approach using FTIR, SEC and design of experiments techniques. *J Pharm BIomed Anal* **2012,** *57,* 143-152.

48. Bhambhani, A.; Kissmann, J.; Joshi, A.; Volkin, D.; Kashi, R.; Middaugh, C. Formulation design and high-throughput excipient selection based on structural integrity and conformational stability of dilute and highly concentrated IgG1 monoclonal antibody solutions. *J Pharm Sci* **2012,** *101* (3), 1120-1135.

49. Zhang-van Enk, J.; Mason, B.; Yu, L.; Zhang, L.; Hamouda, W.; Huang, G.; al, e. Perturbation of thermal unfolding and aggregation of human IgG1 Fc fragment by Hofmeister anions. *Mol Pharm* **2013,** *10* (2), 619-630.

50. Kim, N.; Remmele, R.; Liu, D.; Raazinkov, V.; Fernandez, E.; Roberts, C. Aggregation of anti-streptavidin immunoglobulin gamma-1 involves Fab unfolding and competing growth pathways mediated by pH and salt concentration. *Biophys Chem* **2013,** *172,* 26-36.

51. Buchner, J.; Renner, M.; Lilie, H.; Hinz, H.; Jaeninke, R.; Kiefhabel, T.; et al. Alternatively folded states of an immunoglobulin. *Biochemistry* **1991,** *30* (28), 6922-6929.

52. Lilie , H.; Buchner, J. Domain interactions stabilize the alternatively folded state of an antibody Fab fragment. *FEBS Lett.* **1995,** *362* (1), 43-46.

53. Thies, M.; Kammermeier, R.; Richter, K.; Buchner, J. The alternatively folded state of the C(H)3 domain. *J Mol Bio* **2001,** *309* (5), 1077-1086.

54. Feige, M.; Simpson, E.; Herold, E.; Bepperling, A.; Heger, K.; Buchner, J. Dissecting the alternatively folded state of the antibody Fab fragment. *J Mol Biol* **2010,** *339* (5), 719-730.

55. Corisdeo, S.; Wang, B. Functional experssion and display of an antibody Fab fragment in Escherichia coli: study of vector designs and culture conditions. *Protein Expr Purif* **2004,** *34* (2), 270-279.

56. Kim, J.; Iyer, V.; Joshi, S.; Volkin, D.; Middaugh, C. Improved data visualization techniques for analysing macromolecule structure changes. *Protein Sci* **2012,** *21* (10), 1540-1553.

57. Hendler, R.; Shrager, R. Deconvolutions based on singular value decompositions and the pseudoinverse: a guide for beginners. *J Biochem Biophys Methods* **1994,** *28* (1), 1-33.

58. Goh, C.; Lan, N.; Douglas, S.; Wu, B.; Echols, N.; Smith, A.; et al. Mining the structural genomics pipeline: identification of protein properties that affect high-throughput experimental analysis. *J Mol Biol* **2004,** *336* (1), 115-130.

59. Idicula-Thomas, S.; Balaji, P. Understanding the relationship between primary structure of proteins and its propensity to be soluble on overexpression in Escherichia coli. *Protein Sci* **2005,** *14* (3), 582-592.

60. Conchillo-Sole, O.; de Groot, N.; Aviles, F.; Vendrell, J.; Daura, X.; Ventura, S. Aggrescan: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. *BMC Bioinformatics* **2007,** *8,* 65.

61. Trovato, A.; Chiti, F.; Marian, A.; Seno, F. Insight into the structure of amyloid fibrils from the analysis of globular proteins. *Plos Computational Biology* **2006,** *2* (12), 1608-1618.

62. Lauer, T. M.; Agrawal, N. J.; Chennamsetty, N.; Egodage, K.; Helk, B.; Trout, B. L. Developability Index: A Rapid in silico Tool for the Screening of Antibody Aggregation Propensity. *J. Phar. Sci* **2012,** *101* (1), 102-115.

63. Cho, H.-S.; Mason, K.; Ramyar, K. X.; Stanley, A. M.; Gabelli, S. B.; Denney, D. W.; Leahy, D. J. Structure of the extracellular region of HER2 alone and in complex Herceptin Fab. *Nature* **2003,** *421,* 756-760.

64. Moiani, D.; Salvalaglio, M.; Cavallotti, C.; Bujacz, A.; Redzynia, I.; Bujacz, G.; Dinon, F.; Pengo, P.; Fassina, G. Structural Characterization of a Protein A Mimetic Peptide Dendrimer Bound to Human IgG. *J. Phys. Chem. B.* **2009,** *113* (50), 16268-16275.

65. Olsson, M. H. M.; Sondergaard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *J. Chem. Theory Comput.* **2011,** *7* (2), 525-537.

66. Tartaglia, G. G.; Cavalli, A.; Pellarin, R.; Caflisch, A. Prediction of aggregation rate and aggregation-prone segments in polypeptide sequences. *Protein Science* **2005,** *14,* 2723-2734.

67. Zhang, A.; Singh, S. K.; Shirts, M. R.; Kumar, S.; Fernandez, E. J. Distinct Aggregation Mechanisms of Moloclonal Antibody Under Thermal and Freeze-Thaw Stresses Revealed by Hydrogen Exchange. *Pharm Res* **2012,** *29,* 236-250.

68. Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids;* Clarendon Press: Oxford, 1987.

69. Zhu, J.; Ofek, G.; Yang, Y.; Zhang, B.; Louder, M. K.; Lu, G.; McKee, K.; Pancera, M.; Skinner, J.; Zhang, Z.; Parks, R.; Eudailey, J.; Lloyd, K. E.; Blinn, J.; Alam, S. M.; Haynes, B. F.; Simek, M.; Burton, D. R.; Koff, W. C.; NISC Comparative Sequencing Program; Mullikin, J. C.; Mascola, J. R.; Shapiro, L.; Kwong, P. D. Mining the antibodyome for HIV-1–neutralizing antibodies with next-generation sequencing and phylogenetic pairing of heavy/light chains. *PNAS* **2012,** *110* (16), 6470-6475.

70. Huang, J.; Ofek, G.; Laub, L.; Louder, M. K.; Doria-Rose, N. A.; Longo, N. S.; Imamichi, H.; Bailer, R. T.; Chakrabarti, B.; Sharma, S. K.; Alam, S. M.; Wang, T.; Yang, Y.; Zhang, B.; Mingueles, S. A.; Wyatt, R.; Haynes, B. F.; Kwong, P. D.; Mascola, J. R.; Connors, M. Broad and potent neutralization of HIV-1 by a gp41-specific human antibody. *Nature* **2012,** *491,* 406-412.

71. Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **2008,** *4* (3), 435-447.

72. Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters. *Proteins: Structure, Function, and Bioinformatics* **2006,** *65,* 712-725.

73. Aggarwal, R. What's fueling the biotech engine-2012 to 2013. *Nature Biotechnology* **2014,** *32,* 32-39.

74. Joshi, V.; Shivach, T.; Kumar, V.; Yadav, N.; Rathore, A. Avoiding antibody aggregation during processing: estabilishing hold times. *Biotechnol J* **2014,** *9,* 1195-1205.

75. Ratanji, K.; Derrick, J.; Dearman, R.; Kimber, I. Immunogenicity of therapeutic proteins: Influence of aggregation. *J Immunotoxicol* **2014,** *11,* 99-109.

76. Shire, S. Formulation and manufacturability of biologics. *Curr Opin Biotechnol* **2009,** *20,* 708-714.

77. Roque, C.; Sheung, A.; Rahman, N.; Ausar, S. Effect of polyethylene glycol conjugation on conformational and colloidal stability of a monoclonal antibody antigen-binding fragment (Fab'). *Mol Pharm* **2015,** *12,* 562-575.

78. Kayser, V.; Chennamsetty, N.; Voynov, V.; Forrer, K.; Helk, B.; Trout, B. L. Glycosylation influences on the aggregation propensity of therapeutic monoclonal antibodies. *Biotechnol J* **2011,** *6,* 38-44.

79. Perchiacca, J.; Tessier, P. Engineering Aggregation-Resistant Antibodies. *Annu Rev Chem Biomol Eng* **2012,** *3,* 263-286.

80. Wang, X.; Das, T. K.; Singh, S. K.; Kumar, S. Potential aggregation prone regions in biotherapeutics: A survey of commercial monoclonal antibodies. *mAbs* **2009,** *1,* 254-267.

81. Hurwitz, H.; Fehrenbacher, L.; Novotny, W.; Cartwright, T.; Hainsworth, J.; Heim, W.; Berlin, J.; Baron, A.; Griffing, S.; Holmgren, E.; Ferrara, N.; Fyfe, G.; Rogers, B.; Ross, R.; Kabbinavar, F. Bevacizumab plus irinotecan, fluorouracil, and leucovorin for metastatic colorectal cancer. *N Engl J Med* **2004,** *350,* 2335-2342.

82. Los M; Roodhart, J.; Voest, E. Target Practice: Lessons from Phase III Trials with Bevacizumab and Vatalanib in the Treatment of Advanced Colorectal Cancer. *Oncologist* **2007,** *12,* 443-450.

83. Ciulla, T. A.; Rosenfeld, P. J. Antivascular endothelial growth factor therapy for neovascular age-related macular degeneration. *Curr Opin Opthalmol* **2009,** *20,* 158-165.

84. Paul, M.; Lahlou, A.; Carvalho, M.; Blanchet, B.; Astier A. Thermal stability of two monoclonal antibodies: cetuximab and bevacizumab. *Eur J Oncol Pharm* **2008,** *2,* 37.

85. Oliva, A.; Llabres, M.; Farina, J. B. Capability measurement of size-exclusion chromatography with a light-scattering detection method in a stability study of bevacizumab using capability indices. *J Chromatogr A* **2014,** *1353,* 89-98.

86. Cromwell, M.; Gazzano-Santoro, H. Protein Aggregation and Potency. http://www.iirusa.com/upload/wysiwyg/P1198_Images/IIR_P1198_Cromwell.pdf.

87. Latypov, R. F.; Hogan, S.; Lau, H.; Gadgil, H.; Liu, D. Elucidation of Acid-induced Unfolding and Aggregation of Human Immunoglobulin IgG1 and IgG2 FC. *J Biol Chem* **2012,** *287,* 1381-1396.

88. Wang, X.; Kumar, S.; Buck, P. M.; Singh, S. K. Impact of deglycosylation and thermal stress on conformational stability of a full length murine IgG2a monoclonal antibody: Observations from molecular dynamics Simulations. *Proteins* **2013,** *81,* 443-460.

89. Li, C. H.; Narhi, L. O.; Wen, J.; Dimitrova, M.; Wen, Z.; Li, J.; Pollastrini, J.; Nguyen, X.; Tsuruda, T.; Jiang, Y. Effect of pH, Temperature, and Salt on the Stability of Escherichia coli- and Chinese Hamster Ovary Cell-Derived IgG1 Fc. *Biochem* **2012,** *51,* 10056-10065.

90. Zheng, K.; Bantog, C.; Bayer, R. The impact of glycosylation on monoclonal antibody conformation and stability. *mAbs* **2011,** *3,* 568-576.

91. Sinclair, A. M.; Elliott, S. Glycoengineering: the effect of glycosylation on the properties of therapeutic proteins. *J Pharm Sci* **2005,** *94,* 1626-1635.

92. Pepinsky, R. B.; Silvian, L.; Berkowitz, S. A.; Farrington, G.; Lugovskoy, A.; Walus, L.; Eldredge, J.; Capili, A.; Mi, S.; Graff, C.; Garber, E. Improving the solubility of anti-LINGO-1 monoclonal antibody Li33 by isotype switching and targeted mutagenesis. *Protein Science* **2010,** *19,* 954-966.

93. Voynov, V.; Chennamsetty, N.; Kayser, V.; Helk, B.; Forrer, K.; Zhang, H.; Fritsch, C.; Heine, H.; Trout, B. L. Dynamic fluctuations of protein-carbohydrate interactions promote protein aggregation. *PLos One* **2009,** *4,* e8425.

94. Muller, Y.; Chen, Y.; Christinger, H.; Li, B.; Cunningham, B.; Lowman, H.; de Vos, A. VEGF and the Fab fragment of a humanized neutralizing antibody: crystal structure of the complex at 2.4 A resolution and mutational analysis of the interface. *Structure* **1998,** *6,* 1153-1167.

95. Humphrey, W.; Dalke, A.; Schulten, K. VMD - Visual Molecular Dynamics. *J. Molec. Graphics.* **1996,** *14,* 33-38.

96. Woods Group. GLYCAM Web. http://www.glycam.org (accessed May 10, 2015).

97. Case, D. A.; Darden, T. A.; Cheatham, T. E. I.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhange, W.; Merz, K. M.; al, e. *Amber 12.*

98. Kirschner, K. N.; Yongye, A. B.; Tschampel, S. M.; Gonzalez-Outeiriño, J.; Daniels, C. R.; Foley, B. L.; Woods, R. J. GLYCAM06: A generalizable biomolecular force field for Carbohydrates. *J. Comput. Chem.* **2008,** *29,* 622-655.

99. Berendsen, H. J.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984,** *81,* 3684.

100. Agrawal, N.; Helk, B.; Trout, B. A computational tool to predict the evolutionarily conserved protein-protein interaction hot-spot residues from the structure of the unbound protein. *FEBS Letters* **2014,** *588,* 326-333.

101. Jenkins, N. Modifications of therapeutic proteins: challenges and prospects. *Cytotechnology* **2007,** *53,* 121-125.

102. Arnold, J. N.; Wormald, M. R.; Sim, R. B.; Rudd, P. M.; Dwek, R. A. The Impact of Glycosylation on the Biological Function and Structure of Human Immunoglobulings. *Annu Rev Immunol* **2007,** *25,* 21-50.

103. Hristodorov, D.; Fischer, R.; Linden, L. With or Without Sugar? (A)glycosylation of Therapeutic Antibodies. *Mol Biotechnol* **2013,** *54,* 1056-1068.

104. Hari, S. B.; Lau, H.; Razinkov, V. I.; Chen, S.; Latypov, R. F. Acid-induced aggregation of human monoclonal IgG1 and IgG2: molecular mechanism and the effect of solution composition. *Biochemistry* **2010,** *49,* 9328-9338.

105. Hristodorov, D.; Fischer, R.; Joerissen, H.; Muller-Tiemann, B.; Apeler, H.; Linden, L. Generation and Comparative Characterization of Glycosylated and Aglycosylated Human IgG1 Antibodies. *Mol Biotechnol* **2013,** *53,* 326-335.

106. Wu, S. J.; Luo, J.; O'Neil, K. T.; Kang, J.; Lacy, E. R.; Canziani, G.; Baker, A.; Huang, M.; Tang, Q. M.; Raju, T. S.; Jacobs, S. A.; Teplyakov, A.; Gilliland, G. L.; Feng, Y. Structure-based engineering of a monoclonal antibody for improved solutbility. *Protein Eng Des Sol* **2010,** *23,* 643-651.

107. Mattu, T. S.; Pleass, R. J.; Willas, A. C.; Kilian, M.; Wormald, M. R.; Lellouch, A. C.; Rudd, P. M.; Woof, J. M.; Dwek, R. A. The glycosylation and structure of human serum IgA1, Fab, and Fc regions and the role of N-glycosylation on Fc alpha receptor interactions. *J Biol Chel* **1998,** *273,* 2260-2272.

108. Jefferies, R. Glycosylation as a strategy to improve antibody0based therapeutics. *Nat Rev Drug Discov* **2009,** *8,* 226-234.

109. Qian, J.; Liu, T.; Yang, L.; Daus, A.; Crowley, R.; Zhou, Q. Structural characterization of N-linked oligosacchatieds on monoclonal antibody cetuximab by the combination of orthogonal matrix-assisted laser desorption/ionization hybrid quadrupole time-of-flight tandem mass spectrometry and sequential enzymatic diges. *Anal Biochem* **2007,** *364,* 8-18.

110. Beck, A. Biosimilar, biobetter and next generation therapeutic antibodies. *mAbs* **2011,** *3,* 107-110.

111. Testa, B.; Mayer, J. M. *Hydrolysis in Drug and Prodrug Metabolism;* VHCA: Zurich, Switzerland, 2003.

112. Marcus, F. Preferential cleavage at aspartyl-prolyl peptide bonds in dilute acid. *Int. J. Peptide Protein Res.* **1985,** *25,* 542-546.

113. Singh, S. Impact of Product-Related Factors on Immunogenicity of Biotherapeutics. *J Pharm Sci* **2011,** *100* (2), 354-387.

114. Kamerzell, T. J.; Li, M.; Arora, S.; Ji, J. A.; Wang, J. The Relative Rate of Immunoglobulin Gamma 1 Fragmentation. *J Pharm Sci* **2011,** *100,* 1341-1349.

115. Van Buren, N.; Rehder, D.; Gadgil, H.; Matsumura, M.; Jacob, J. Elucidation of Two Major Aggregation Pathways in an IgG2 Antibody. *Journal of Pharmaceutical Sciences* **2009,** *98* (2).

116. Liu, H.; Gaza-Bulseco, G. Fragmentation of a Recombinant Monoclonal Antibody at Various pH. *Pharmaceutical Research* **2008,** *25* (8), 1881-1890.

117. Kahne, D.; Still, W. C. Hydrolysis of a Peptide Bond in Neutral Water. *J. Am. Chem. Soc.* **1988,** *110,* 7529-7534.

118. Cordoba, A. J.; Shyong, B.-J.; Breen, D.; Harris, R. J. Non-enzymatic hinge region fragmentation of antibodies in solution. *J Chrom B* **2005,** *818,* 115-121.

119. Bryant, R. A. R.; Hansen, D. E. Direct Measurement of the Uncatalyzed Rate of Hydrolysis of a Peptide Bond. *J. Am. Chem. Sic.* **1996,** *118,* 5498-5499.

120. Wang, B.; Cao, Z. Mechanism of Acid-Catalyzed Hydrolysis of Formamide from Cluster-Continuum Model Calculations: Concerted versus Stepwise Pathway. *J. Phys. Chem. A* **2010,** *114,* 12918-12927.

121. Krug, J. P.; Popelier, P. L. A.; Bader, R. F. W. Theoretical Study of Neutral and of Acid and Base Promoted Hydrolysis of Formamide. *The Journal of Physical Chemistry* **1992,** *96,* 7604-7616.

122. Gorb, L.; Asensio, A.; Tunon, I.; Ruiz-Lopez, M. F. The Mechanism of Formamide Hydrolysis in Water from Ab Initio Calculations and Simulation. *Chem. Eur. J.* **2005,** *11,* 6743-6753.

123. Radzicka, A.; Wolfenden, R. Rates of Uncatalyzed Peptide Bond Hydrolysis in Neutral Solution and the Transition State Affinities of Proteases. *J. Am. Chem. Soc.* **1996,** *118,* 6105-5109.

124. Antonczak, S.; Ruiz-Lopez, M. F.; Rivail, J. L. Ab Initio Analysis of Water-Assisted Reaction Mechanisms in Amide Hydrolysis. *J. Am. Chem. Soc.* **1994,** *116,* 3912-3921.

125. Pan, B.; Ricci, M. S.; Trout, B. L. A Molecular Mechanism of Hydrolysis of Peptide Bonds at Neutral pH Using a Model Compound. *J. Phys. Chem. B* **2011,** *115,* 5958-5970.

126. Hehre, W. J.; Radom, L.; Schleyer, P. v. P.; Pople, J. A. *Ab Initio Molecular Orbital Theory;* Wiley: New York, 1986.

127. Koch, W.; Holthausen, M. C. *A Chemist's Guide to Density Functional Theory;* Wiley-VCH: Weinheim, 2000.

128. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. . J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Pople, J. A. *Gaussian 03, Revision D.01;* Gaussian Inc.: Wallingford, 2004, 2004.

129. Shao, Y.; Gan, Z.; Epifanovsky, E.; Gilbert, A. T. B.; Wormit, M.; Kussmann, J.; Lange, A. W.; Behn, A.; Deng, J.; Feng, X.; Ghosh, D.; Horn, M. G. P. R.; Jacobson, L. D.; Kaliman, I.; Gill, P. M. W.; Head-Gordon, M. Advances in molecular quantum chemistry contained in the Q-Chem 4 program package. *Molecular Physics* **2015,** *113* (2), 184-215.

130. Becke, A. A new mixing of Hartree–Fock and local density-functional theories. *J. Chem Phys* **1993,** *98,* 1372-1378.

131. Barone, V.; Cossi, M. Quantum calculation of molecular energies and energy gradients in solution by a conductor solvent model. *J. Phys. Chem. A* **1998,** *102,* 1995-2001.

132. Gonzalez, C.; Schlegel, H. B. An Improved Algorithm for Reaction-Path Following. *J. Chem. Phys.* **1989,** *30,* 2154-2161.

133. Gonzalez, C.; Schlegel, H. B. Reaction-Path Following in Mass-Weighted Internal Coordinates. *J. Phys. Chem.* **1990,** *94,* 5523-5527.

134. Chai, J.-D.; Head-Gordon, M. Systematic optimization of long-range corrected hybrid density functionals. *J. Chem. Phys.* **2008,** *2008,* 084106.

135. Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Generalized Born Solvation Model SM12. *J. Chem. Theory Comput.* **2013,** *9* (1), 609-620.

136. Wood, G. P. F.; Santiso, E. E.; Trout, B. L. Assessment of a Genotype Meta-Heuristic for Cluster Optimizations. *In production.*

137. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004,** *25* (9), 1157-1174.

138. Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993,** *98,* 10089.

139. Arnett, K. L.; Harrison, S. C.; Wiley, D. C. Crystal structure of human CD3-e/d dimer in complex with UCHT1 single-chain antibody fragment. *Proc. Natl. Acad. Sci. USA* **2004,** *101,* 16268-16273.

140. Wedekind, J. E.; Trame, C. B.; Dorywalska, M.; Koehl, P.; Raschke, T. M.; McKee, M.; FitzGerald, D.; Collier, R. J.; Mckay, D. B. Refined crystallographic structure of Pseudomonas aeruginosa exotoxin A and its implications for the molecular mechanism of toxicity. *J. Mol. Biol. 314,* 823-837.

141. Lee, J. E.; Kuehne, A.; Abelson, D. M.; Fusco, M. L.; Hart, M. K.; Saphire, E. O. Complex of a protective antibody with its Ebola virus GP peptide epitope: unusual features of a V lambda x light chain. *J. Mol. Biol.* **2008,** *375,* 202-216.

142. Lowe, D. C.; Gerhardt, S.; Ward, A.; Hargreaves, D.; Anderson, M.; Ferraro, F.; Pauptit, R. A.; Pattison, D. V.; Buchanan, C.; Popovic, B.; Finch, D. K.; Wilkinson, T.; Sleeman, M.; Vaughan, T. J.; Mallinder, P. R. Engineering a high-affinity anti-IL-15 antibody: crystal structure reveals an α-helix in VH CDR3 as key component of paratope. *J. Mol. Biol.* **2011,** *406,* 160-175.

143. Guan, Y.; DeVico, A. L.; Lewis, G. K.; Pazgier, M. Crystal structure of human N12-i2 Fab, an ADCC and neutralizing anti-HIV-1 Env antibody. *To be published.*

144. Li, S.; Wang, H.; Peng, B.; Zhang, M.; Zhang, D.; Hou, S.; Guo, Y.; Ding, J. Efalizumab binding to the LFA-1 alphaL I domain blocks ICAM-1 binding via steric hindrance. *PNAS* **2009,** *106,* 4349-4354.

145. Blech, M.; Peter, D.; Fischer, P.; Bauer, M. M.; Hafner, M.; Zeeb, M.; Nar, H. One target-two different binding modes: structural insights into gevokizumab and canakinumab interactions to interleukin-1β. *J. Mol. Biol.* **2013,** *425,* 94-111.

146. The MathWorks. *MATLAB 2015a;* Natick, MA, 2015.

147. Berens, P. CircStat: A Matlab Toolbox for Circular Statistics. *J. Stat. Soft.* **2009,** *31* (10), 1-21.

148. Frishman, D.; Argos, P. Knowledge-based Secondary Structure Assignment. *Proteins: Structure, Function, Genetics* **1995,** *23,* 566-579.

149. Bell, J. *Machine Learning: Hands-On for Developers and Technical Professionals,* 1st ed.; John Wiley & Sons, Inc.: Indianapolis, IN, 2015.

150. Robinson, N. E.; Robinson, A. B. *Molecular Clocks Deamidation of Asparaginyl and Glutaminyl Residues in Peptides and Proteins;* Althouse Press: Cave Junction, 2004.

151. Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A* **1988,** *38,* 3098.

152. Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron-Density. *Phys. Rev. B. Condens. Matter* **1988,** *37,* 785-789.

153. Hamprecht, F. A.; Cohen, A. J.; Tozer, D. J.; Handy, N. C. Development and assessment of new exchange-correlation functionals. *J. Chem. Phys.* **1998,** *109,* 6264-6271.

154. Ernzerhof, M.; Scuseria, G. E. Assessment of the Perdew-Burke-Ernzerhof Exchange-Correlation Functional. *J. Chem. Phys.* **1999,** *110,* 5029-5036.

155. Adamo, C.; Barone, V. Toward Reliable Density Functional Methods without Adjustable Parameters: The PBE0 Model. *J. Chem. Phys.* **1999,** *110,* 6158-6170.

156. Boese, A. D.; Martin, J. M. L. Development of Density Functionals for Thermochemical Kinetics. *J. Chem. Phys.* **2004,** *121,* 3405-3416.

157. Zhoa, Y.; Truhlar, D. G. The M06 Suite of Density Functionals for Main Group Thermochemistry, Thermochemical Kinetics, Noncovalent Interactions, Excited States,

and Transition Elements: Two New Functionals and Systematic Testing of Four M06-Class Functionals and 12 Other Function. *Theor. Chem. Acc.* **2008,** *120,* 215-241.

158. Tao, J.; Perdew, P.; Staroverov, V. N.; Scuseria, G. E. Climbing the Density Functional Ladder: Nonempirical Meta-Generalized Gradient Approximation Designed for Molecules and Solids. *Phys. Rev. Lett.* **2003,** *91,* 146401.

159. Peverati, R.; Truhlar, D. G. Improving the Accuracy of Hybrid Meta-GGA Density Functionals by Range Separation. *J. Phys. Chem. Lett.* **2011,** *2* (21), 2810-2817.

160. Henderson, T. M.; Janesko, B. G.; Scuseria, G. E. Generalized gradient approximation model exchange holes for range-separated hybrids. *J. Chem. Phys.* **2008,** *128,* 194105.

161. Heyd, J.; Scuseria, G. E.; Ernzerhof, M. Hybrid functionals based on a screened Coulomb potential. *J. Chem. Phys.* **2003,** *118,* 8207.

162. Chai, J.-D.; Head-Gordon, M. Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections. *Phys. Chem. Chem. Phys.* **2008,** *10,* 6615-6620.

163. Grimme, S. Semiempirical GGA-Type Density Functional Constructed. *J. Comp. Chem.* **2006,** *27* (15), 1787-1799.

164. Chai, J.-D.; Head-Gordon, M. Long-range corrected double-hybrid density functionals. *J. Chem. Phys.* **2009,** *131,* 174105.

165. Purvis, G. D.; Bartlett, R. J. A full coupled-cluster singles and doubles model: The inclusion of disconnected triples. *J. Chem. Phys.* **1982,** *76,* 1910.

166. Slater, J. C. A Simplification of the Hartree-Fock Method. *Phys. Rev.* **1951,** *81,* 385.

167. Møller, C.; Plesset, M. S. Note on an Approximation Treatment for Many-Electron Systems. *Phys. Rev.* **1934,** *46,* 618.

168. Feyereisen, M.; Fitzgerald, G.; Komornicki, A. Use of approximate integrals in ab initio theory. An application in MP2 energy calculations. *Chem. Phys. Lett.* **1993,** *208* (5-6), 359-363.

169. Ditchfield, R.; Hehre, W. J.; Pople, J. A. Self-Consistent Molecular Orbital Methods. 9. Extended Gaussian-type basis for molecular-orbital studies of organic molecules. *J. Chem. Phys.* **1971,** *54,* 724.

170. Frisch, M. J.; Pople, J. A.; Binkley, J. S. Self-Consistent Molecular Orbital Methods. 25. Supplementary Functions for Gaussian Basis Sets. *J. Chem. Phys.* **1984,** *80,* 3265-3269.

171. Bryantsev, V. S.; Diallo, M. S.; Goddard III, W. A. Calculation of Solvation Free Energies of Charged Solutes Using Mixed Cluster/Continuum Models. *J. Phys. Chem. B.* **2008,** No. 112, 9709-9719.

10. APPENDIX A: ADDITIONAL INFORMATION ON DEVELOPMENT OF THE DEVELOPABILITY INDEX

10.1. Effect of SAP Radius on DI and classifications:

Previous work using SAP had found that aggregation prone regions can be found using both 10 Å and 5 Å. As these works focused on visually finding regions that are aggregation prone, not attempting to quantify how aggregation prone, several radii were attempted. These initially included 5 Å, 10 Å, and 15 Å. As 5 Å was found to be the best, several other radii were attempted around 5 Å, these include: 2 Å, 4 Å, 6 Å, and 7 Å. One interesting effect of the variation of the SAP radius is that the relative values for the SAP Score also changed. The behavior of the SAP Score is the same for all mAbs as the SAP radius is varied. It depends on the size and magnitude of the hydrophobic patches.

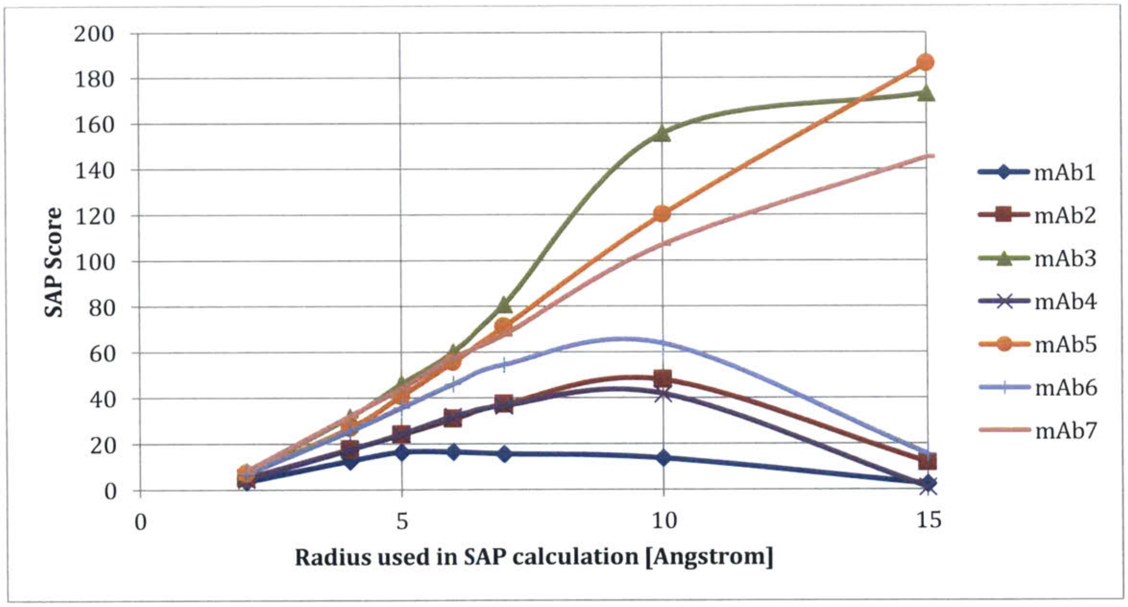

Figure 10-1 - SAP Score at different SAP Radii for 7 mAbs.

In order to determine which radii is the best, each was used in a data regression separately. The 40 °C data was used at it showed the most variance in aggregation propensities. The number of correctly classified antibodies was used to quantify which of the radii was the best. After using each radius in the DI procedure, a range of radii appeared to work equally well, from 4 Å to 7 Å. 5 Å was selected from this range; other values in the range would also work.

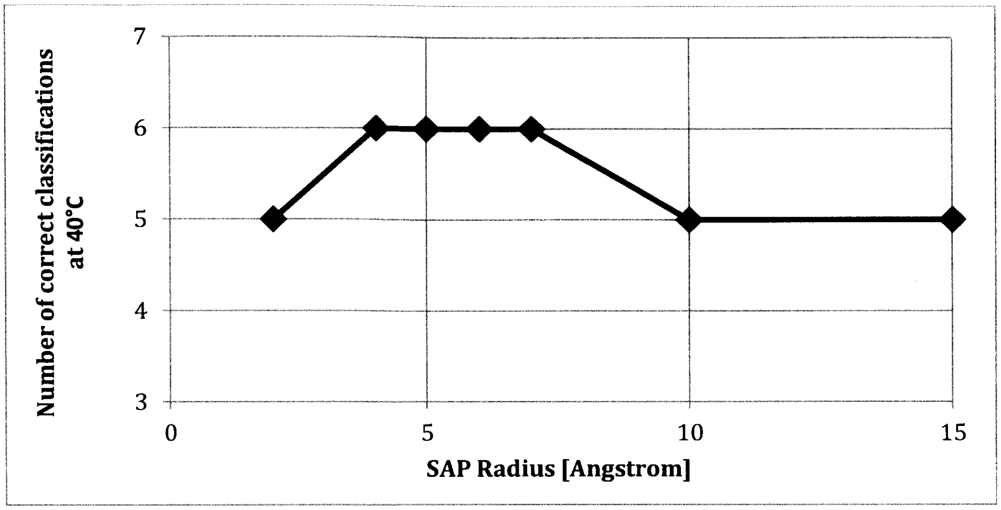

Figure 10-2 - Number of correct classifications (based on 7 data points) at 40°C at several SAP Radii.

10.2. Variations in the SAP Score cutoff value

During the calculation of the SAP Score, there is a step where only a portion of the atoms are considered. Originally the cutoff for the SAP value was 0. This cutoff was varied in an attempt to determine its effect on the final DI fittings. This was done is a similar manner to the variations in SAP radius.

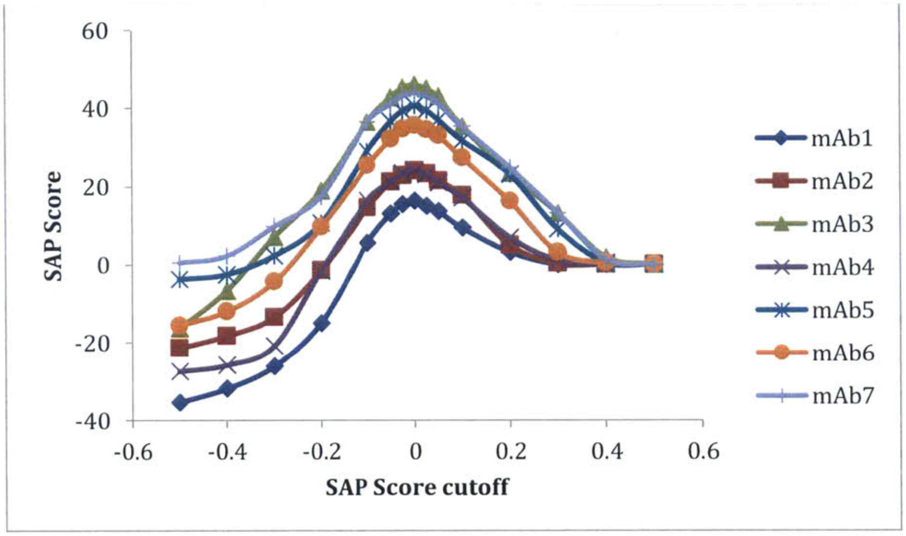

Figure 10-3 - SAP Score as the cutoff value for the sum of SAP values that determines the SAP Score

Variation of the SAP Score cutoff did result in different classifications based on the cutoff. There is a wide range of values that results in similar number of errors. In this case between -0.2 and 0.1 all incorrectly classify one mAb.

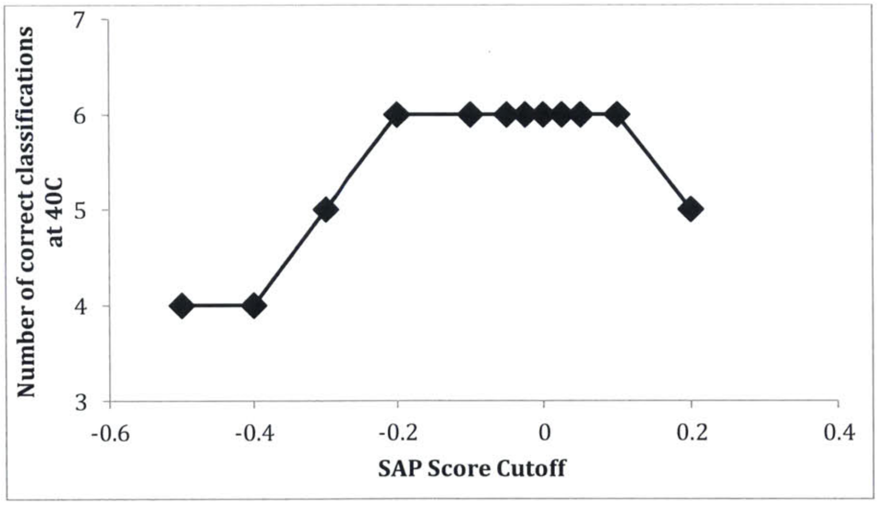

Figure 10-4 - Number of correct classifications (based on 7 data points) of 40 °C data using the given SAP Score cutoff value

This accounts for a large change in the percent of atoms included in the SAP Score calculation, approximately 76% of observed SAP values in the CDR fall within this range.

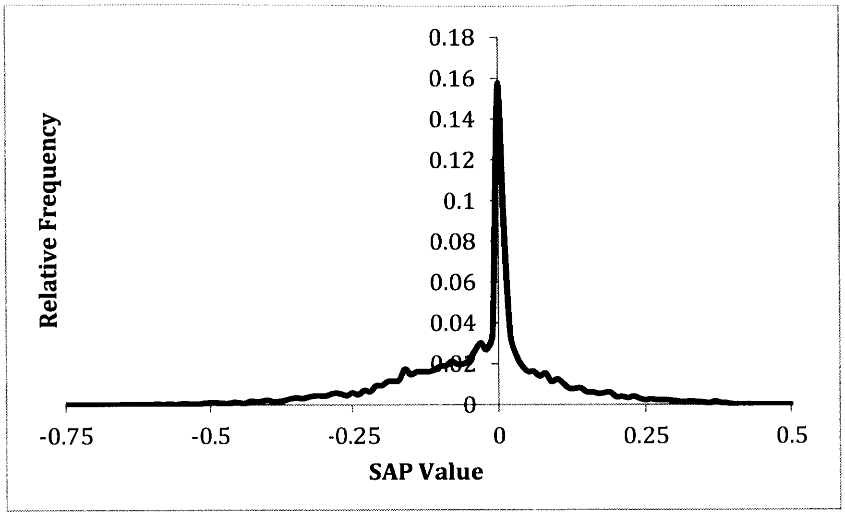

Figure 10-5 - Relative frequency of different SAP values for all the atoms in the CDRs of the seven mAbs, mAb1 through mAb7, used in the original fitting.

10.3. Additional Functional forms for DI

In addition to those presented here, other functional forms were tried. These were left out of the report for one of two reasons; the equations had mathematical issues that prevented their use or performed worse than simply using SAP as a predictor.

The first of these additional equations was the use of the natural logarithm:

$$\mathrm{DI} = \ln\left(\left[\mathrm{mAb\ SAP\ Score}\right] - \beta \times \left[\mathrm{mAb\ Net\ Charge}\right]^2\right)$$

Equation 10-1

This equation, and other similar equations, was not considered a good equation after it was noted that a mAb with a very high charge could cause the term inside the logarithm to be negative. As the logarithm of a negative number is not defined, this equation, and those like it, was not considered for use as the functional forms of DI.

Another equation that was considered was:

$$\mathrm{DI} = \frac{\left[\mathrm{mAb\ SAP\ Score}\right]}{\left[\mathrm{mAb\ Net\ Charge}\right]^2}$$

Equation 10-2

This equation performed worse than the SAP Score alone. In fact, if the DI was constant and always returned high stability, it would perform equally well.

Below is a table of the results of all fittings, including the function, the number of correct classifications and the values of β and the scaling constant c. "c" is the scaling factor applied to the aggregation propensity centroids to scale the kinetic rate values to the DI scale. In order to get the new DI classifications, multiply the aggregation propensity cutoff in Table 2-2 by c, these will be the cutoff value for the DI if that function is used.

Table 10-1 - Fitted parameters and number of correct classifications (based on 7 data points) for the tested DI functions

DI Function:	**Number of correct classifications**	**β (40°C)**	**C (40°C)**
DI = [Antibody SAP Score] - β x [Antibody Net Charge]^2	6	0.0498	4.6977x10^4
DI = [Antibody SAP Score] - β x [Antibody Net Charge]	6	0.7973	6.5409 x10^4
DI = exp ([Antibody SAP Score] - β x [Antibody Net Charge]^2	5	0.0756	10
DI = exp ([Antibody SAP Score] - β x [Antibody Net Charge])	5	1.7208	10
$\mathrm{DI} = \frac{[\text{mAb SAP Score}]}{[\text{mAb Net Charge}]^2}$	6	0.0607	88.3862

10.4. SAP Map for few Representative MAbs

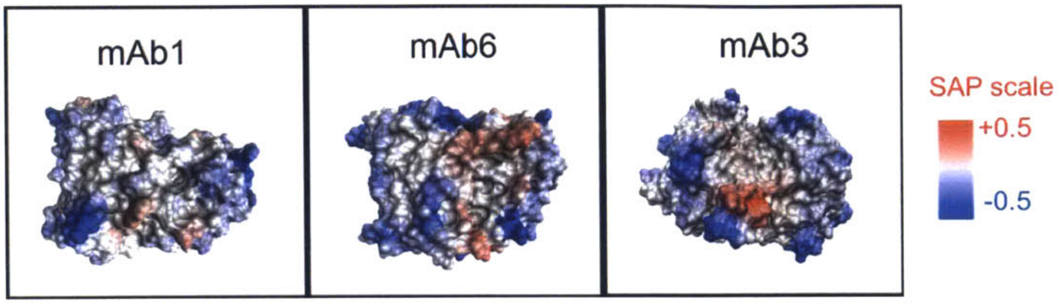

Figure 10-6 - SAP of variable regions (Fv) of mAb1 (Low SAP Score, 16.3), mAb6 (Medium SAP Score, 35.6), and mAb3 (High SAP Score, 46.0), R=5 Å. Red indicates hydrophobic regions while blue are hydrophilic regions.

10.5. Effect of Variations in the mAb Framework

There are few differences in the sequence of Fv domain outside of the CDR. For instance, the constant region of the light chain varies by 63 amino acids between κ and λ. In the case of the seven mAbs used in the fitting, mAb1 through mAb7, there is a wide range of values for the sequence identity and sequence similarity of the mAbs. In particular, if only the mAbs that are IgG1 with a κ side chain are used, the average sequence identity between the framework (the Fv not including the CDR) of two mAbs is 81%, but this can be as low as 68% or as high as 91%. Similarly, a wide range is observed in the sequence similarity, ranging from 83% to 94% (mean: 89%). The values are given below in the tables Table 10-2 and Table 10-3.

Table 10-2 - The sequence identity for the framework IgG1 κ mAbs.

	mAb1	mAb2	mAb3	mAb5	mAb6
mAb1	100%				
mAb2	68%	100%			
mAb3	76%	79%	100%		
mAb5	71%	83%	85%	100%	
mAb6	82%	84%	91%	86%	100%

Table 10-3 - The sequence similarity between the sequences for the framework of IgG1 κ mAbs

	mAb1	mAb2	mAb3	mAb5	mAb6
mAb1	100%				
mAb2	83%	100%			
mAb3	89%	86%	100%		
mAb5	85%	89%	93%	100%	
mAb6	89%	92%	94%	93%	100%

In order to test the hypothesis that the CDR is sufficient to capture the differences between mAbs, a SAP Score for the entire Fv was calculated. The SAP Score for the Fv is defined as the sum of the positive SAP values in the Fv, including the CDR. While one could calculate a SAP Score for the entire mAb, there were no differences in the sequence of any of the constant regions of the heavy chain, or of the constant region of the light chain (given either a κ or a λ chain). There is a remarkable correlation between the values calculated for Fv and the CDR (correlation coefficient of 95%, see Figure 10-7). The high correlation between the SAP Score based only on CDR residues and the SAP Score based on the Fv suggests that the mutations outside the CDR have little effect on the SAP Score of the mAb. Because the SAP Score of the CDR is sufficient to capture the differences in the SAP over the surface of the protein and requires a calculation over a small region, the SAP Score based on the CDR was used.

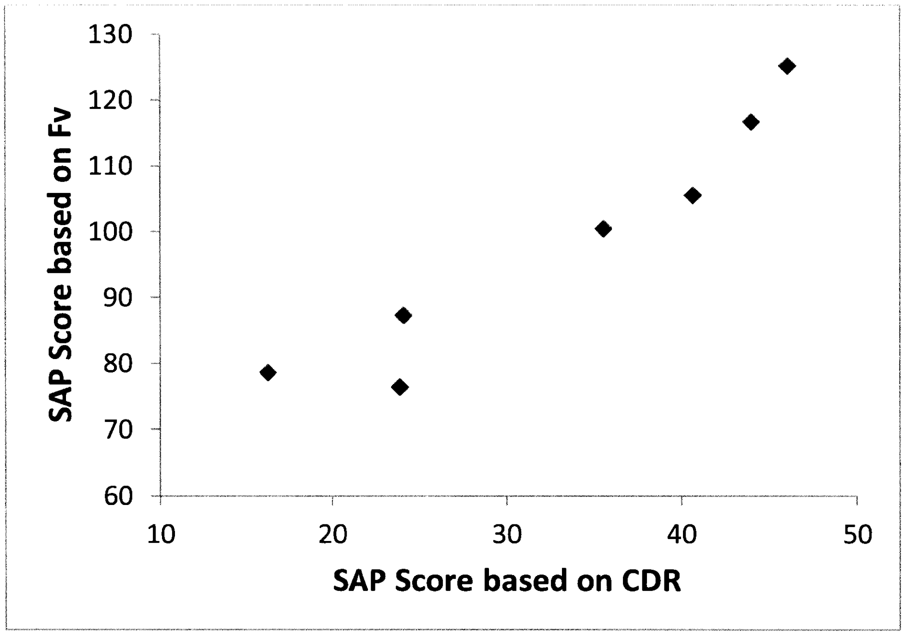

Figure 10-7 - SAP Score based on Fv vs. SAP Score based on Fv

10.6. Experimental Conditions for Stability Studies

Table 10-4 - List of experimental conditions for stability studies for each mAb, multiple experiments are available for mAb1, mAb2, mAb3, and mAb10

mAb	Additives	pH	Initial mAb Concentration [mg/mL]	Initial Percent of Aggregate [%]
mAbl	20 mM L-histidine	6.5	150	1.7
			75	1.1
mAb2	20 mM L-histidine	6.5	150	0.8
			75	0.3
mAb3	20 mM L-histidine	6.5	150	1.5
			75	0.9
mAb4	10 mM L-histidine	6.0	55	0.8
mAb5	10 mM L-histidine	6.0	70	0.8
mAb6	10 mM L-histidine	6.0	60	0.8
mAb7	20 mM L-histidine	6.0	40	3.0
mAb8	10 mM L-histidine	6.2	65	0.4
mAb9	20 mM L-histidine 100mM Trehalose	5.0	60	0.3
mAb10	25mM Phosphate	6.2	150	2.2
			75	1.9
mAb11	10 mM L-histidine	5.5	60	2.9
mAb12	10 mM L-histidine	5.5	75	0.3

11. APPENDIX B: ADDITIONAL INFORMATION ON THE UNDERSTANDING OF THE NON-ENZYMATIC HYDROLYSIS

11.1. Investigation of the Formamide Pathway

11.1.1. Methods

In order to determine the minimal level of theory to produce accurate system energies and structures, the impact of functional choice of energy evaluation and the basis set for structure optimization was studied. In order to quantify their impacts, the hydrolysis of formamide was investigated, as formamide has previously been studied and both stable species and transition states are known. The structures for the rate determining step of hydrolysis, the formation of the diol, from Wang and Cao (***120***) were used as fixed structures to investigate the impact of energy evaluation method on the system energy. A wide range of functionals were tested to see which reproduced the difference in reaction and activation energies calculated with CCSD(T)/6-311+G**. All energy evaluation methods were tested with the 6-311++G(3df,3pd), unless otherwise noted. The tested energy methods included:

Generalized Gradient functionals (GGA): BLYP (***151***), PBE (***152***)

Hybrid-GGA functionals: B3LYP (***130***), Becke Half & Half, HCTH (***153***), PBE0 (***154,155***)

Meta-GGA functionals: BMK (***156***) M06, M06-2X (***157***), TPSS (***158***)

range corrected: M11 (***159***), μPBE (***160***), ωPBE (***161***), ωB97, ωB97X (***134***), ωB97X-D (***162***)

hyper functionals: B2PLYP (***163***), ωB97X-2LP (***164***)

and wave-function based methods: CCSD(T) (***165***), HF (***166***), MP2 (***167***), RIMP2 (***168***)

The impact of basis set was evaluated by testing several Pople basis sets, starting with 6-31G* (***169***) and adding in various addition polarized basis sets, diffuse basis sets (***170***), or switching to the triple zeta 6-311 basis set. These were then compared to the results for optimizations with the largest basis set 6-311++G(3df,3pd), specifically the relative Gibbs Free energies of the reaction and activation energies evaluated with MP2/6-311++G(3df,3pd), the value of the imaginary frequency, and the displacements associated with each atom in the imaginary frequency. All optimization were done using the B3LYP functional and the C-PCM implicit solvent (***131***) in Gaussian.

11.1.2. Results

11.1.2.1. Selection of Functional for Energy Evaluation

Due to the size of the peptides, the smallest tripeptide is 50 atoms in size not including the surrounding waters, a model system (formamide) was used to quantify the impact of energy evaluation. In order to evaluate the impact of the energy evaluation method, several were methods tested on the model system. The structures were taken from the supplemental information of (***120***), these structures were optimized with B3LYP/6-311+G(2df,2p) with a CPCM optimization. The absolute errors in ΔE_0 are presented in Figure 11-1 for the tested functionals. While more accurate energy evaluation methods than the reference (CCSD(T)/6-311+G**) exists, a wide range of methods reproduce the activation and reaction energies. Many methods reproduce the energy within 1 kcal/mol of the target method, these accurate methods include BMK, M06, M11, ωB97, ωB97X, ωB97X-D, ωB97X-2LP, CCSD(T)/6-31+G** and MP2. The most accurate of these methods was ωB97X, which on average deviated from the CCSD(T)/6-311++G(3df,3pd) by 0.04 kcal/mol. This method was selected for use in all following energy evaluations due to the moderate cost of the method, and it and its related functionals (ωB97, ωB97X-D, ωB97X-2LP) were all reasonably accurate.

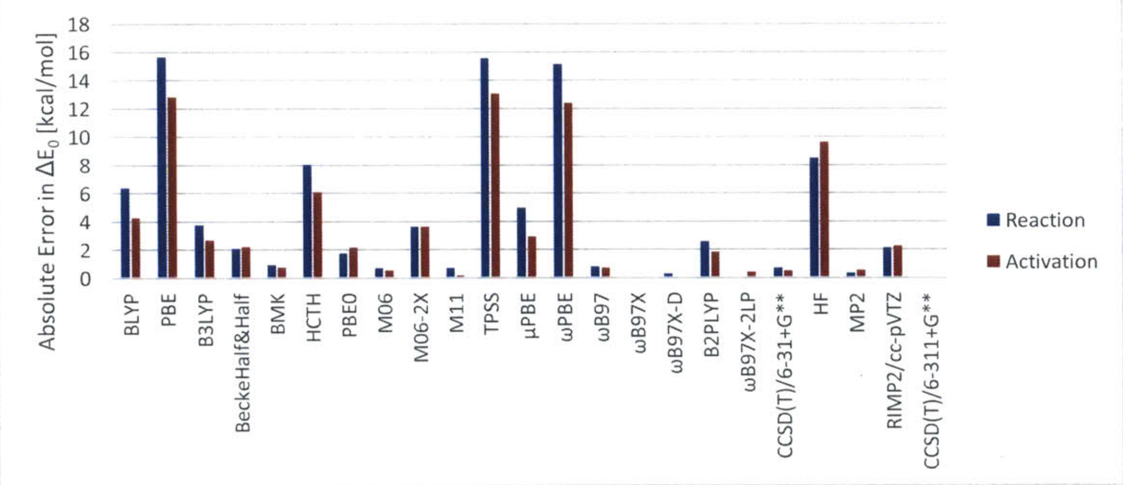

Figure 11-1 - Absolute deviation for ΔE_0 for the transition state (in red) and diol intermediate (in blue) compared to CCSD(T)/6-311+G energies. Holding the basis set fixed (6-311++G(3df,3pd)), except for RIMP2, which used cc-pVTZ and rimp2-cc-pVTZ as the auxiliary basis set, CCSD(T) which used 6-31+G**.**

11.1.2.2. Selection of Basis Set for Optimization

In addition to quantifying the error level of theory on energy accuracy, the impact of basis set on final energy was quantified. This was tested by optimizing the formamide system with different basis sets, and comparing the results to larger basis sets. Specifically, the energy of species, the frequencies and the correlation between the displacements associated with the imaginary frequency between the given basis set and the 6-311++G(3df,3pd). These results are tabulated in Table 11-1. From this table, the results with the 6-31G* basis set does not appear to be sufficient, as while it produces similar frequencies it does not accurately reproduce the energy suitably well. The smallest basis set the results in energies similar to the largest basis set 6-311++G(3df,3pd) is 6-31+G**. It reproduces the relative energy of species within 1 kcal/mol, and gets the activation and reaction energies within 1 kcal/mol. Additionally, it reproduces the negative frequency with little difference compared to the largest basis set. Because the 6-31+G** basis set best reproduces the structures of the largest basis set, it will be used for all the following calculations.

Table 11-1 – Tabulated values for the error in energy resulting from the use of different basis sets. The energy was evaluated using MP2/6-311++G(3df,3pd)//B3LYP/specified basis set, the harmonic approximation and a C-PCM as an implicit solvent. All values are relative to the values calculated with the 6-311++G(3df,3pd) basis set. Frequencies are evaluated with the B3LYP functional and the stated basis set.

Basis Set	Error in 2d-P [kcal/mol]	Error in 2d-R [kcal/mol]	Error in TS2-d [kcal/mol]	Frequency of imaginary frequencies (Hz)	Correlation of displacements in imaginary frequencies
6-31G*	5.7	17.9	6.5	-669.822	0.989
6-31+G**	-0.4	-1.0	-0.2	-573.596	0.998
6-311G(d,p)	4.4	1.2	6.0	-604.239	0.982
6-311+g(2df,2p)	-0.2	-0.3	-0.6	-639.589	1.000

11.2. Impact of GA on Free Energies

Due to the limits of MM modeling of the system, the GA may not represent the global minimum on the potential energy surface using any quantum method. To test this, 10 separate populations were created for the (H$^+$)PA(OH)$_2$A. These were evolved using the GA for approximately 8000 iterations (a fixed amount of compute time) and the most stable structure was taken, and the free energy of the system was evaluated. These were tabulated in the Table 11-2. A range of 3 kcal/mol was found between the six different conformations (four of the GA populations resulted in structures found in other populations).

Table 11-2 - The MM energy and Free Energy for 6 Conformations out of GA. All energies are in kcal/mol and relative to the lowest MM energy species. Energy Evaluated using M06/6-311++G(3df,3pd)//B3LYP/6-31G*, CPCM solvation, and the harmonic approximation

Conformation	MM Energy	Free Energy
1	1.48	-2.59
2	0.99	-2.27
3	0.00	0.0
4	1.02	-0.88
5	0.55	-0.90
6	0.91	0.63

11.3. Naming Conventions

Due to the number of species, and the variations in sequence being studied, a naming convention was used to clarify the species being talked about. Presently there are two key pieces of information contained in the name. The first is the sequence of the originating peptide. For instance, PDA refers to a peptide containing three residues, beginning with a proline, then an aspartic acid, and finally alanine. The other important information, is to differentiate between the various species in the reaction; this is denoted by parenthesis. For instance, the $(-OH)_2$ denotes the diol intermediate, while the (H^+) denotes the presence of a proton. The location in the name also caries significance, if a (H^+) is between two residues, which indicates the presence of a proton, bound to the carbonyl group on the backbone between the two residues. NH+ indicates the protonation of the nitrogen between the two amino acids. While a space in the name indicates cleavage of the amide bond. The final mechanism, those forming a furane ring are denoted with a (Cyc) after the aspartic acid to denote the formation of the ring. To illustrate this naming convention, please see Figure 11-2.

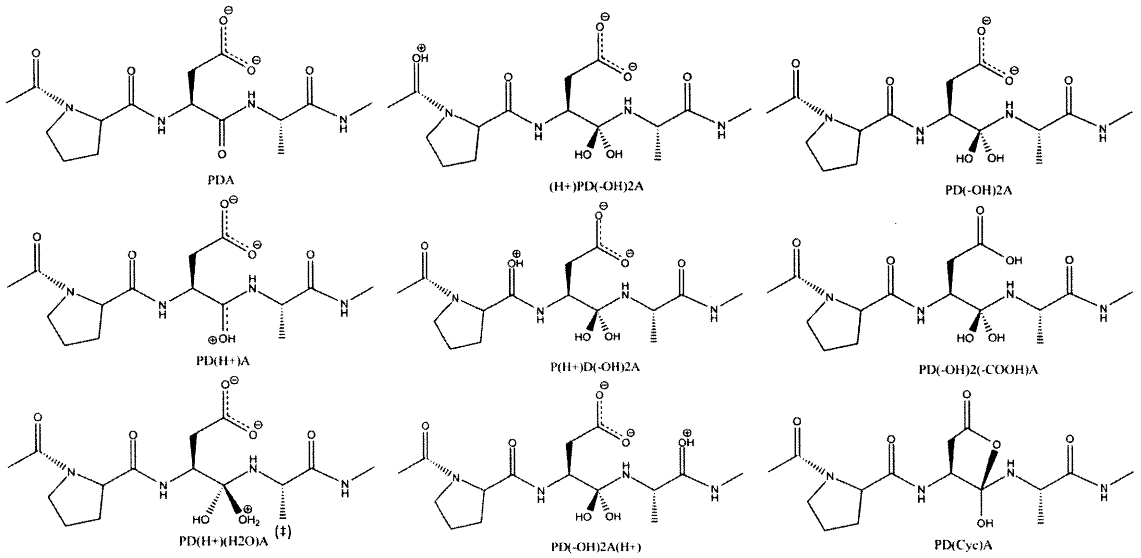

Figure 11-2 – Structures and naming convention applied to species in PDA based hydrolysis reaction

11.4. Tested Predictors

A wide range of predictors were tested for use in differentiating reactive and non-reactive carboxylic acid containing residues, either an aspartic or glutamic acid. These predictors can be classified into three categories of properties, those based on the primary sequence and secondary structure of the residue, those based on the residues interaction with its surroundings, and the relative configuration of the residue.

The first class of predictors were those based on the sequence and secondary structure of the protein. These included: if this residue was an aspartic acid or glutamic acid, if the preceding residue was proline, if the following residue was proline, the charge of the preceding residue, the charge of the following residue, and the hydrophobicity of both the preceding and following residue. Another predictor was the secondary structure classification of the residue, specifically the percentage of time a residue was in a turn, an extended conformation, an isolated bridge, an α helix, a 3_{10} helix, a π helix, or part of a random coil of the MD simulation as classified by VMD (*95*). These were included to determine if certain sequences or secondary structural elements accelerated or hindered hydrolysis.

The next class of predictors were those measuring the interactions between the protein and its environment. These included the number of waters within a given distance, all distances from 2 to 8Å, at 0.25Å increments, of the carbon, oxygen, and nitrogen atoms of the amide bond. The solvent available surface area of the backbone carbon, backbone oxygen, and sidechain were also tested as predictors. These were included to quantify the solvent interaction with the residue, and more specifically to what extent was the residue was buried. In addition to interactions with the solvent, interactions with the rest of the protein were also quantified. These included the number of hydrogen bonds between the residue of interest and the rest of the protein. This count was separated into three domains, the number of hydrogen bonds between the sidechain and the rest of the protein, the number of hydrogen bonds between the backbone and the rest of the protein, and the number of hydrogen bonds between the backbone carbonyl group and the rest of the protein.

The remaining predictors were measures of the internal geometry of the residue, these included a number of distances, angles, and dihedral angles. These included six atomic distances; those between Os (the carboxylic acid oxygen closest to the amide bond) and the C (the residue's backbone carbon), the distance between Os and Nn (the nitrogen of the following residue), the distance between Os and O (the backbone oxygen), the distance between Cg (the gamma carbon) and O, the distance between Cg and C, and the distance between Cg and Nn. Six angles were included; Os-C-O, Os-C-Ca (where Ca is the alpha carbon), Cb-Cg-Os (where Cb is the beta carbon), Os-Ca-C, Nn-C-Os, and Cg-Cb-Os. Seven dihedral angles were also used as predictors; Cg-Cb-Ca-C, Os-Cb-Ca-C, Cb-Ca-C-O, and Cb-Ca-C-Os, C-Ca-Cb-Os, Ca-Cb-Cg-Os, and Cb-Ca-C-Nn. Three more dihedrals were included that were used to predict the secondary structure including, these were the φ, ψ, and ω dihedrals. The last grouping of variables were the number of sidechain atoms within a given distance of each atom of the amide bond.

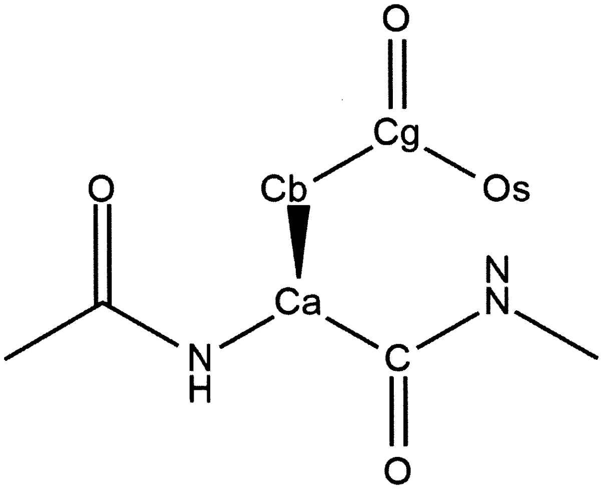

Figure 11-3 – Labeling of atoms in aspartic acid residue for use in machine learning predictors, Os is the closest of the carboxylic oxygen atoms. Cg, Cb, and Ca are the residue's gamma, beta, and alpha carbon's respectively. C and O are the backbone carbonyl carbon and oxygen atoms.

11.5. Attempts at Machine Learning

Several attempts were made to use machine learning to find the best predictors and separation criteria for the prediction of hydrolysis sites. A wide range of predictors were tested, see section 11.4, and a range of classifiers were tested. These included: classification/identity trees, support vector machines, and clustering. All data fitting was done using MATLAB ® and the MATLAB Machine learning toolbox (***146***). Several combinations of training and testing sets were done using the data available for prot-X, prot-Y, and prot-Z. However, none performed well when applied to a training set. This is likely due to overfitting of the data due to the large number of tested predictors. A few of the tested combinations are listed in Table 11-3.

Table 11-3 – Results of several machine learning models. Models are marked by the different classification algorithms used, the protein data used to train the model, the data used to test the model and the performance of the model based on the number of true non-reactive predicted bonds, the number of falsely predicted reactive bonds, the number of falsely predicted non-reactive

bonds, and the number of correctly classified bonds. The number before the parenthesis is the number of that bond in the training set, while the number in the parenthesis are the overall number.

Algorithm Used	Training Set	Test Set	True Non-Reactive Bond	False Reactive Bond	False Non-Reactive Bond	True Reactive Bond
Clustering	prot-X	prot-Y and prot-Z	236 (294)	0 (0)	8 (8)	0 (6)
	prot-X and prot-Z	prot-Y	110 (294)	0 (0)	4 (14)	0 (0)
	prot-X and prot-Y	prot-Z	126 (294)	0 (0)	4 (14)	0 (0)
Support Vector Machine	prot-X	prot-Y and prot-Z	236 (294)	0 (0)	8 (10)	0 (4)
	prot-X and prot-Z	prot-Y	109 (293)	1 (1)	4 (4)	0 (10)
	prot-X and prot-Y	prot-Z	125 (293)	1 (1)	4 (5)	0 (9)
	prot-X, prot-Y, and prot-Z	None	(294)	(0)	(1)	(13)
Classification Tree	prot-X	prot-Y and prot-Z	236 (294)	0 (0)	8 (8)	0 (6)
	prot-X and prot-Z	prot-Y	110 (294)	0 (0)	4 (5)	0 (9)
	prot-X and prot-Y	prot-Z	122 (289)	4 (5)	4 (7)	0 (7)
	prot-X, prot-Y, and prot-Z	None	(294)	(0)	(2)	(12)