A Protein Sequence-Properties Evaluation Framework for Crystallization Screen Design

by

David S. Dougall

Biology, BS, University of Pittsburgh, 1991

Marine-Estuarine Environmental Sciences, MS, University of Maryland, 1994

Submitted to the Graduate Faculty of

School of Medicine in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2007

UNIVERSITY OF PITTSBURGH

School of Medicine

This dissertation was presented

by

David S. Dougall

It was defended on

October 27, 2006

and approved by

Gregory F. Cooper, M.D. Ph.D.
Associate Professor, Department of Biomedical Informatics, University of Pittsburgh

Jim Lyons-Weiler, Ph.D.
Assistant Professor, Department of Biomedical Informatics, University of Pittsburgh

Jerrold H. May, Ph.D.
Professor, Katz Graduate School of Business, University of Pittsburgh

John Rosenberg, Ph.D.
Professor, Department of Biological Sciences, University of Pittsburgh

Dissertation Advisor: Vanathi Gopalakrishnan, Ph.D.
Assistant Professor, Department of Biomedical Informatics, University of Pittsburgh

A Protein Sequence-Properties Evaluation Framework for Crystallization Screen Design

David S. Dougall, PhD

University of Pittsburgh, 2007

The goal of the research was to develop a Protein-Specific Properties Evaluation (PSPE) framework that would aid in the statistical evaluation of variables for predicting ranges of and prior probability distributions for protein crystallization conditions. Development of such a framework is motivated by the rapid growth and evolution of the Protein Data Bank. Features of the framework that has been developed include (1) it is an instantiation of the "scientific method" for the framing and testing of hypotheses in an informatics setting, (2) the use of hidden variables, and (3) a negative result is still useful.

The hidden variables examined in this study are related to the estimated net charge (Q) of the proteins under consideration. The Q is a function of the amino acid composition, the solution pH, and the assumed pK_a values for the titratable amino acid residues. The protein's size clearly has a significant impact on the magnitude of the Q. Therefore, two additional variables were introduced to mitigate this effect, the specific charge ($\overline{Q}$) and the average surface charge density (σ).

The principal observation is that proteins appear to crystallize at low values of $\overline{Q}$ and σ. One problem with this observation is that "low" is a relative term and the frame of reference requires careful examination. The results are sufficiently weak that no prospective predictions appear possible although information of this type could be included with other weak predictors in a Bayesian predictor scheme. Additional work would be required to establish this; however that work is beyond the scope of the dissertation. Although many statistically significant correlations

among Q-related quantities were noted, no evidence could be developed to suggest they were anything other than those expected from the additional information introduced with the hidden variables.

Thus, the principal conclusions of this PSPE analysis are that (1) $\overline{Q}/\sigma$ and other Q-related variables are of limited value as prospective predictors of ranges of values of crystallization conditions. Although this is a negative result, it is still useful in that it allows attention to be directed into more productive avenues.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

First and foremost, I would like to thank my Ph.D. advisor, Vanathi Gopalakrishnan for all of her guidance and encouragement over the years. Next, I would like to thank John Rosenberg (Department of Biological Sciences, University of Pittsburgh) for his guidance and helpful comments early on, especially on the discussion. Dr. Rosenberg's ideas and keen insight formed the foundation for all of this research, which I am extremely grateful. At his suggestion, I began examining the effects of charge on crystallization, which then led to his suggestion to examine the estimated specific charge and estimated surface charge density. Many hours of discussion were also spent over the past seven years with Dr. Rosenberg's computational crystallography group, which included Dan Hennessy and Eric Williams, in addition to Vanathi and me. I would also like to thank the other members of my committee, Greg Cooper, Jim Lyons-Weiler, and Jerry May. I have taken classes from each of them over the years and have walked away a more well-rounded and knowledgeable person.

In addition to the people directly involved with my Ph.D. research, I would like to thank the people of the Department of Biomedical Informatics, especially Greg Cooper, Chuck Friedman, Cindy Gadd, and Mike Becich. Their hard work and dedication transformed the Center for Biomedical Informatics into the Department of Biomedical Informatics during my tenure here. While they worked the frontlines, a good support staff followed their lead. I will always have fond memories of Joe Cummings, Toni Porterfield, Cleat Szczepaniak, Rose-Ann Thomas, Beth LaRotonda, Kim Barnhart, Christen Reid, Bill Milberry, and many others. I would also like to thank my fellow Biomedical Informatics Graduate Students over the years.

I also would like to thank the National Library of Medicine for funding me for five years with a Pre-Doctoral Fellowship through the Center for Biomedical Informatics at the University of Pittsburgh (Medical Informatics Training Grant number NLM/NIDR 5 T15 LM/DE07059).

Lastly, but certainly not least, I would like to thank my family, wife Angela and children, Alaina and Graham, for sharing my time with others. One of the smartest and most beautiful people I know, Angela has supported me throughout my journey. I definitely could not have done it without her.

GLOSSARY

2Step:– Two-step clustering

AA_{au}: Asymmetric unit amino acid composition

Angstrom (Å): $1x10^{-10}$ meters

A_S: Solvent Accessible surface area; measured in nm^2

Controllable: experimental variables that can be controlled by the researcher

Discretization: the process of transforming a continuous variable into a categorical variable

$diff_{lim}$: resolution limit of diffraction, measured in Angstroms (Å)

e: electron units of charge

Features: independent variables

Givens: variables that are known or can be estimated prior to any crystallization attempts

GRAVY: Grand Average of the Hydropathy, measure of hydrophobicity (Kyte and Doolittle, 1982)

HHE: Henderson-Hasselbach Equation

Hidden Features: independent variables that are not within the database of interest, but can be calculated or estimated with the *Given Features* and/or *Given Observables*

Hidden Controllables: environmental conditions that are not within the database of interest, but can be set experimentally

Hidden Observables: dependent variables that are not within the database of interest, but can be calculated or estimated with the *Given Features* and/or *Given Observables*

Isoelectric point: the solution pH where a protein exhibits a net charge of zero

kDa: kilodaltons

ln: natural log

$ln(MW_{au})$: natural log of the asymmetric unit Molecular Weight

m*e*: millielecton units of charge

mmCIF: macromolecular Crystallographic Information File

MW: Molecular Weight in kilodaltons (unless otherwise noted)

MW_{au}: Asymmetric unit Molecular Weight in kilodaltons

nm: nanometer; $1x10^{-9}$ meters

NMR: Nuclear Magnetic Resonance

nrPDB: Non-redundant Protein Data Bank

$nrPDB_{10.04.05}$: Non-redundant PDB training set created based on the PDB and nrPDB on 10/04/2005, which consisted of 4,114 proteins

Observables: Experimental outcome variable

PDB: Protein Data Bank

PDB-REPRDB: The Representative Protein chains from the PDB database

pI_{est}: Estimated isoelectric point

pH_{cryst}: Reported pH of crystallization

pK_a: Acid-dissociation constant, where an equal number of titratable residues are ionized and deionized

Primary Features: The main *Feature* available from the database (PDB); in this case the asymmetric unit sequence

Primary Observable: The main *Observable* available from the database, pH_{cryst} for this dissertation

Proxy Variable: a measured variable used to infer the value of a variable of interest

Q: Estimated net charge

Q_{cryst}: Estimated net charge at the reported pH of crystallization

$\overline{Q}$: Estimated charge in electrons/kilodaltons (*e*/kDa)

$\overline{Q}_{cryst}$: Estimated specific charge in electrons/kilodaltons at the reported pH of crystallization

$\widetilde{Q}$: scaled estimated net charge

PSPE Framework: Protein Sequence-Properties Evaluation Framework

σ: Estimated surface charge density in millelectrons/nm^2 (m*e*/nm^2)

σ_{cryst}: Estimated surface charge density at the reported pH of crystallization in millelectrons/nm^2

SC: Specific Charge

SCD: Surface Charge Density

Charge Range Test: the Charge Range Test measures the amount of proteins whose estimated specific charge or estimated average surface charge density values fall within a given range of the group mean

SD: Standard Deviation

SE: Standard Error

SG: Structural Genomics

SOM: Self-Organizing Map

Supervised SOM: Supervised Self-Organizing Map developed for this dissertation to determine the optimum number of clusters of a one-dimensional vector of estimated surface charge density curves.

Surface charge density: the estimated net charge of the protein relative to its size in nm^2

Target Protein: a protein that a researcher is trying to crystallize

TargetDB: Structural Genomics (SG) database listing all proteins undergoing structural determination and tracking their process through the SG pipeline

1.0 INTRODUCTION

X-ray crystallography is the primary method of choice for modeling the three-dimensional (3D) structure of proteins and other biological macromolecules, such as deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). Determining the 3D structure of a protein can provide important details, such as its function and mode of action (how it works), and can aid in drug design and discovery (McPherson, 1999). Currently, there are more than forty compounds used as drugs that have been discovered or identified using structural-based methods. Many of these compounds have been approved or are undergoing clinical trials for the treatment of human disease; examples include drugs against HIV/AIDS (HIV proteases: Nelfinavir, Amprenavir, and Lopinavir), chronic myelogenous leukemia (Imatinib), and non-small cell lung cancer (Erlotinib; Congreve, et al., 2005).

1.1 PROBLEM

There are several steps involved in obtaining the 3D structure of a protein and breakdowns in this process can occur for any number of reasons. Crystallizing the protein along with obtaining a soluble form of the protein are the major bottlenecks in the structure determination process. Growing a crystal suitable for X-ray diffraction studies can be very time consuming, costly, and difficult, as demonstrated by the low success rate. Crystal growth is often problematic because of the large search space. The experimenter has to determine the correct concentrations of the protein, salt(s), and buffer, as well as the temperature and pH. Additionally, the optimal crystallization conditions appear to vary idiosyncratically from protein to protein raising the question: What, if anything, can be predicted about the crystallization behavior before the fact and if so, how can that be used to facilitate the process?

The current 'state of the art' methods for crystallizing a protein consist of randomly searching the possible conditions, using an ordered search by combining all pairwise combinations of two or more solution variables at several levels each (grid screens), or using the solution conditions that have worked for other previously crystallized proteins (the sparse-matrix approach). Because of the large amount of parameters to search and the usually limited amount of protein, researchers try to maximize the probability of obtaining a crystal by using the sparse-matrix approach. Companies, such as Hampton Research, Molecular Dimensions, Emerald BioStructures, or Jena Bioscience, all market crystallization screens based on the sparse-matrix approach. The sparse-matrix approach for initial crystallization screens is based on the assumption that whatever solution conditions that have previously succeeded in crystallizing proteins have a higher probability in crystallizing new target proteins, a one-size-fits-all approach. These commercial screens are also available for special subclasses of proteins, such as membrane proteins or nucleic acids, and are relatively easy to use. However, even with these methods, the success rate of crystallization remains quite low.

The sparse-matrix method creates a somewhat repetitive set of conditions for several reasons. First, certain areas of the search space are over sampled, such as the abundance of high molecular weight polyethylene glycols (PEG) in screens. Second, a few of the experiments within a screen provide most of the crystallization success rate. Page et al. (2003) demonstrated that removing 75% of 480 screening conditions would have no effect on the number of proteins crystallized in a study of 539 *Thermotoga maritima* proteins. For these reasons, alternate methods need to be developed to condense screens even further to contain smaller subsets (Kimber et al., 2003; Page et al., 2003; Wooh et al., 2003; Page and Stevens, 2004).

One such method to developing alternative screening strategies is to determine what characteristics of a protein make it amenable to the structural determination process. Researchers are realizing the importance of recording the conditions used to grow the crystals. The Protein Data Bank (PDB; Berman et al., 2000) is now recording many of these parameters, but most are still optional upon deposition of a new 3D structure. Now with more available data due to funding for the Protein Structure Initiative (PSI) by the National Institute of General Medical Sciences (NIGMS) and a few independent researchers, such as John Rosenberg at the University of Pittsburgh, some plausible explanations can be explored in more detail. These projects are collecting information about successes (where to search) as well as failures (where

not to search). This dissertation developed a framework to frame and test hypotheses about protein crystallization behavior. If the hypotheses are validated on a retrospective data set, they can be used to suggest regions in the crystallization search space that are more likely to result in the formation of crystals of the proteins under study.

McPherson (1999) lists 36 parameters, 12 physical, 12 chemical, and 12 biochemical, which affect the outcome of crystallization experiments (Table 1.1). Each of these variables has multiple levels and may be either continuous, such as concentration and pH, or discrete, such as precipitant and salt type. Additionally, the interaction between these variables is not fully understood. Although many of the variables in Table 1.1 are not typically examined, the possible combinations of all parameters would be extremely large and unmanageable with limited protein, time, labor, and cost.

Table 1.1 Parameters influencing protein crystallization outcome.

Physical	**Chemical**	**Biochemical**
Temperature	**pH**	**Isoelectric point**
Surfaces	Precipitant Type	Macromolecular stability
Methodology	Precipitant Concentration	Source of Macromolecule
Gravity	Ionic Strength	Ligands, inhibitors, effectors
Pressure	Specific Ions	Proteolysis / Hydrolysis
Time	Degree of Supersaturation	Chemical Modifications
Vibrations / Sound	Macromolecular Concentration	Posttranslational modifications
Electrostatic / Magnetic Fields	Metal Ions	Genetic Modifications
Dielectric Properties of the Medium	Reductive / Oxidative Environment	Macromolecular Purity
Viscosity of the Medium	Cross-linkers/Polyions	Aggregation State
Homogeneous or Heterogeneous Nucleants	Detergents / Surfactants / Amphophiles	Inherent symmetry of the macromolecule
Rate of Equilibration	Non-macromolecular Impurities	History of the sample

* Taken from McPherson, 1999

Unfortunately, most of these parameters in Table 1.1 are not recorded or in a format suitable for information extraction. The conditions used to grow the crystals are starting to appear with more frequency in the macromolecular crystallographic information file (mmCIF),

but this information is voluntary and far from complete (Section 2.4). Additionally, only successful conditions are reported within the PDB. Negative examples would provide an additional wealth of information (Edwards et al., 2000; Rupp, 2003; Page and Stevens, 2004; Rupp and Wang, 2004), but are usually only reported within laboratory notebooks. Knowing where not to search in the parameter space, would allow the researcher to focus more of his/her efforts on areas where crystallization is more probable. Additionally, a large amount of protein would be required to screen all of these combinations, which would often not be available or cost effective due to the time and effort required for such a screening. Therefore, this research focused on features that are sequence based and easy to calculate, such as molecular weight and isoelectric point, and do not require the use of any of the protein sample.

1.2 SIGNIFICANCE

The approaches currently used to grow crystals are costly and have low success rates. For example, the estimated average cost of obtaining a single 3D protein structure is $100,000-$125,000 (Burley and Bonanno, 2002; Uehling, 2005). Coupled with success rates ranging from 1-23% (Chayen and Saridakis, 2002; Lesley et al., 2002; Rupp and Wang, 2004; Couzin, 2005; Page et al., 2005; Walter et al., 2005) it is clear that an alternative method for selecting crystallization conditions needs to be developed. This dissertation developed a hypothesis testing framework, the Protein Sequence-Properties Evaluation Framework (Chapter 4), which was used to demonstrate that most proteins do crystallize at a low estimated net charge, or more specifically a low estimated specific charge ($\overline{Q}$) and estimated average surface charge density (σ). One problem with this observation is that "low" is a relative term and the frame of reference requires careful examination. The results are sufficiently weak that no prospective predictions appear possible although information of this type could be included with other weak predictors in a Bayesian predictor scheme. Additional work would be required to establish this; however that work is beyond the scope of the dissertation.

1.3 THE APPROACH

The goal of this research was to conduct a retrospective study to identify features within a protein's primary sequence that can guide a protein crystallographer in designing and/or optimizing crystallization screens. The proteins examined in this research were a non-redundant set from the Protein Data Bank (PDB), i.e. proteins that have been successfully crystallized. The goal of the research was to develop a Protein-Specific Properties Evaluation (PSPE) framework that would aid in the statistical evaluation of variables for predicting ranges of and prior probability distributions for protein crystallization conditions.

1.4 PROTEIN SEQUENCE-PROPERTIES EVALUATION FRAMEWORK

The development of this framework is largely motivated by the rapid growth and evolution of the Protein Data Bank. The general framework was developed to test hypotheses about the solution conditions leading to the formation of protein crystals, the Protein Sequence-Properties Evaluation (PSPE). This framework is an instantiation of the "scientific method" for the framing and testing of hypotheses in an informatics setting. The goal was to identify protein features that could be used to suggest crystallization solution conditions that would result in an increased success rate of obtaining crystals. Because the reported pH of crystallization (pH_{cryst}) is the most recorded solution parameter, the initial focus was on using the estimated charge of the protein to suggest solution pH ranges for crystallization, but with more information available, any environmental variable could be examined. For example, based on the target protein's estimated net charge, the probability of growing a crystal suitable for diffraction studies over a range of pH values can be calculated when given its protein sequence. Although the PSPE framework was developed for the specific application of protein crystallization, it could be implemented in other areas of interest where researchers are trying to find or explain differences in protein behavior.

1.5 THESES

The overall hypothesis that forms the basis for this work is that there are features present within a protein's primary sequence that can guide a protein crystallographer to intelligently select experimental solution conditions that have a higher probability of generating crystals. This would ideally lead to the development of more successful crystallization screens. In this dissertation, the following theses are specifically developed and tested:

First, it is hypothesized that the Protein Sequence-Properties Evaluation Framework can be used to frame and test hypotheses in-silico about variables that are believed to be important in protein crystallization. These discoveries can then be used to predict ranges of solution *Controllables* in a probabilistic manner. It is hypothesized that this general framework will result in more efficient crystallization screens.

Second, based on the application of the PSPE framework to the examination of solution pH values of previously crystallized proteins, it is hypothesized that a protein's $\overline{Q}$ or σ could be used to suggest solution pH ranges that have a higher probability of generating crystals suitable for diffraction studies. This approach examines the distribution of both the $\overline{Q}$ and σ of previously crystallized proteins to suggest values for an independent dataset comprised of newer PDB entries. These charge values are then translated into pH space using the target protein's $\overline{Q}$ or σ curves. However, the mapping from charge space back into pH space is still an open research problem. Additional experimental validation of these results would be extremely important.

Third, it is hypothesized that 'similar' proteins crystallize under similar solution conditions, including those that result in similar charge values, in terms of both the $\overline{Q}$ and σ. Groups of proteins are identified with the PSPE Framework, using binning or clustering algorithms, such as two-step clustering and self-organizing maps, to increase the accuracy of predicting the $\overline{Q}$ and σ ranges of target proteins. The following protein features were examined to define proteins as being similar: the molecular weight, the estimated isoelectric point (pI_{est}), and the $\overline{Q}$ curve.

Based upon the above theses, specific claims are made:

1. The goal of the research was to develop a Protein-Specific Properties Evaluation (PSPE) framework that would aid in the statistical evaluation of variables for predicting ranges of and prior probability distributions for protein crystallization conditions. Development of such a framework is motivated by the rapid growth and evolution of the Protein Data Bank. Features of the framework that has been developed include:
    a. It is an instantiation of the "scientific method" for the framing and testing of hypotheses in an informatics setting.
    b. The use of hidden variables, *i.e.* parameters which are analytic functions of quantities extracted from the database. Note that one of the hazards of hidden variables is that they introduce additional information and care must be taken to ensure that "discoveries" based on the hidden variables are not simply reflections of that additional information.
    c. A negative result is still useful; *i.e.* the recognition that a variable has minimal utility in predicting ranges or probabilities of crystallization allows energy and attention to be focused elsewhere, where it may be more productively employed.
2. The hidden variables examined in this study are related to the estimated net charge (Q) of the proteins under consideration. The Q is a function of the amino acid composition, the pH of the solution and the assumed pK_a values for the titratable amino acid residues in the protein. Specific variables and observations include:
    a. The size of the protein clearly has a significant impact on the magnitude of the Q; two additional variables were introduced to mitigate this effect:
        i. The specific charge ($\overline{Q}$) is the ratio of the Q to the protein mass, expressed here in units of *e*/kDa (electron units of charge per kilo Dalton).
        ii. The surface charge density (σ) is the ratio of the Q to the estimated surface area of the protein; a convenient unit is m*e*/nm^2 (10^{-3} electron units of charge per square nm). Although the estimation of surface area is difficult, σ facilitates comparisons with other biologically relevant macromolecules.
    b. Additional Q-related quantities examined include:

i. The estimated isoelectric point (pI_{est}), which is the pH at which the Q is zero.

ii. Measures of the shape of the titration curve which is either $\overline{Q}$ or σ expressed as a function of pH.

c. The principal observation of this study is that proteins appear to crystallize at low values of $\overline{Q}$ and σ.

i. One problem with this observation is that "low" is a relative term and the frame of reference requires careful examination.

1. One frame of reference is provided by comparison to the known σ values for nucleic acids and phospholipid bilayers, whose surface charge densities are at least an order of magnitude greater than that of proteins.

a. One problem with this frame of reference is that there is no pH at which proteins are as highly charged as nucleic acids.

b. A more serious problem is that nucleic acids crystallize readily, demonstrating that high σ is not a barrier to crystallization.

2. Another frame of reference is in relation to the mean $\overline{Q}/\sigma$ values at "physiological pH" of 7.4 ($\overline{Q}_{pH=7.4}$); here, there is an approximately three-fold reduction in the mean of $\overline{Q}_{cryst}$ vs. the mean of $\overline{Q}_{pH=7.4}$.

a. One problem with this frame of reference is that the pH reference value, 7.4, is questionable because many proteins function in physiological compartments where the pH is significantly different.

b. A more serious problem is that the standard deviations of the two distributions are more than twice the shift between them. The shift is statistically significant because of the

sample size, but cannot be used to make meaningful predictions about specific proteins.

3. It should be noted that the preponderance of "domain knowledge" would be that the primary factors selecting low net charge are issues of protein folding and stability as well as issues of functionality.

d. The results are sufficiently weak that no protein-specific prospective predictions appear possible although information of this type could be included with other weak predictors in a Bayesian predictor scheme. Additional work would be required to establish this; however that work is beyond the scope of the dissertation.

e. Although many statistically significant correlations among Q-related quantities were noted, no evidence could be developed to suggest they were anything other than those expected from the additional information introduced with the hidden variables (see 1b).

3. Thus, the principal conclusions of this PSPE analysis are:

a. $\overline{Q}/\sigma$ and other Q-related variables are of very limited value as prospective protein-specific predictors of ranges of values of crystallization conditions.

b. Probabilities based on these variables may prove useful in a Bayesian sense, although that has not been demonstrated at this time.

c. Although this is a negative result, it is still useful in that it allows attention to be directed into more productive avenues.

1.6 DISSERTATION OVERVIEW

In Chapter 2, the current methods used to grow a crystal are discussed as well as other relevant research. In Chapter 3, the estimated net charge is identified as an important variable that may be able to be predicted apriori for crystallization. This analysis led to the development of the

Protein Sequence-Properties Evaluation (PSPE) Framework, which is presented in Chapter 4 (Claim 1). A specific application of the PSPE framework, identification of a low $\overline{Q}$ and σ being important for crystallization, are described in Chapter 5. In Chapter 6, various methods of clustering proteins by similarity are explored to improve upon the results in Chapter 5, which are used to generate solution pH priors for future proteins targeted for crystallization attempts. Several applications are presented in Chapter 7, which calculate the predicted pH ranges for crystallization for three proteins undergoing crystallization attempts in the Rosenberg Research Group (Department of Biological Sciences, University of Pittsburgh). A Discussion of the results, methods, and limitations are presented in Chapter 8, while future research opportunities are discussed in Chapter 9. The Python scripts written for this dissertation are found in Appendix A. An examination of the solution pH values of commercial screens was examined in Appendix B., Next, an annotated example on how to use the information developed in this dissertation is demonstrated in Appendix C using three proteins from the test set. Finally, the preliminary results of fitting a curve to the plot of pI_{est} vs. $\overline{Q}_{cryst}$ is presented in Appendix D, while the preliminary results of using a protein's amino acid composition to suggest pH ranges for initial crystallization attempts using Linear Regression and Neural Networks is shown in Appendix E.

2.0 BACKGROUND

Determining a protein's three-dimensional (3D) structure to molecular level detail can provide a wealth of information for biomedical researchers. The detailed 3D structure can give insight into a protein's function at a broad level, mechanism of action at a narrow level, and possible binding partners, regardless if any of these are already known. For example, molecular level detail can be used to determine the actual atoms from the amino acid residues involved in the active site of enzymes. This knowledge can also be used to examine and predict the effects of mutations of these residues, which is an important aspect in understanding the effects of various diseases. Additionally, determining a biological macromolecule's structure is a very important part of the drug design and discovery process, facilitating new treatments for disease and understanding of important biological processes.

2.1 METHODS OF OBTAINING A PROTEIN'S 3D STRUCTURE

There are several methods, both computational and experimental, that are used to determine the three-dimensional (3D) structure of biological macromolecules, primarily proteins and nucleic acids. The computational approaches are based upon the target protein's amino acid sequence and use no physical sample, but cannot provide the level of detail that experimental methods can unless there is a very high level of sequence similarity. Even when a target protein has a sequence similarity of greater than 50% to proteins with a known three-dimensional structure, computational approaches, such as homology modeling, can only predict the structure to a comparable X-ray resolution limit of diffraction of 3.0 Angstroms (Å). When high-level of molecular detail is required, such as that for enzymes, the computational approaches fail to provide enough detail and experimental approaches should be attempted (Burley and Bonanno,

2003). Experimental methods are able to provide a much better level of molecular detail, often approaching the atomic level, ~1.2 Å.

The experimental approaches require use of the actual protein sample. Even though only a few milligrams of protein may be required, this small amount is very difficult and time-consuming to obtain. The primary experimental methods consist of X-ray diffraction and nuclear magnetic resonance (NMR) spectroscopy. Although each of these different methods can each provide varying degrees of molecular detail of the structures, X-ray diffraction is the only one that can approach atomic level detail (McPherson, 2004).

The general experimental steps involved in the structure determination process and their success rates are discussed in Section 2.2. The protein's solubility is discussed in Section 2.3, while the crystallization process and the methods commonly used to obtain crystals are described in Section 2.4. Next, methods that can increase the probability of obtaining crystals for X-ray diffraction studies are discussed in Section 2.5. Finally, the variables involved in the crystallization process are examined methods in Section 2.6.

2.2 STRUCTURE DETERMINATION PROCESS

There are several common steps involved in any experimental protein structure determination process undertaken in the laboratory, whether using X-ray diffraction or NMR. The steps in order include target selection, cloning the protein, expressing the protein in a soluble form, and purifying it. The ability to obtain a purified protein can fail in any of these steps. Once a protein has been purified in sufficient quantities (milligrams), it can undergo the next step of the structural determination process, which is crystallization for the purposes of this dissertation (Section 2.4).

The biggest failure rates in the structural determination process have been reported in the literature as either obtaining a soluble form of the protein (Christendat et al. 2000; Edwards et al. 2000; Lesley et al. 2002; Yee et al., 2002; Goulding and Perry, 2003; Rupp, 2003a; Bourne et al., 2004; Albeck et al., 2005) and/or determining the correct environmental conditions for crystallization (Chayen, 1999; Heinemann et al., 2000; Chayen, 2002; Chayen and Saridakis 2002; Goulding and Perry, 2003; Bourne et al., 2004; Page and Stevens, 2004). Temporal data

available from the Protein Structure Initiative's database, TargetDB (Table 2.1, Westbrook et al., 2003; http://targetdb.pdb.org/statistics/TargetStatistics.html), supports this observation. Several large drop-offs in success rates are observed, such as attempting to obtain a soluble form of the expressed protein (47%) and trying to crystallize a purified protein (40%). There is even less success in obtaining a crystal suitable to diffraction studies from a purified protein (20%). When comparing the amount of proteins that have been cloned to those that have resulted in a diffraction quality crystal, a 4% success rate is observed.

However, these results give an overestimate of the success rate, because many proteins are not selected as targets if they are predicted to be difficult to solve. Proteins are often filtered from selection if they are predicted to be a transmembrane protein, a signal peptide, a coil-coil protein, a protein with low complexity, or a disordered protein. These proteins often have a more difficult time proceeding through the steps of structure determination and crystallizing. Thus, the PDB may not be a true representative sample of the protein universe as described by SWISS-PROT, the database of the protein universe. This is indeed the case, as the PDB contains an under abundance of transmembrane proteins, proteins with low complexity, signal sequences, and disordered regions. Alternatively, the PDB contains an over abundance of enzymes and proteins with disulfide bonds or metal-binding sites (Peng et al., 2004). It has been speculated that the proteins in the PDB are the 'low hanging fruit', meaning the soluble proteins that are relatively 'easy' to clone, express, purify, and crystallize. This illustrates that new methods are needed to improve upon the success rates.

In addition to being a difficult step, crystallization has been estimated as one of the most time consuming steps of the structural determination process (Table 2.1). Using temporal information available from the TargetDB, Bourne et al. (2004) were able to calculate the mean time to complete each step of the structural determination process. Cloning the target protein took 120 days on average, while expressing the protein required only approximately 25 days. After getting the protein expressed, it took an average of 75 days to purify the protein to sufficient quantities, which would also include obtaining a soluble form of the protein. Once the protein was purified, it required a mean of 195 days to obtain a diffraction quality crystal (85 days to get the protein crystallized with another 110 days to obtain a diffraction quality crystal). After obtaining a diffraction quality crystal, another 195 days on average were required to diffract and obtain the crystal structure. Finally, another 130 days were required to analyze the

structure and deposit it into the PDB, which generally includes getting a publication on the solved structure. The whole process from selecting a target to depositing the solved X-ray diffraction structure into the PDB took about 2 years on average. Additionally, a large part of this time was devoted to the crystallization process, approximately 195 days. Any method that can reduce the crystallization time would be of great benefit to the structural community.

Table 2.1 The steps in the determination of the 3D structure of proteins and the success rates[1].

Status	Total Number of Targets	(%) Relative to Cloned Targets	(%) Relative to Expressed Targets	(%) Relative to Purified Targets	(%) Relative to Crystallized Targets	Approximate Mean Time (Days) [2]
Cloned (C)	70,750	100	-	-	-	119 (C)
Expressed (E)	41,501	59	100	-	-	25 (C→E)
Soluble (S)	19,382	27	**47**	-	-	
Purified (P)	14,972	21	36	100	-	75 (S-P)
Crystallized (X)	5,934	8	14	**40**	100	85 (P→X)
Diffraction-quality Crystals (DC)	3,029	4	7	**20**	**51**	110 (X→DC)
Diffraction (D)	2,579	4	6	**17**	**43**	110 (DC→D)
Crystal Structure (CS)	2,384	3	6	**16**	**40**	85 (D→CS)
NMR Assigned	1,150	2	3	8	-	
NMR Structure	1,065	2	3	7	-	
In PDB	3,056	4	7	20	35	130 (CS→PDB) 110 (NMR→PDB)
Work Stopped	14,363	-	-	-	-	

[1] As determined for the Structural Genomics (SG) projects from the TargetDB on 05-19-2006 (Westbrook et al., 2003)
[2] The approximate times for each step were estimated from Bourne et al. (2004).

In an attempt to increase throughput, researchers have examined how a protein's biophysical properties may be correlated to success (1=yes) or failure (0=no) at each step of the process, cloning, expression, solubility, purification, and crystallization. The reason being that if

a protein has a low probability of producing crystals, it should be removed. In each step, the protein's biophysical properties have significantly correlated with success or failure (Goh et al., 2004). The large amount of data being generated by the Structural Genomics (SG) groups has enabled such analysis as data mining to be undertaken. Using the TargetDB, the status of each SG target protein is tracked. This enables one to create a list of proteins that have succeeded or failed at each step. In this dissertation, a protein's biophysical properties are used to predict a solution pH range that has a higher probability of producing crystals.

As mentioned previously, there is often a bias beginning at the first step, selecting the protein targets for analysis. Once a protein is selected for analysis, various expression systems (usually bacterial organisms) are used to get the protein expressed, i.e. made by the organism. A variety of protein properties such as size, amino acid composition (Christendat et al., 2000; Braun et al., 2002; Goh et al., 2004; Luan et al., 2004), isoelectric point (pI; Braun et al., 2002; Goh et al., 2004), secondary structure composition (Christendat et al., 2000), the presence of particular protein domains, subcellular location (Braun et al., 2002), hydrophobicity (Goh et al., 2004; Luan et al., 2004), signal sequence , and whether the protein is conserved across different organisms (Goh et al., 2004) have correlated with successful expression of the protein. For example, as the size increases, there is a decrease in expression. Of the amino acids shown to predict expression, the acidic amino acids, Glutamic acid and Aspartic acid, apparently play a major role (Christendat et al., 2000; Goh et al., 2004). Although a protein may be expressed, it may not be of use if it is not soluble.

Often proteins are expressed in an insoluble form and packaged into inclusion bodies by the cell. The proteins placed in these vesicles are often the result of insoluble aggregates of misfolded or denatured proteins. There are several reasons these insoluble masses may be formed. When the host organism forms a foreign protein, it may lack the appropriate chaperone or binding partner necessary to properly fold the protein. Additionally, bacteria cells lack the mechanism for post-translational modifications that may be required for eukaryotic proteins. However, there are several methods that are available to try to resolubilize the proteins located in the inclusion bodies, including adding a denaturing agent, such as urea. An insoluble protein presents problems in terms of purification and the possibility of misfolding. A variety of studies have been performed that have examined a protein's biophysical properties and whether the protein is soluble or insoluble upon expression. The composition of certain amino acids has

shown to be positively (Arginine, Glutamine, Glutamic acid, Isoleucine, and Leucine) or negatively (Asparagine, Cysteine, Glycine, Methionine, Phenylalanine, Proline, Serine, Threonine, and Tyrosine) correlated with solubility (Christendat et al., 2000; Luan et al., 2004; Idicula-Thomas et al., 2006; Idicula-Thomas and Balaji, 2005). Many of these correlations seem logical such as positive correlation of Arginine, Glutamine, and Glutamic acid with solubility. These amino acids are polar or charged and can favorably interact with the water molecules in the liquid phase. Similarly, the non-polar and hydrophobic residues would be expected to be negatively correlated with solubility. Therefore, it was not unexpected that this was the case with Phenylalanine, Methionine, and Proline. However, there were some correlations that did not appear to make sense, such as the positive correlation between solubility and Isoleucine and Leucine (both hydrophobic amino acids). This was also observed with the negative correlation between solubility and Cysteine, Serine, Threonine, and Tyrosine.

Additional features that are predictive of whether an over-expressed protein will be soluble (Yes/No) include the dipeptide (Idicula-Thomas et al, 2006) or tripeptide frequency (Idicula-Thomas and Balaji, 2005), and length (Goh et al., 2004). Other variables that had a positive correlation with solubility include Aliphatic Index (Idicula-Thomas & Balaji, 2005; Idicula-Thomas et al, 2006), Instability Index (both the whole protein and the N-terminal region), the net charge (Idicula-Thomas et al, 2006), and the secondary structure (α-helices; Idicula-Thomas and Balaji, 2005). Similarly, other variables have a negative correlation with solubility including hydrophobicity (Christendat et al., 2000; Bertone et al., 2001; Bussow et al., 2004; Dyson et al., 2004; Luan et al., 2004;), the presence of a signal sequence, transmembrane helices (Luan et al., 2004), low complexity regions (Dyson et al., 2004), molecular weight (Dyson et al., 2004), and secondary structure (β-sheets; Idicula-Thomas and Balaji, 2005).

Once a protein is expressed in a soluble form, it is relatively easy to purify, assuming that the protein is tagged for easy purification. Similar to the earlier steps, the ability to purify an expressed protein (Yes/No) has also been correlated with the protein's biophysical properties. Braun et al. (2002) examined the ability to purify human proteins in a bacterial expression system, Escherichia coli (E. coli). They found that proteins with certain domains, like ras-like domains (n=15) or protein kinases (n=10) in their data set had an 80%+ success rate for purification, while proteins from the seven-transmembrane-domain-receptors (n=4), Ephrin (n=4), or tumor necrosis factor (TNF) domains (n=4) all failed the purification process. In these

cases, the domains were defined using the Pfam database (Bateman et al., 2004). Subcellular localization, as defined by Gene Ontology (GO; The Gene Ontology Consortium, 2000), was also found to correlate with propensity to purify. Cytoskeleton (n=6) and DNA associated proteins (n=38) had a high probability of purifying (75%+), while integral membrane and extracellular proteins had a lower probability (25%) of purifying (Braun et al., 2002). In a later study using decision trees and random forests, Goh et al. (2004) found decision rules based upon the amino acid composition of Aspartic acid + Glutamic acid, Asparagine + Glutamine, and the small hydrophobic residues (Glycine + Alanine + Valine + Leucine + Isoleucine), along with pI and proteins that are conserved across organisms predictive of the purification step. Following the rules developed, a protein could be predicted as being able to be purified or not.

Once a protein has been purified in sufficient quantities, crystallization attempts are performed. Crystallization is considered the last step, at which a protein's biophysical properties correlate with success, because once a 'good' crystal is obtained, it should just be a matter of analysis time to solve the 3D structure.

Crystallization, which is the focus of this dissertation, also had biophysical properties associated with success or failure. Similar to other steps in the process, the amino acid content was predictive of the propensity to crystallize (Christendat et al., 2000; Bertone, et al., 2001; Canaves et al., 2004; Goh et al., 2004). This has been shown for both individual amino acids composition, such as Alanine, Asparagine, Aspartic acid, Tyrosine, Serine, and Methionine, and the cumulative composition of 'similar' amino acids, such as all charged residues, the acidic (Aspartic acid + Glutamic acid) and/or the basic residues, small hydrophobic residues (Glycine + Alanine + Valine + Leucine + Isoleucine). This is especially true for the acidic and basic amino acids, which are the ones that have side chains that can be charged. For example, Bertone et al. (2001) found that proteins with an Aspartic acid composition greater than 5% have a higher probability of crystallizing. The correlation of the aliphatic amino acids, Alanine, Valine, Isoleucine, and Leucine, which give a measure of the protein's stability (Aliphatic Index; Ikai, 1980), with solubility, might also not be unexpected. Proteins that are more stable might be hypothesized to crystallize more readily. Conversely, Christendat et al. (2000) found that proteins with an Asparagine composition greater than 3.5% have a lower probability of crystallizing. Canaves et al. (2004) found that acidic proteins, based on pI values 5.1-7.5, had a higher probability of crystallizing. Additionally, they found that signal peptides, proteins with

transmembrane helices, those with a GRAVY value (hydrophobicity measure; Kyte and Doolittle, 1982) below –1.0 or above –0.1, or proteins with low complexity regions are less likely to crystallize. The lower probability of the first three to crystallize is probably due to the presence of a significant number of hydrophobic residues within these proteins. Proteins with regions of low complexity are another issue. These proteins are generally non-globular and have mobile domains. Low complexity proteins are often elongated, displaying extended coils and helical structures (Wootton, 1994).

Differences were also found based on the protein's secondary structure and amino acid length. Examining proteins whose structures have been solved by X-ray and NMR or NMR only, Valafar et al. (2002) concluded that proteins with β-sheets were less likely to crystallize than proteins with α-helices. This was inferred by few protein structures with β-sheets being solved by both methods (Valafar et al., 2002). Idicula-Thomas and Balaji (2005) also found that proteins with a higher occurrence of amino acids that have a propensity to form β-sheets displayed lower solubility. Not surprisingly, the length of the protein was also correlated with propensity to crystallize (Christendat et al., 2000; Bertone et al., 2001; Canaves et al., 2004). Proteins with a sequence length <80 residues (short proteins) and >560 residues (long proteins) crystallize less frequently (Canaves et al., 2004). Therefore, it was felt that similar biophysical properties might also indicate what solution conditions could lead to success, i.e. crystals.

2.3 PROTEIN SOLUBILITY

In order to proceed with crystallization attempts, the target protein should have a moderate level of solubility, usually ≥10 milligrams/milliliter (mg/ml). A protein's solubility is the maximum concentration in liquid phase in equilibrium with the solid phase as either a precipitate or a crystal, such that there is no net loss or gain of the solid phase. The protein's solubility is determined by its biophysical properties, the solvent, and the interactions between the two. The protein's solubility is a key factor in understanding its crystallization behavior. Currently, there are no methods to predict the solubility of a protein. This can only be determined experimentally, which is tedious and requires the use of the protein sample.

2.3.1 Solute Properties

The solubility of the solute (protein) is determined by its biophysical properties, including its hydrophobicity and net charge. The hydrophobic amino acids (Alanine, Isoleucine, Leucine, Methionine, Phenylalanine, and Valine) fold into the interior of the protein to minimize contact with any polar (water) or charged (salts) solvent components. This is believed to be a key factor in the folding of globular proteins. Because of this, proteins with large stretches of hydrophobic regions, such as transmembrane proteins, generally have a low solubility. In order to solubilize membrane proteins, researchers have to include a detergent in the solution to achieve maximal solubility. While the hydrophobic residues tend to fold into the interior, the hydrophilic residues are found on the surface where they can interact with the solvent.

The protein's charge has been known to be a major determinant of its solubility. There are certain residues on the proteins that can be protonated or deprotonated, such as the amino- and carboxyl-terminus, respectively. Additionally, the side chains of the acidic (Aspartic acid, Glutamic acid, Cysteine, and Tyrosine) and basic residues (Histidine, Lysine, and Arginine) can take on a negative or positive charge, respectively. The charge of these residues is largely controlled by the solution pH. The solution pH gives a measure of the amount of free protons (H^+) in the solution. When there is an excess amount of H^+ in the solution, the basic residues will pick up a proton and become positively charged. Alternatively, when there is a low abundance of protons, the carboxyl groups of Glutamic acid and Aspartic acid will give up a proton and become negatively charged. At neutral pH values, the acidic residues are negatively charged and the basic residues are positively charged. These charged molecules then can form hydrogen bonds with the water molecules to neutralize the charge and shield nearby molecules from the charge. The overall effect of the solution pH on a protein's charge is observed by calculating an estimated titration curve.

The estimated titration curve plots the proteins estimated net charge (Q) as a function of the solution pH (Figure 2.1). The point along the curve where the protein has a zero net charge is the pI_{est}. This curve is based upon the amino acid composition of the protein and is calculated using the Henderson-Hasselbach Equation (HHE; Equation 2.1). The HHE uses model pK_a values for each titratable residue, which represents the pH value where 50% of the residues would be charged and 50% would be uncharged. Thus, a protein's solubility can change

dramatically around the pK_a values. Additionally, a protein's solubility is usually minimal at the pI_{est}. This is largely due to the reduction of repulsive electrostatic interactions between protein molecules. Therefore, it was hypothesized that there would be a correlation between the pI_{est} and the pH_{cryst}.

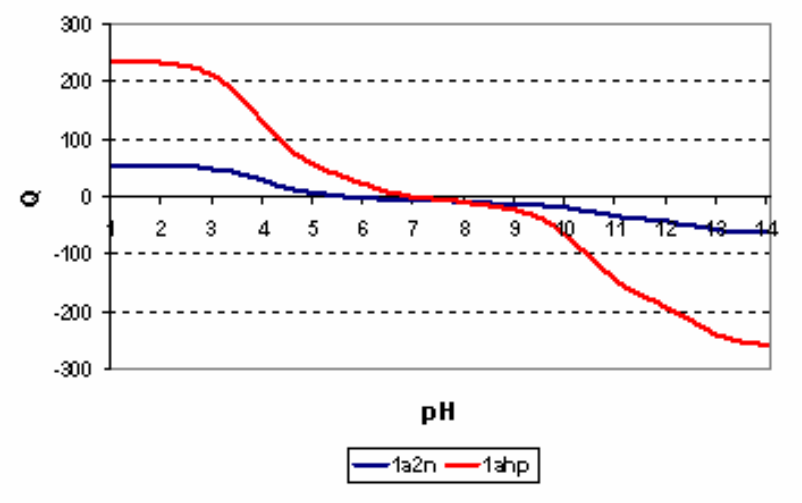


Figure 2.1. An example estimated titration curve for PDB proteins 1A2N and 1AHP.

When calculating the estimated titration curve, several assumptions are made, which may or may not be true. The first assumption is that all like amino acids have the same pK_a value. For example, all Aspartic acids have the same pK_a value and 50% are negatively charged at a pH of 3.65, its pK_a value. Secondly, it is assumed that all titratable groups can be protonated or deprotonated. This is not always the case, because not all charged amino acids are present on the surface. Some charged amino acids are required to interact with other molecules, which requiring the residue to have a charged state, which may cause the pK_a of that residue to shift several pH units. The next assumption is that all charged groups are accessible, i.e. on the surface. Finally, for the estimation of net charge it is assumed that the protein is monomeric, because until the structure is solved the residues located in the interface between molecules can not be known (Ries-Kautt and Ducruix, 1997; 1999).

Equation 2.1 $$Q(AA_{au}, pH) = \Sigma_{(+)}^{\alpha NH_2,H,K,R} n_{(+)} * \frac{10^{-pH}}{10^{-pKa} + 10^{-pH}} - \Sigma_{(-)}^{\alpha COOH,C,D,E,Y} n_{(-)} * \frac{10^{-pKa}}{10^{-pKa} + 10^{-pH}}$$

It has long been hypothesized that there should be a correlation between the pH_{cryst} and its pI, where its estimated net charge is zero. This was for two main reasons: (1) a protein usually exhibits minimum solubility at its pI and (2) a low net charge lowers the probability of unfavorable electrostatic interactions between protein molecules (McPherson, 1999). This was demonstrated in early protein solubility literature when crystalline pepsin (Northrop, 1930), crystalline hemoglobin (Green, 1931a), crystalline catalase (Sumner and Dounce, 1937), crystalline β-lactoglobulin (Gronwall, 1942), crystalline insulin (Fredericq and Neurath, 1950), and more recently with amorphous fibrinogen (Leavis and Rothstein, 1974), amorphous lysozyme (Shih et al., 1992), crystalline ovalbumin (Judge et al., 1996), and crystalline ribonuclease Sa (Shaw et al., 2001) all displayed a minimum solubility at their pI. As the solution pH was moved away in either direction from the pI, the protein's solubility increased. It is well known that crystallizing a protein requires a slow reduction in solubility for the protein to come out of solution as a well-ordered crystal.

However, the hypothesized correlation between pI_{est} and pH_{cryst} has not been supported in practice or by the literature (McPherson, 1999; Page et al., 2003; Kantardjieff and Rupp, 2004). In a large scale structural genomics approach to protein crystallization, Page et al. (2003) attempted to crystallize 539 Thermotoga maritima open-reading frames (ORFs) that had been successfully cloned, expressed, and purified. Of the 465 proteins that did crystallize (86% success rate), the correlation between pI_{est} and pH_{cryst} was quite low ($r = 0.01$).

Meanwhile, Kantardjieff and Rupp (2004) took a data mining approach and examined 9,596 proteins that had previously been crystallized from the PDB and similarly found a low correlation ($r < 0.10$) between the pI_{est} and pH_{cryst}. Although this correlation was statistically significant, such a low correlation is unlikely to provide much information in determining initial crystallization conditions.

2.3.2 Solvent Properties

In addition to the protein's biophysical properties, the solvent may interact with the protein or other solvent compounds to alter the protein's solubility. Solvent properties, such as pH, temperature, and ionic strength of the solution, play critical roles in determining the

protein's solubility. Water molecules form hydrogen bonds with the protein, which keeps the protein solvated. While some solvent components act to increase the solubility, others act to decrease the solubility. Alternatively, a compound may increase the solubility at one concentration, but decrease the solubility at another concentration. This is the case with salts. Salts have been known to change the protein's solubility, stability, and activity.

The protein's solubility can generally be increased with a low concentration of salt, a phenomenon called salting-in. A low concentration of salt stabilizes the protein through non-specific electrostatic interactions. The salt ions act to shield the charge or bind to the protein, allowing the protein molecules to come closer together (decrease repulsion between protein molecules). After the initial increase in solubility, additional salt will decrease the solubility (salting-out). The decrease in solubility is mainly due to hydrophobic interactions caused by the competition of the salt ions and the protein for the solvent (water) to remain solvated. This forces the protein molecules to interact with each other. These protein-protein interactions are what eventually cause the protein to come out of solution, either as a crystal or a precipitate. The particular effects of a salt vary depending upon the type of cation and anion.

The effectiveness of a given salt to reduce the protein's solubility is predicted to some degree by the Hofmeister series and the charge of the protein. The Hofmeister series ranks the ability of different anions and cations to stabilize (salt-out) or destabilize (salt-in) proteins. The cations rank $NH_4^+ > K^+ > Na^+ > Cs^+ > Li^+ > Mg^{2+} > Ca^{2+} > Ba^{2+}$ in their efficiency to lower a protein's solubility, regardless of the protein's charge (Cacace et al., 1997; Ries-Kautt and Ducruix, 1999). The effect of salts on a protein's solubility is mainly driven by the anions. However, the effects are additive over all the ions in the liquid phase. When considering the effect of the anions, the charge of the protein matters. When the protein has a net negative charge, the ability of the anion to lower solubility follows the Hofmeister series, $F^- > PO_4^{3-} > SO_4^{2-} > CH_3COO^- > Cl^- > Br^- > I^- > CNS^-$. This is generally the case for acidic proteins, where the solution pH is above their pI and the net charge is negative. However, when the protein has a net positive charge the Hofmeister series is reversed, often referred to inversion of the Hofmeister series. This is usually the case with basic proteins, when the solution pH is below their pI, resulting in a net positive charge.

2.4 PROTEIN CRYSTALLIZATION

Once a protein is available in a purified form, several experimental methods can be applied to determine its three-dimensional (3D) structure (Table 2.2). The most widely used method as determined by the frequency of structures in the PDB remains X-ray crystallography (84.7%), while the second most popular method is Nuclear Magnetic Resonance (NMR) spectroscopy (14.7%). Currently, these two methods account for 99.4% of the structures in the PDB. An advantage of NMR over crystallography is that the molecule is studied in its native state in the liquid phase or within a membrane. There is no need to grow a crystal, which is a difficult task as discussed throughout this dissertation. However, NMR is limited in its ability to determine the structure of monomeric proteins <40 kilodaltons (kDa) in size or multi-meric proteins <60 kDa (Widmer and Jahnke, 2004), while there are no protein size limits for X-ray diffraction studies. An additional drawback is that the protein studied needs to be highly soluble for NMR studies. However, many proteins are not highly soluble as observed by the fraction of soluble proteins (47%) obtained from the expressed proteins in Table 2.1. As a comparison to X-ray diffraction, the length of data collection for NMR ranges from 45-60 days, while analysis takes an additional 6-12 months (Eisenstein et al., 2000). This dissertation focuses primarily on protein crystallization and X-ray diffraction because of its popularity and widespread use among structural biologists. However, it is hypothesized that similar methods and findings from this dissertation may be applied in the future to the other methods, such as NMR.

In order to determine the X-ray structure of the protein, the purified protein has to first be crystallized. This usually involves an undersaturated liquid phase and slowly creating a supersaturated state, where there is more protein present in the liquid phase than can be maintained in equilibrium (Section 2.3). This leads to the formation of a solid phase, ideally in the form a large well-ordered crystal suitable for X-ray diffraction studies. However, when the protein forms a solid phase, the formation of crystals is not the only possible result. If the protein solution is very highly supersaturated, the protein may form an amorphous precipitate. Alternatively, the protein may undergo a phase transition as a shower of microcrystals or separate into a gel (liquid-liquid phase separation). However, these negative results may still be informative. For example, crystallization is often preceded by a liquid-liquid phase separation

(Section 2.5.2). A slight adjustment in solution parameters around these regions of phase transitions (liquid-solid or liquid-liquid) may result in a large crystal.

Typically, proteins are initially screened against 100's to 1000's of different solution conditions. For instance, the Hauptmann Woodward Medical Research Institute (HWI) in Buffalo, New York typically screens each protein against 1,536 different solution conditions (Luft et al., 2003; http://www.hwi.buffalo.edu/). These initial screens may lead to diffraction quality crystals or may indicate where follow-up experiments should be directed from the observation of some form of crystalline material ('hits') in a particular solution. The realistic goal of the initial screen is to find areas in the crystallization search space where protein crystals are obtained and then further optimized for quality. The follow-up experiments (optimization) usually narrow down the ranges/levels of two to three variables examined around the 'hits,' trying to produce one large good quality crystal, i.e. Grid screens.

Historically, crystallization conditions were initially searched in a lab specific manner with individual researchers having their own preference for reagents and conditions. This changed with the development of the sparse matrix crystallization screens by Jancarik and Kim (1991). These screens were designed using conditions that had succeeded in crystallizing other proteins as compiled from the literature, initially sampling pH values approximately every 1.0 unit from 4.5 to 8.5 (4.6, 5.6, 6.5, 7.5, and 8.5). In this method, a large number of conditions are examined by varying the partial combinations of the concentrations and type of salt, buffer, and precipitants. These screens caught on and were commercialized and are currently used by many researchers due to their ease of use and proven track record. However, this approach is more of a 'one-size-fits all' approach, assuming that all proteins will follow a similar pattern. Since then, other researchers have developed and published their own sparse matrix screens (Doudna et al., 1993; Cudney et al., 1994; Zeelen et al., 1994; Scott et al., 1995; Harris et al., 1995; Brzozowski and Walton, 2001; Radaev and Sun, 2002; Iwata, 2003; Majeed et al., 2003; Tran et al., 2004; Gao et la., 2005; Radaev et al., 2006).

Others have taken a more mathematical/statistical approach to crystallization screen design. These approaches typically design screens composed of an ordered selection from the experimental parameter space. These approaches include the factorial and incomplete factorial experiments (Carter and Carter, 1979; Carter et al., 1988; Sedzik, 1995; Shieh et al., 1995), along with orthogonal arrays (Kingston et al., 1994), or Bayesian approaches (Hennessy et al., 2000).

The complete factorial experiment samples every combination of available parameters, which is typically not feasible in terms of time, cost, and protein volume. The incomplete factorial takes an ordered subset of the complete factorial. After coding the results, including the failures, statistical analysis is performed to identifying the important solution parameters that are involved in forming a high-quality crystal.

Bayesian approaches examine the occurrences of each solution parameter in other proteins (Hennessy et al., 2000; Rupp and Wang, 2004). Again, these proteins can have varying degrees of similarity to the protein of interest. These successes are tabulated and used as priors to predict the probability of success for a new protein. These probabilities can be combined over multiple features.

Reducing the time spent in searching areas for crystallization conditions that are not likely to produce crystals would increase the chances of yielding a high quality crystal. Focusing more attention on the areas that have a higher probability of yielding crystals can do this. A unique approach for optimization was presented recently by DeLucas et al. (2003). They used the results of incomplete factorial experiments and fed the coded results into a neural network, which was then used to predict the outcomes of experimental conditions not previously seen by the trained neural network. The trained neural network correctly predicted crystallization conditions for the test set conditions in all cases. This approach again shows the importance of negative examples, i.e. no crystals or the presence of phase transitions, which are used in training the neural networks.

2.5 METHODS TO INCREASE THE PROBABILITY OF GROWING A CRYSTAL

There are several methods that have been used to increase the probability of growing a crystal. The most widely used method is to simply use the crystallization solution conditions that have worked previously for other proteins (sparse-matrix approach; Section 2.5.1). Another method used is to observe the protein's phase behavior as a function of varying solution parameters, such as the protein and precipitating agent concentrations, which can then be used to create a phase diagram. This diagram can then indicate what concentration values may be more suited for forming (nucleation) and growing crystals (Section 2.5.2). Alternatively, the protein-protein

interactions in the liquid phase are measured using light scattering techniques (Section 2.5.3). Although these methods have proven useful for identifying the conditions that can lead to the formation of crystals, the creation of a phase diagram and light scattering studies are not routinely performed. One reason is that these two methods require the use of the protein sample, which may be in short supply.

2.5.1 Crystal Screens

Since Jancarik and Kim's original sparse matrix screen, others have developed more specific sparse matrix screens for RNA (Doudna et al., 1993; Scott et al., 1995 Golden et al., 1997), immunoglobulins (Harris et al., 1995), enzymes (Brzozowski & Walton, 2001), protein-protein complexes (Radaev & Sun, 2002; Radaev et al., 2006), and membrane proteins (Iwata, 2003). These studies and others demonstrated that particular classes of macromolecules have a preference for solution conditions that lead to successful crystallization (Samudzi et al., 1992; Farr et al., 1998; Hennessy et al. 2000; Jurisica et al., 2001; Gilliland et al. 2002; Kimber et al. 2003; Goh et al., 2004; Rupp and Wang, 2004). The successful experimental conditions that have been used to grow the crystal can often be found in the literature, the Biological Macromolecule Crystallization Database (BMCD; Gilliland et al., 1994), or the PDB. Most successful conditions are listed within the published article where the 3D structure was first described. However, it would take a huge amount of effort to obtain all of this information from free text, including a large amount not available in electronic format.

Gary Gilliland, who manually extracted the information from the published literature into his notebook, originally created the BMCD in the late 1980's, which had information on 600 different biological macromolecules (Gilliland, 1988). The BMCD contains detailed information on the conditions (solution and environmental) used to grow the crystals. This effort has been abandoned as no new crystal entries have been deposited into the BMCD since 1997. Since that time, the PDB has been including fields for crystallization solution parameters.

However, the value of using the BMCD for crystallization screen design was shown in later studies. Using the BMCD, certain classes or families of proteins were shown to have a tendency to crystallize under a more narrow range of environmental conditions (Samudzi et al., 1992; Farr et al., 1998; Hennessy et al. 2000). For example, Hennessey et al. (2000) found that

proteins they described as 'ligand binding proteins' and 'enzymes' had significantly different pH_{cryst} distributions. This prior knowledge is encoded as Bayesian priors and used to generate probability distributions for various solution parameters, including pH and temperature. These probabilities are further combined over multiple crystallization variables as more data, both positive and negative, is collected. The solution probabilities are then rank-ordered by the probability of successfully obtaining a crystal suitable for diffraction studies to suggest regions in the crystallization search space more likely to produce well-ordered crystals. These rank-ordered conditions can then be chosen for the initial crystallization attempts. This is accomplished by selecting the conditions that have a higher probability in generating crystals. Even a collection of weak predictors is informative and can aid in crystallization design (Hennessy et al., 2000).

Information on a limited number of crystallization parameters can also be found in the PDB. Each structure within the PDB has a set of standard information collected, including a limited set of crystallization parameters, and put into a file, the macromolecular Crystallographic Information File (mmCIF), which provides standard annotations for data uniformity. The experimental conditions that have been used to grow the crystal are starting to appear more frequently in the mmCIF files (Table 2.2); however, most are still optional (Bourne et al., 1997). The pH_{cryst} is the most recorded crystallization parameter for PDB structures solved by X-ray diffraction, 21,602/26,995 (80%) on 11/14/2005. The pH_{cryst} is obtained from the __exptl_crystal_grow.pH field of the mmCIF file (Table 2.2a). Other experimental parameters, such as the method and temperature, may be listed within the __exptl_crystal_grow or __exptl_crystal_grow_comp fields. The __exptl_crystal_grow.detail field is in the form of free text, while the __exptl_crystal_grow_comp field contains more structured text components, such as the component name, concentration, and volume (Table 2.2b). However, these fields are still optional and not highly populated.

Table 2.2 mmCIF (macromolecular Crystallography Information File) specifications for crystallization conditions (http://ndbserver.rutgers.edu/mmcif/).

(a) __exptl_crystal_grow

Item Name	Mandatory Code	Example Entry
_exptl_crystal_grow.crystal_id	Yes	1
_exptl_crystal_grow.apparatus	No	Linbro plates
_exptl_crystal_grow.atmosphere	No	Room air
_exptl_crystal_grow.details	No	
_exptl_crystal_grow.method	No	Hanging drop
_exptl_crystal_grow.method_ref	No	McPherson et al., 1988
_exptl_crystal_grow.pH	No	5.5
_exptl_crystal_grow.pressure	No	
_exptl_crystal_grow.pressure_esd	No	
_exptl_crystal_grow.seeding	No	macroseeding
_exptl_crystal_grow.seeding_ref	No	Stura et al., 1989
_exptl_crystal_grow.temp	No	298
_exptl_crystal_grow.temp_details	No	?
_exptl_crystal_grow.temp_esd	No	
_exptl_crystal_grow.time	No	2-4 days

(b) __exptl_crystal_grow.comp fields

Item Name	Mandatory Code	Example
_exptl_crystal_grow.comp.conc	No	0.1 ml
_exptl_crystal_grow.comp.crystal_id	Yes	1
_exptl_crystal_grow.comp.details	No	in 3 mM NaAzide
_exptl_crystal_grow.comp.id	Yes	1
_exptl_crystal_grow.comp.name	No	Acetic acid
_exptl_crystal_grow.comp.sol_id	No	Solution A
_exptl_crystal_grow.comp.volume	No	0.1 ml

2.5.2 Phase Diagrams

Another method that may guide the crystallographer to selecting solution conditions that have a higher probability of generating crystals, but is not yet common practice, is the creation of a phase diagram (solubility diagram; Odahara et al., 1994; Saridakis et al., 1994; Shaw Stewart and Khimasia, 1994; Galkin and Vekilov, 2001; Santesson et al., 2003; Saridakis and Chayen, 2003; Asherie, 2004; Collins et al., 2004; Saijo et al., 2005; Sommer and Larson, 2005; Vivares et al, 2005). A phase diagram typically displays the resulting phase (solid, liquid, etc.) of the target protein as a function of 2-3 solution parameters, often the protein's concentration by the precipitating agent's concentration.

A phase diagram allows a researcher to determine the target protein's solubility curve, which represents the border between undersaturation and supersaturation. Along the solubility curve, the protein in the liquid phase is at equilibrium with the solid phase (precipitate or crystal). In areas below the solubility curve (undersaturation), crystals will not form, because the protein is stable in the liquid phase. However, in areas above the solubility curve, there is more protein in the liquid phase than the system can contain at equilibrium, which causes some of the protein to form a solid phase until equilibrium is reached.

There are three areas of interest above the solubility curve, the precipitation, nucleation, and metastable zones. At very high levels of supersaturation, precipitate will spontaneously form in the solution, the precipitation zone. The nucleation zone occurs at slightly lower levels of supersaturation, where crystals will form (nucleate). This is the area that needs to be found for successful crystallization. Depending upon the level of saturation within this area, there could be a few or many crystal forming. In the metastable zone, which is immediately above the solubility curve, crystals cannot form, but they can grow. Knowledge of the supersaturation zones should increase the probability of generating protein crystals.

From the various observations of phase transitions of the test protein, the borders between the saturation zones are assessed. These observations can suggest areas of supersaturation that are more likely to generate crystals, i.e. where nucleation takes place. Nucleation is usually preceded by the formation of a metastable liquid-liquid phase separation. This involves the formation of a droplet that has a high protein concentration and the remaining solution that has a low protein concentration. It is within this droplet that the nucleation of a crystal takes place

(Haas and Drenth, 1999). Alternatively, regions where the protein is undersaturated (no observed phase separation) should be avoided, because these conditions are unlikely to yield crystals in a timely manner. However, the creation of a phase diagram is tedious and requires the use of the protein sample, which may be in short supply.

2.5.3 Light Scattering

Light scattering techniques are another experimental method that is used as a pre-screening method and can explain protein crystallization behavior. These methods measure the interactions of protein in the liquid phase and can indicate whether the solution in which a protein resides can produce crystals (Kam et al., 1978). The two most popular methods are dynamic light scattering (DLS) and static light scattering (SLS). For both of these methods, the scattering of light through the protein solution is measured using different protein concentrations at either a fixed (DLS) or different (SLS) scattering angles. Due to their tedious nature and use of protein, which may be in short supply, these methods are not routinely performed prior to the initial crystallization attempts.

Dynamic Light Scattering (DLS) measures the dispersity of the protein in the liquid phase. Depending upon the amount of light scattered, this measurement gives an indication of the aggregation state of the protein molecules, indicating whether there are different sized protein aggregates. If only one size aggregate is present (monodisperse), there is a higher probability of crystallization. However, if more than one size of aggregate is present in the solution (polydisperse), crystals generally will not form. Additionally, an interaction parameter is calculated which signifies whether the interactions between molecules are attractive (negative values) or repulsive (positive values). This technique can also be used to test individual additives, whether they increase or decrease the aggregation, to give an idea whether the additive may promote crystallization (Veesler and Boistelle, 1999; Wilson, 2003).

Another light scattering technique that has been used successfully to determine whether a solution can promote crystallization is Static Light Scattering (SLS). SLS is used to measure the second virial coefficient (B_{22}), which is a dilute solution parameter that gives a measure of the adhesive hard sphere potential between two protein molecules. Similar to the measurement of polydispersity by DLS, negative B_{22} values indicate an attractive interaction between protein

molecules in the liquid phase, while positive values indicate a repulsive interaction. George and Wilson (1994) measured the B_{22} of nine different proteins in solutions that grew crystals and discovered that using a wide range of precipitating agents, crystals formed only in a narrow region of slightly negative B_{22} values ($-1x10^{-4}$ to $-8x10^{-4}$ moles*milliliters/gram^2), the 'crystallization slot,' where protein molecules are slightly to moderately attractive. Interactions that are more negative have a greater probability of forming amorphous precipitate than crystals. There is also some variation within the crystallization slot, with fewer larger crystals being formed at the more positive end of the slot and many small crystals being formed at the more negative end (Wilson, 2003). This method was also shown to work with membrane proteins (Hitscherich et al., 2000), a special subset of proteins that are difficult to crystallize.

In addition to its use for predicting crystallization conditions, a correlation between B_{22} and solubility was also demonstrated for a number of proteins, including lysozyme (Rosenbaum and Zukoski, 1996; George et al., 1997; Guo et al., 1999; Piazza and Pierno, 2000), ovalbumin (Guo et al., 1999; Demoruelle et al., 2002), and serum albumin (George et al., 1997; Demoruelle et al., 2002). This correlation matches quite well between the 'crystallization slot' of B_{22} and the observed liquid-liquid phase separation (Haas and Drenth, 1999). The second virial coefficient (B_{22}) depends strongly on the solution pH and ionic strength (Haynes et al., 1992). This method has been suggested to replace the direct measurements of solubility, which can often be laborious.

2.5.4 This Dissertation

Similar to the other methods discussed in Sections 2.5.1-2.5.3, this dissertation aims to increase the probability of forming a protein crystal in the initial screen. However, unlike the creation of a phase diagram (Section 2.5.2) or using light scattering techniques (Section 2.5.3), the methods discussed here use no protein sample. The methods used here are purely computational. Any method that increases the probability of crystal formation, should decrease the time and cost of the structure determination process. The unique approach presented here attempts to use protein sequence information to suggest the most likely pH regions, which if successful should increase the probability of obtaining a crystal. This is unlike the typical screening procedures, which generally use a one-size-fit all approach (Section 2.5.1). The resulting knowledge can also be

used to gain insight into the crystallization process and explain the idiosyncratic nature of proteins (why some solution conditions are more effective at generating crystals for certain proteins than others).

2.6 TYPES OF CRYSTALLIZATION VARIABLES

In order to begin to understand the crystallization process, one must first understand the types of variables that are involved and are available. McPherson (1999) lists thirty-six variables that are believed to play an important role in the crystallization process that were listed in Table 1.1. These variables are broken down further into two types of crystallization variables that will be examined and discussed, *Givens/Features* (Section 2.6.1) and *Controllables* (Section 2.6.2). The results from any crystallization experiment are the *Observables* (Section 2.6.3). Knowledge of these variables and their interactions should help in obtaining a desirable outcome, the formation of a crystal suitable for diffraction studies.

2.6.1 Givens

Givens refer to those variables that are known or can be estimated prior to any crystallization attempts. Most of the features listed in the Biochemical section of Table 1.1 would be considered *Givens*. The *Givens* examined in this dissertation include *Features* about the protein and previously successful experimental conditions. The *Primary Given Feature* for protein crystallization is the amino acid sequence, which is available from the PDB. From a protein's amino acid sequence, other biophysical properties can be calculated or estimated, such as the molecular weight, estimated titration curve, and the pI_{est}, which are not present in the PDB, *Hidden Features*. Determining the importance of each *Feature* in the crystallization process could lead to a better understanding of the process as a whole and could possibly be used for improving the success rate for future crystallization attempts. Currently, no one has demonstrated a useful correlation between a protein's biophysical properties (*Features*) and the specific solution conditions used for crystallization (*Controllables*), although several researchers

have suggested that there should be a link between the two (Heinemann et al., 2000; Kimber et al., 2003; Canaves et al., 2004; Rupp and Wang, 2004; Bussow et al., 2004). However, there have been some recent studies showing that a protein's biophysical properties can give an idea as to whether a protein is amenable to each step in the structural determination process (Section 2.2).

2.6.2 Controllables

Controllables refer to the variables that are manipulated in the experiment. These include the pH, temperature, type of precipitating agent, and concentration. A more complete list is found in the Physical and Chemical fields of Table 1.1. As mentioned previously, most of these variables are not present within the major structural databases and are usually only found within individual laboratory notebooks. The solution pH was the initial *Controllable* examined, because of its known importance to the crystallization process and being the most reported crystallization solution condition. Because the solution pH largely controls the charge of the protein, as demonstrated by the estimated titration curve, the estimated net charge (Q) and related variables (the estimated specific charge and estimated average surface charge density) can also be considered as *Hidden Controllables*. If enough data was available in the PDB any *Controllable* can be examined.

2.6.3 Observables

Observables refer to the experimental observations resulting from the interactions of *Features* and *Controllables*, i.e. one experimental well/setup. The results of a given experiment (i.e. well) could be in the form of success or failures, i.e. presence of crystal, 1=yes or 0=no or a more detailed crystal quality scale (Table 2.3), such as that suggested by Carter (1999). *Observables* reported in the form of successes, as listed in the PDB, could include the resolution limit of diffraction ($diff_{lim}$), R factor, and the asymmetric unit composition (number of chains and the molecular weight), which may be different than the biological unit. A good example is presented at the PDB website (http://www.rcsb.org/pdb/ biounit_tutorial.html) using hemoglobin, which

has a biological unit composed of four protein chains. However, the crystal's asymmetric unit can be composed of 1 biological unit (2HHB; four chains), 2 biological units (1HHO; eight chains), or a portion of a biological unit (1HV4; two chains). Observations could also be recorded for failures, such as the presence or absence of precipitates or phase separation. Because it is readily available and its use as the primary indicator of quality, the $diff_{lim}$ was the only observable examined in this dissertation. The $diff_{lim}$ is the closest resolution that two objects can be determined as different.

The distribution of pH_{cryst}, Q_{cryst}, $\overline{Q}_{cryst}$, and σ_{cryst} for previously crystallized proteins would also be considered as *Given Observables*, because they are known or can be estimated prior to any crystallization attempts for a target protein. The goal of this dissertation was to find variables from sequence information (*Features*), such as the amino acid composition, molecular weight, and pI_{est}, which are known apriori in the lab, to predict a pH range (*Controllable*) that will result in crystallization, pH_{cryst} (*Observable*), which is not known apriori in the lab.

Table 2.3 An example crystal quality scale for a given experiment result (well) as suggested by Carter (1999).

Observation	**Score**
Cloudy/Amorphous Precipitate	1.0
Gelatinous/Particulate Precipitate	2.0
Oils	3.0
Spherulites	4.0
Needles	5.0
Plates	6.0
Prisms	7.0

2.7 BACKGROUND SUMMARY

Protein crystallization still remains the most common method of determining the 3D structure of biological macromolecules. However, the crystallization process is a complex event that is not completely understood. The conditions that lead to the successful crystallization of one protein have no effect on another protein. Currently, most researchers try a one-size-fits-all approach to

identifying these solution conditions. While occasionally successful, alternative methods need to be developed to increase the success rate of generating a crystal suitable for X-ray diffraction studies.

Similar to other methods presented in Section 2.5, this dissertation aims to increase the probability of obtaining a protein crystal in the initial screen. Unlike the other methods discussed in Sections 2.5.1-2.5.3 that use the protein sample to determine the solution conditions (*Controllables*) most likely to produce crystals, the unique approach presented here uses protein sequence information (*Features*) to suggest the most likely pH regions to result in the formation of a crystal. Additionally, the approach presented in Chapter 4 is examined on several proteins from the test set (Appendix C) or target proteins whose structure remains to be solved (Chapter 7). The results also give some insight into the crystallization process by offering plausible explanations (Chapter 8). For these reasons, the time and cost of the structure determination process should hopefully decrease.

3.0 INITIAL ANALYSIS OF THE PDB

Within the last couple of years more reported crystallization data has been available in the Protein Data Bank (PDB), therefore a retrospective analysis was performed to retest the relationship between a protein's estimated isoelectric point (pI_{est}) and the reported pH of crystallization (pH_{cryst}). The long held belief that the two would be related has stemmed from the fact that a protein generally has its minimum solubility at the pI_{est} (Section 2.5.1). Because protein crystallization requires the reduction of protein solubility, this seemed quite logical. However, previous attempts (McPherson, 1999; Page et al., 2003; Kantardjieff and Rupp, 2004) have found little evidence of this.

The solution pH is known to largely exhibit its effects of controlling the charge of the protein (causal relationship). Because of this, the estimated net charge at the pH_{cryst} (Q_{cryst}) was estimated using the protein's amino acid sequence, the assumed pK_a values of the titratable amino acid residues, and the pH_{cryst}. From this information, the estimated titration curve is calculated and the Q_{cryst} value inferred from the curve by setting $Q_{cryst} = Q$, where the $pH = pH_{cryst}$. Initially, it was hypothesized that there would be a linear relationship between Q_{cryst} and protein *Features*, particularly the molecular weight and the pI_{est}.

In order to examine this relationship, groups of proteins were formed on the basis of their asymmetric unit molecular weight (MW_{au}) or pI_{est}. Then the distributions of both *Given Features* (MW_{au} and pI_{est}) and *Given Observables* (pH_{cryst} and Q_{cryst}) were examined between the groups to determine whether there are differences in the distribution of *Given Observables*, especially the estimated net charge of the protein, based upon a protein's molecular weight or pI_{est}. The hypothesis was that the estimated net charge (Q) of a target protein over the pH range (pH 1.0-14.0) can be used to guide researchers into selecting the solution pH ranges that are

more likely to result in crystallization. This can theoretically be accomplished by setting the solution pH (*Controllable*) to a value that equals the *Given Observable*, $pH = pH_{cryst}$, by using the solution pH that results in the Q (*Hidden Controllable*) being equal to the Q_{cryst} (*Hidden Observable*), $Q = Q_{cryst}$.

3.1 METHODS

The methods used here for *Feature* extraction, *Feature* construction, and case selection were incorporated into the Protein Sequence-Properties Evaluation Framework (Chapter 4). For more information on the process, see Chapter 4. Version #107 of the Protein Data Bank (January 2004) contained information on 23,792 structures. There were 14,468 structures in version #107 of the PDB that had a pH_{cryst} value. After case selection, the original list structures was reduced to a final data set of 11,518 proteins, labeled PDB_v107. The PDB_v107 data set was allowed to contain redundant proteins. A second data set of non-redundant PDB entries was created using the non-redundant PDB (nrPDB; updated 02/03/04; Holm and Sander, 1998), as described in Section 4.6.3. The final nrPDB_v107 data set contained 3,957 proteins.

After the creation of both data sets, the PDB_v107 and nrPDB_v107, various features were calculated as described in Section 4.5. The slope of the estimated titration curve at both the pI_{est} and pH_{cryst} was calculated, using the pH values 0.1 above and below the pI_{est} and their corresponding charge values. In an attempt to determine whether there is a bias in the data sets, the distributions of variables in both data sets were compared using a Kolmogorov-Smirnov (KS) test (Section 4.9). After creating variables that were not present in the original data set, the intercorrelations among all variables were examined using Spearman's rho (non-parametric) correlations (Section 4.6).

After examining the correlation among variables, several attempts were made to group proteins by similarity and then examine differences in crystallization behavior, pH_{cryst} and Q_{cryst} values. Here, proteins are described as 'similar' by displaying similar values of their MW_{au}, pI_{est} (Section 3.2.1), or estimated titration curves (Section 3.2.2). First, all protein structures were

binned into groups using their MW_{au} (Section 3.2.1.1) or pI_{est} (Section 3.2.1.2). Then the combined effects of each variable, MW group by pI_{est} group, was examined (Section 3.2.1.3).

Next, the ability to use the whole estimated titration curve, rather than one point along the curve (pI_{est}), was used to suggest regions of the pH search space were explored (Section 3.2.2). The estimated titration curves were calculated for all proteins selected from the PDB in units of 0.2 (Figure 3.1a; Section 4.5.3). These estimated titration curves were then clustered using self-organizing maps (SOMs; Kohonen, 2001; Section 4.7.2.2). This unsupervised clustering method appeared to accurately cluster the groups as visually judged by the titration curves; however the clusters primarily separated the proteins by their molecular weight, as judged by either end of the estimated titration curve (pH 1.0 and pH 14.0). In order to put all titration curves on the same scale, each Q value along the titration curve was scaled by subtracting the minimum Q value (Q_{min} always at pH 14.0) and divided by the difference between the maximum Q value (Q_{max} always at pH 1.0) and Q_{min} (Equation 3.1). Thus, a scaled titration curve has a maximum $\widetilde{Q}$ value of 1.0 and a minimum $\widetilde{Q}$ value at 0.0 (Figure 3.1b). This would in theory allow for all proteins to be equally compared, removing much of the molecular weight effect.

Equation 3.1 $$\widetilde{Q} = \frac{Q - Q_{min}}{Q_{max} - Q_{min}}$$

One drawback of the traditional SOM algorithm is that the user has to pre-specify the number of clusters usually with no prior information. This was not an issue if a dynamic SOM is used, which allows the data to determine the appropriate number of clusters. For this procedure, the GSOM package (Hsu et al., 2003) was used (kindly provided by Art Hsu). Growth is controlled by the within cluster error and a growth threshold. When the accumulated error value exceeds the growth threshold within a cluster, the cluster is split into two clusters.

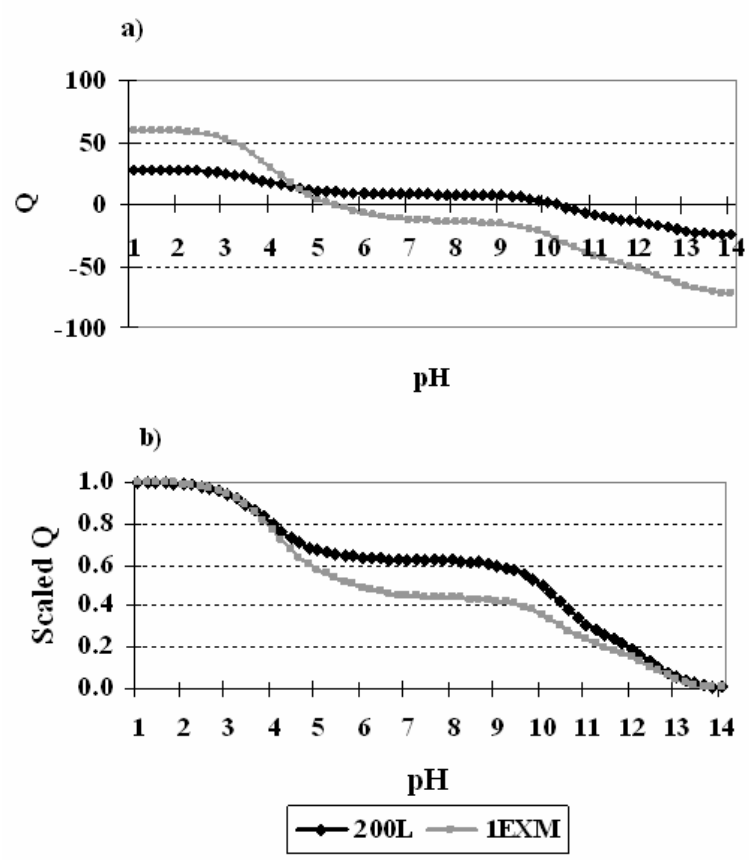


Figure 3.1.The (a) Q or (b) scaled estimated net charge ($\widetilde{Q}$) curves for PDB proteins 200L and 1EXM.

3.2 RESULTS

Two data sets were created and examined in the Preliminary Results to determine the effects of bias in the Protein Data Bank (PDB). The two data sets only differ in the amount of redundant proteins within each set. The PDB_v107 data set allows for any level of redundant proteins, such as the many different lysozyme mutants present in the PDB. The second data set, nrPDB_v107, addresses this issue by removing redundant protein structures with a sequence similarity of BLAST p-value $<10^{-80}$, which reduced the data set by approximately 66% (3,957 nrPDB_v107/11,518 PDB_v107).

First, the variable distributions between the data sets were examined. Several variables, including the slope of the titration curve at both the pI_{est} and pH_{cryst}, and the slope of the scaled

titration curve (SS) at both the pI_{est} and pH_{cryst} had negatively skewed distributions. The distribution of MW_{au} was positively skewed. Therefore, non-parametric correlations (Spearman's rho) and statistical measures (Kolmogorov-Smirnov Z test) were used for analyzing the data. An α-level of 0.001 was used for significance to reduce spurious significant findings due to the large number of cases.

The means and standard deviations (SD) of all variables are shown in Table 3.1. While there appeared to be slight differences in the mean values between the data sets, the actual distributions of most variables were significantly different ($p<0.001$). No differences in the distributions of Q_{cryst} and the slope of the titration curves at both the pI_{est} and pH_{cryst} were found. Additionally, there should not be any differences in the random variable and indeed there were not. It was not clear why the slopes of the titration curves at both the pI_{est} and pH_{cryst} were not significantly different between the data sets, while scaling the slope resulted in a significant difference between data sets. The differences may be due to the MW_{au} and pI_{est} distributions being significantly different. A significant difference between $diff_{lim}$ distributions was also found. This is not unexpected, because the nrPDB_v107 data set is comprised of the 'best' non-redundant structures from the PDB_v107 data set.

Table 3.1 Comparing the PDB_v107 and nrPDB_v107 datasets.

Data Set	n		MW_{au} [b]	pI_{est} [b]	Slope at pI_{est} [b]	SS at pI_{est} [b]	pH_{cryst} [a]	Slope at pH_{cryst} [b]	SS at pH_{cryst} [b]	Q_{cryst} [b]	$diff_{lim}$ [a]	Random Number
PDB_ v107	11,518	Mean	67.7	6.4	-14.5	-0.09	6.7	-9.0	-0.05	-5.3	2.12	0.01
		SD	79.6	1.6	20.6	0.06	1.3	14.2	0.04	24.5	0.48	1.00
nrPDB_ v107	3,957	Mean	65.1	6.3	-15.1	-0.09	6.7	-9.3	-0.05	-5.2	2.08	0.00
		SD	82.9	1.7	21.6	0.06	1.3	15.4	0.05	25.6	0.51	1.02
KS		Z	3.0	2.8	1.6	4.4	2.2	1.7	5.6	0.8	2.6	1.1
Test [c]		p<	0.0001	0.0001	0.009	0.0001	0.0002	0.007	0.0001	0.485	0.0001	0.153

[a] Extracted from the PDB

[b] Calculated using sequence information

[c] Variables distributions were considered as being significantly different if $p < 0.001$ as determined by a KS Test.

The pH_{cryst} histograms for both the PDB_v107 (n = 11,518) and nrPDB_107 (n=3,597) data sets are shown in Figure 3.2b. From the histograms, a saw-tooth pattern was observed at approximately every 0.5 pH unit from pH 4.5-9.0. There appears to be little differences between the two graphs except a larger peak at pH 4.5 for the PDB_v107 data set. This is believed to be caused by the preponderance of lysozyme structures within the full data set. Many of these proteins and their mutants had a pH_{cryst} of 4.5. The distribution of pI_{est} also demonstrated interesting differences (Figure 3.2a). Although these figures show many similarities, there are two points in the PDB_v107 data set that stand out, peaks at a pI_{est} of 9.0 and 10.1. These values correspond to the estimated pI values for hen egg white lysozyme and T4 lysozyme, respectively. It was thought that this would be a potential bias, so further analysis primarily focused on the nrPDB_v107 data set.

After detecting a potential bias in the full data set, the intercorrelations were examined between the *Given Features* and the *Given Observables* for the nrPDB_v107 data set (Table 3.2). *Givens* refer to the variables that are known prior to any crystallization attempts in the lab (Section 2.6.1). Such variables would include a protein's biophysical properties (MW_{au}, pI_{est}, slope of the estimated titration curve at the pI_{est}, and the slope of the scaled titration curve at the pI_{est}). The *Given Observables* refer to the solution conditions that have succeeded in crystallizing proteins as reported in the PDB. The primary *Observable* is the pH_{cryst}, which can also be used with the local *Feature* (primary amino acid sequence) to calculate some *Hidden Features*, such as the estimated net charge, slope of the titration curve, and scaled slope of the estimated titration curve, along with some *Hidden Observables*, such as the Q_{cryst}. The resolution limit of diffraction ($diff_{lim}$) is an *Observable*, which is a quality outcome measure.

Table 3.2 The Spearman's rho correlation values for the variables examined using the PDB_v107 data set (black font) or the nrPDB_v107 data set (red font).

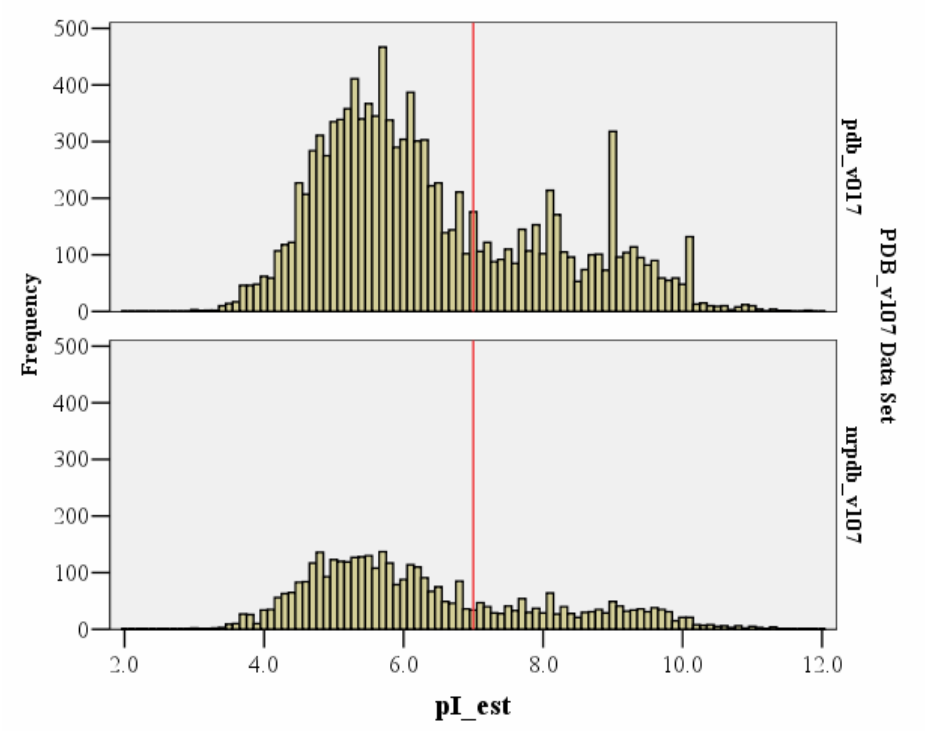
Variable	MW_{au}	pI_{est}	Slope at pI_{est}	SS[a] at pI_{est}	pH_{cryst}	Slope at pH_{cryst}	SS[a] at pH_{cryst}	Q_{cryst}	$diff_{lim}$	Random #
MW (kDa)		-0.043	**-0.692****	0.189**	0.092**	**-0.644****	0.084**	**-0.252****	**0.357****	-0.009
pI_{est}	-0.134**		**0.402****	**0.524****	0.061	0.058	0.031	**0.661****	0.017	-0.024
Slope at pI_{est}	**-0.674****	**0.439****		**0.510****	-0.024	**0.518****	-0.007	**0.440****	**-0.258****	-0.012
SS[a] at pI_{est}	**0.230****	**0.439****	**0.494****		0.074**	-0.073**	0.071**	**0.301****	0.078**	-0.038
pH_{cryst}	0.096**	0.046**	-0.038**	0.069**		**0.336****	**0.566****	**-0.475****	0.049	0.003
Slope at pH_{cryst}	**-0.644****	0.090**	**0.479****	-0.125**	**0.312****		**0.658****	-0.018	**-0.231****	0.009
SS[a] at pH_{cryst}	0.084**	-0.008	-0.049**	0.040**	**0.538****	**0.657****		**-0.305****	0.048	-0.004
Q_{cryst}	**-0.290****	**0.677****	**0.459****	**0.265****	**-0.483****	-0.007	**-0.323****		-0.086**	-0.017
$diff_{lim}$	**0.391****	-0.055**	**-0.275****	0.109**	0.067**	**-0.235****	0.078**	-0.139**		-0.019
Random #	-0.008	-0.008	0.012	0.003	-0.001	0.010	0.006	0.000	0.008	

[a] SS = scaled slope, where the maximum and minimum values for the estimated titration curves were 1 and 0, respectively.

[b] Any correlation greater than 0.20 or less than -0.20 is in bold, while any correlation greater than 0.5 or less than -0.5 are additionally highlighted.

** p-value <0.0001

(a)



(b)

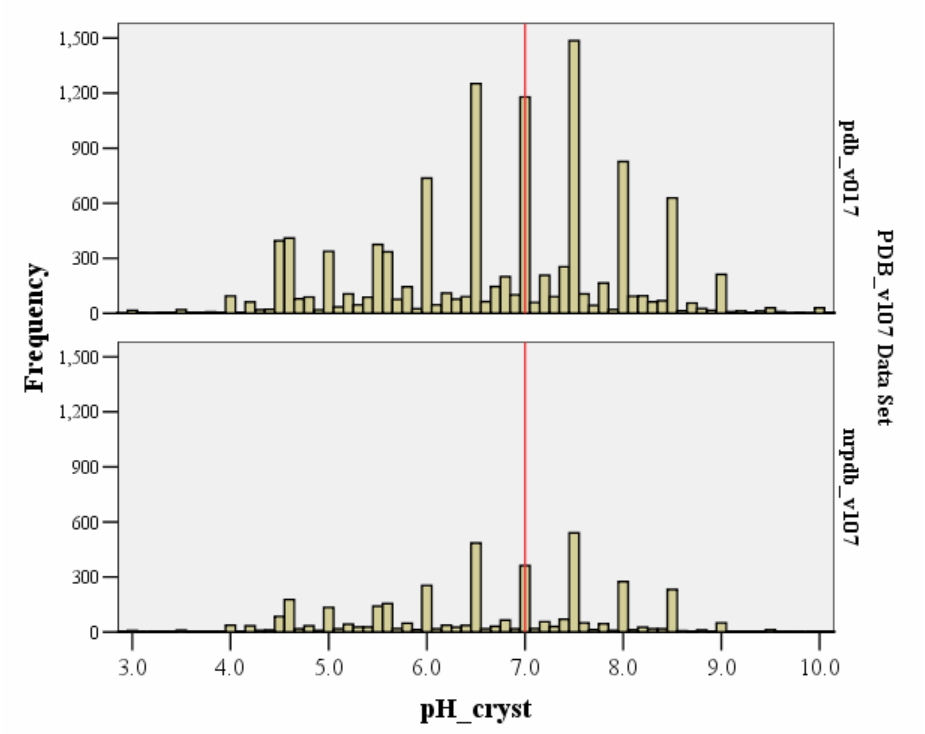


Figure 3.2: The frequency distributions for (a) pI_{est} and (b) pH_{cryst} for both data sets.

The protein's MW_{au} was highly correlated with measures based on the estimated titration curve, slope of the titration curve at both the pI_{est} (r = -0.674) and the pH_{cryst} (r = -0.644) as demonstrated in Table 3.2.

3.2.1 Binning Proteins

The PDB provides the amino acid sequence of all protein chains in the structure's asymmetric unit. From this list, various variables are calculated, including the MW_{au} and the estimated titration curve, from which the pI_{est} is obtained. The effects of molecular weight (MW_{au}) and pI_{est} on the pH_{cryst} were examined by grouping proteins with 'similar' values of each individual variable. It was thought that proteins with 'similar' MW or pI_{est} values would crystallize under similar conditions, including the solution pH. Because the solution pH controls the protein's Q, the estimated net charge at the pH_{cryst} (Q_{cryst}) was also examined. Due to differences in the full and non-redundant data sets' distributions, only the nrPDB_v107 data set was used here. Scatterplots demonstrate the relationships between the MW_{au}, the pI_{est}, the pH_{cryst}, and the Q_{cryst} (Figure 3.3). As previously demonstrated in Table 3.2, little correlation was found between MW_{au} x pH_{cryst} (Figure 3.3a) and pI_{est} x pH_{cryst} (Figure 3.3b). However, a much stronger correlation was observed between MW_{au} x Q_{cryst} (Figure 3.3c). Lower MW proteins had a Q_{cryst} distribution more tightly centered on zero. As the proteins grew in size, the Q_{cryst} values spread out in both the negative and positive directions. An interesting pattern was also observed for the scatterplot of pI_{est} x Q_{cryst}, which were also significantly correlated (Figure 3.3d). Proteins with a pI_{est} around 7.0 have a Q_{cryst} near zero. As the pI_{est} increased, the Q_{cryst} became more positive, while the Q_{cryst} became more negative if the pI_{est} decreased below 7.0. A high correlation was also observed between pH_{cryst} and Q_{cryst} (Figure 3.3e). Proteins with a high pH_{cryst} are more negatively charged, while proteins with a lower pH_{cryst} generally had a positive Q_{cryst}. The interesting patterns of MW_{au} and pI_{est} with Q_{cryst} indicated that these *Features* might be able to

predict more probable Q_{cryst} values that result in the formation of crystals. Therefore, various groups of proteins were formed based upon their MW and pI_{est} values and differences were then examined in their Q_{cryst} distributions.

3.2.1.1 Binning by MW

Proteins were binned into several groups based upon their MW_{au}, pI_{est}, or estimated titration curve. The first method examined simply binning the nrPDB_v107 by their ln(MW) distribution into three groups, 'Small,' 'Average,' and 'Large.' The ln(MW) distribution was chosen over the MW distribution, because the ln(MW) distribution was relatively normal compared to the highly skewed MW_{au} distribution. A protein was considered 'Average' if its ln(MW) value was within one standard deviation (SD) of the mean value based on all 3,957 proteins. 'Small' proteins were those with a ln(MW) value greater than 1 SD below the mean, while 'Large' proteins were those whose ln(MW) value was greater than 1 SD above the mean.

When the nrPDB_v107 proteins were separated into three groups by their size, $ln(MW_{au})$, no significant differences were observed in their pI_{est} and random number distributions (Table 3.3a). The random number distribution, which served as an internal control, had no differences as expected. Although no differences were found with the pI_{est} distributions, significant differences were found between all slopes (scaled or not) of the estimated titration curves at both the pI_{est} and pH_{cryst}. Although pI_{est} is not affected by the size of the protein, the shape of the titration curve is. As the proteins grew larger, the slope at the titration curve at both the pI_{est} and pH_{cryst} became increasingly negative. This also corresponds to different shaped scaled titration curves. Because a larger protein can have more charged residues, they can take on a greater range of Q values and hence Q_{cryst} values, which were also significantly different between MW groups. Differences were also found in the pH_{cryst} and $diff_{lim}$ distributions. Smaller proteins were found to crystallize under significantly lower experimental pH conditions. As the proteins grow larger, the $diff_{lim}$ increased. This last finding is well described in the literature and also served as an internal control. Not finding this correlation would lead to more questions about the analysis.

The next method split the 'Small' and 'Large' proteins into an additional group each, creating five ln(MW) groups. If the protein was greater than 2 SD below or above the mean ln(MW) value, the protein was labeled as 'Very Small' or 'Very Large,' respectively. The five MW group results were very similar to that of the three MW groups (Table 3.3b). Again, no differences were found in the distributions of the pI_{est} or the random number. The slopes at both the pI_{est} and pH_{cryst} were significantly different for each group, as the values become increasingly negative as the size increases. This same pattern was also observed for the Q_{cryst} distributions. Additionally, sporadic differences were observed in the slopes of the scaled titration curves at both the pI_{est} and pH_{cryst}. This led us to question whether the scaled titration curves were indeed removing the effects of molecular weight. Similar effects of size on the $diff_{lim}$ distributions were also observed with the 5 ln(MW) groups, with each group having a significantly different $diff_{lim}$ distribution with the smaller proteins having a better $diff_{lim}$, i.e. lower values. Now, it became apparent that the 'Very Small' proteins were the ones that had a low pH_{cryst}, 5.9±1.2. All other groups had a mean pH_{cryst} between 6.5-6.8. There were also some slight differences between the other MW groups and the pH_{cryst}, with the 'Small' proteins reportedly crystallized at slightly more acidic conditions than the 'Average' or 'Large' groups.

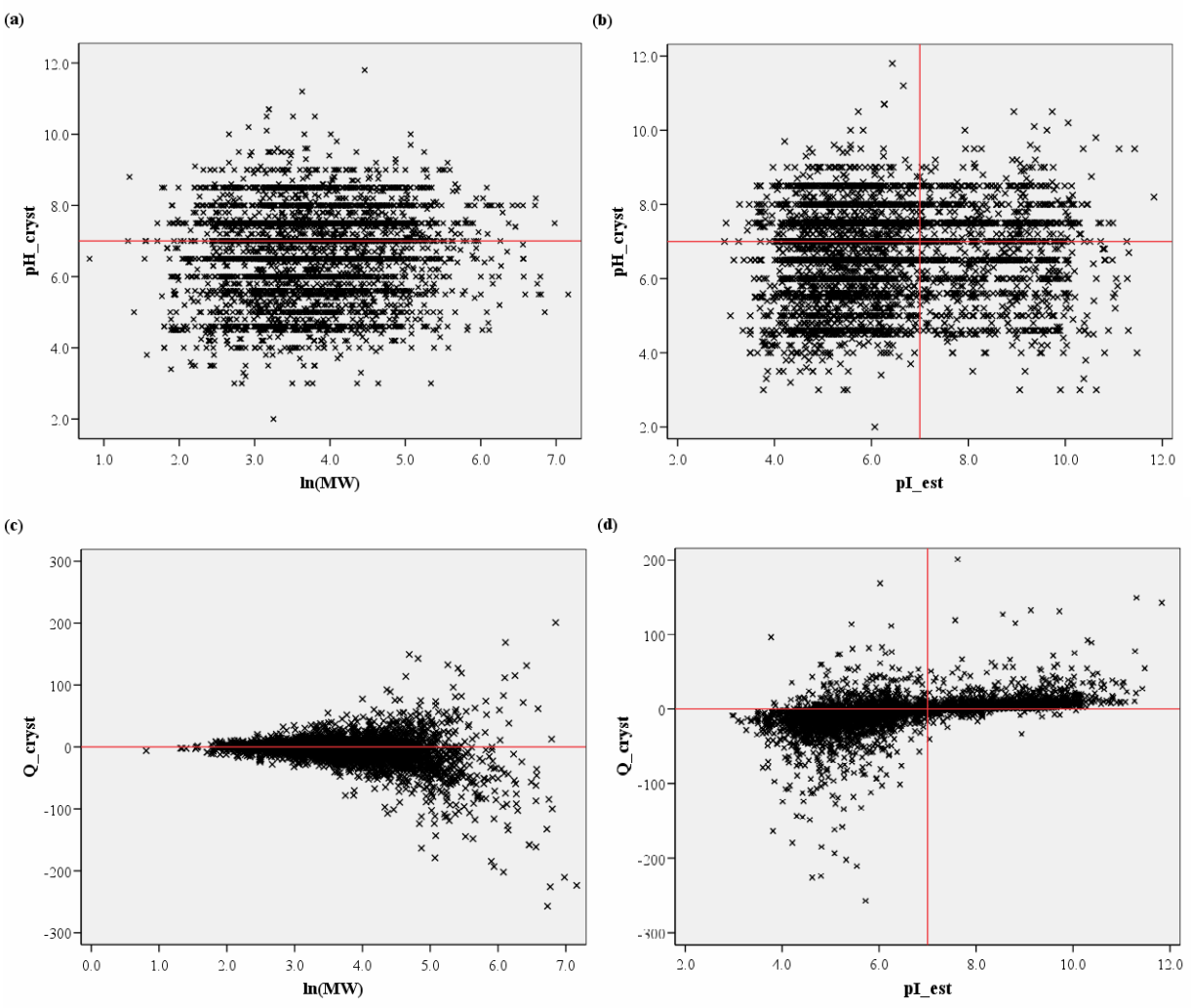


Figure 3.3: Scatterplots of the nrPDB_v107 data showing the distributions of the MW, pI_{est}, pH_{cryst}, and Q_{cryst}.

Table 3.3 The mean and standard deviation (SD) of the examined variables for each of the nrPDB_v107 groups separated by their MW_{au} into either (a) three or (b) five groups.

(a) 3 MW groups

MW_{au} Group	n		pI_{est}	Slope at pI_{est}	SS at pI_{est}	pH_{cryst}	Slope at pH_{cryst}	SS at pH_{cryst} [b]	Q_{cryst}	$diff_{lim}$ [a]	Random Number
Small	656	Mean	6.4 ^{A}	-4.5 ^{A}	-0.12 ^{A}	6.4 ^{A}	-2.5 ^{A}	-0.06 ^{A}	-0.5 ^{A}	1.8 ^{A}	0.0 ^{A}
		SD	2.0	3.3	0.08	1.4	2.5	0.06	6.8	0.4	1.0
Average	2,670	Mean	6.3 ^{A}	-11.6 ^{B}	-0.09 ^{B}	6.7 ^{B}	-7.1 ^{B}	-0.05 ^{B}	-3.3 ^{B}	2.1 ^{B}	0.0 ^{A}
		SD	1.7	9.6	0.06	1.3	7.5	0.05	15.8	0.5	1.0
Large	631	Mean	6.1 ^{A}	-40.6 ^{C}	-0.07 ^{C}	6.8 ^{C}	-25.6 ^{C}	-0.05 ^{C}	-17.9 ^{C}	2.4 ^{C}	0.0 ^{A}
		SD	1.3	41.5	0.05	1.2	30.0	0.04	53.1	0.6	1.0
KW		x^2	1.2	1354.2	118.7	28.6	1204.4	21.7	173.8	344.0	0.2
Test		$p<$	0.539	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.918

b) 5 MW groups

MW Group	n		MW_{au}	pI_{est}	Slope at pI_{est}	SS at pI_{est}	pH_{cryst}	Slope at pH_{cryst}	SS at pH_{cryst}	Q_{cryst}	$diff_{lim}$	Random
Very Small	78	Mean	6.5 ^{A}	6.3 ^{A}	-2.4 ^{A}	-0.10 ^{A}	5.9 ^{A}	-1.4 ^{A}	-0.07 ^{AB}	-0.4 ^{A}	1.6 ^{A}	0.0 ^{A}
		SD	1.0	1.9	2.1	0.09	1.2	1.5	0.06	4.2	0.5	1.1
Small	575	Mean	13.2 ^{B}	6.4 ^{A}	-4.8 ^{B}	-0.12 ^{B}	6.5 ^{B}	-2.6 ^{B}	-0.06 ^{A}	-0.5 ^{B}	1.9 ^{B}	0.0 ^{A}
		SD	2.6	2.0	3.3	0.07	1.4	2.6	0.06	7.1	0.4	1.0
Avg.	2,674	Mean	47.2 ^{C}	6.3 ^{A}	-11.6 ^{C}	-0.09 ^{C}	6.7 ^{C}	-7.1 ^{C}	-0.05 ^{B}	-3.3 ^{C}	2.1 ^{C}	0.0 ^{A}
		SD	22.5	1.7	9.6	0.06	1.3	7.5	0.05	15.8	0.5	1.0
Large	529	Mean	148.2 ^{D}	6.2 ^{A}	-31.2 ^{D}	-0.07 ^{D}	6.8 ^{C}	-19.8 ^{D}	-0.05 ^{C}	-13.2 ^{D}	2.4 ^{D}	0.0 ^{A}
		SD	36.9	1.3	23.4	0.05	1.2	19.6	0.04	39.3	0.5	1.0
Very Large	101	Mean	445.3 ^{E}	6.0 ^{A}	-90.0 ^{E}	-0.07 ^{D}	6.8 ^{BC}	-55.8 ^{E}	-0.05 ^{BC}	-43.1 ^{E}	2.6 ^{E}	0.0 ^{A}
		SD	204.6	1.2	70.7	0.04	1.2	50.6	0.04	93.9	0.7	1.1
KW		x^2	2713.5	2.10	1393.9	126.1	45.2	1243.2	21.5	192.2	356.7	0.29
Test		$p<$	0.0001	0.717	0.0001	0.0001	0.0001	0.0001	0.0003	0.0001	0.0001	0.990

Note: Groups labeled with different letters (A, B, C, D, or E) have significantly different distributions ($p<0.01$) as determined by a KS Test.

3.2.1.2 Binning by pI_{est}

A similar method of discretization was used for the protein's pI_{est}. Initially, the general description of 'Acidic,' 'Neutral,' and 'Basic' proteins was taken from Ries-Kautt and Ducruix (1999). Proteins were considered 'Neutral' if their pI_{est} was between 6.0 and 8.0. Those proteins with a $pI_{est} \leq 6.0$ were labeled 'Acidic,' while those with a $pI_{est} \geq 8.0$ were labeled 'Basic.' As seen in Table 3.4, the majority of proteins were considered 'Acidic' (53.6%), while another sizable portion of proteins were 'Neutral' (26.7%). Only a small portion of the proteins were labeled as 'Basic' (19.7%). This was also observed in the Baseline pI_{est} distribution (Figure 3.2a).

Table 3.4 Discretization of the nrPDB_v107 proteins by their pI_{est}.

pI_{est} Group	n		MW	Slope at pI_{est}	SS at pI_{est}	pH_{cryst}	Slope at pH_{cryst}	SS at pH_{cryst}	Q_{cryst}	$diff_{lim}$	Random Number
Acidic	2,122	Mean	70.0 ^{A}	-20.3 ^{A}	-0.12 ^{A}	6.6 ^{A}	-10.3 ^{A}	-0.06 ^{A}	-14.6 ^{A}	2.1 ^{A}	0.0 ^{A}
		SD	91.2	26.5	0.06	1.3	17.7	0.05	26.2	0.5	1.0
Neutral	1,057	Mean	68.8 ^{A}	-8.4 ^{B}	-0.04 ^{B}	6.7 ^{A}	-9.7 ^{A}	-0.05 ^{B}	0.8 ^{B}	2.1 ^{A}	0.0 ^{A}
		SD	80.5	10.4	0.03	1.2	13.4	0.04	16.8	0.5	1.0
Basic	778	Mean	46.7 ^{B}	-9.8 ^{C}	-0.09 ^{C}	6.7 ^{A}	-5.8 ^{B}	-0.05 ^{C}	12.4 ^{C}	2.0 ^{A}	0.0 ^{A}
		SD	55.4	12.3	0.06	1.3	9.2	0.04	21.9	0.5	1.1
KW Test		x^2	103.2	660.8	1280.4	5.3	122.0	29.0	1629.3	7.6	1.7
		$p<$	0.0001	0.0001	0.0001	0.072	0.0001	0.0001	0.0001	0.022	0.424

Note: Groups labeled with different letters (A, B, or C) have significantly different distributions ($p<0.01$) as determined by a KW Test and followed by a KS Test if significantly differences were found.

No statistically significant differences were observed in the pH_{cryst}, $diff_{lim}$, or random number distributions. The 'Basic' proteins were also smaller, averaging only 47 kDa, while the 'Acidic' and 'Neutral' proteins were considerably larger, ~70 kDa. Differences were also observed in the slopes of the titration curves at both the pI_{est} and pH_{cryst}. The 'Acidic' proteins had a much steeper slope at the pI_{est} than did the other two groups. This might be expected as the pK_a values of Aspartic Acid and Glutamic Acid are 3.65 and 4.25 respectively. This causes the titration curves to decrease rapidly around these pH values. Again, the differences in the titration curve carried over to the scaled titration curves, which were significantly different between all

pI_{est} groups. Finally, while no differences were observed in pH_{cryst} distributions, significant differences in the Q_{cryst} distributions were observed between each pI_{est} group. 'Acidic' proteins were crystallized with a global net negative charge, while 'Basic' proteins were generally crystallized with a net positive charge. Neutral proteins were crystallized with a net charge near zero, the isoelectric point. This was not surprising given the different curves for these pI_{est} groups. A majority of the estimated titration curve was negative for 'Acidic' proteins, while 'Basic' proteins had a large range of positively charged values. This led us to investigate separating the proteins by their titration curve, rather than one point along the curve, the pI_{est}.

3.2.1.3 Binning by pI x MW

After observing differences in variables based upon separation of proteins by their MW or pI_{est}, proteins were separated into groups based upon the combination of their size and pI_{est}. For this analysis the three ln(MW) Groups (Section 3.2.1.1) and three pI_{est} Groups (Section 3.2.1.2) were used. The cross product, ln(MW) x pI_{est}, resulted in the formation of 9 groups (Table 3.5). The number of proteins in each group was quite variable, with the 'Average Acidic' group containing 36% of the proteins and the 'Large Basic' group containing only 1.8%.

For this analysis, only the pH_{cryst} and Q_{cryst} results are presented (Table 3.6). Similar to the previous analysis, sporadic differences are observed in the pH_{cryst} distributions among groups. All groups based on a similar size, such as 'Small,' have similar pH_{cryst} distributions regardless of the pI_{est} group. However, when the Q_{cryst} distributions were examined, significant differences were found among all groups. This led us to believe that the Q_{cryst} is an important crystallization variable that could be used for modeling solution pH conditions.

Table 3.5 Discretization of $\ln(MW_{au})$ and pI_{est} variables within the nrPDB_v107 data set. (a) The number and percentage of proteins in each group when each size group was crossed with each pI_{est} group.

pI_{est} Bin		$\ln(MW_{au})$ Bins			**Total**
		Small	**Average**	**Large**	
Acidic	Count	332	1424	366	2122
	% of Total	8.4%	36.0%	9.2%	53.6%
Neutral	Count	152	713	192	1057
	% of Total	3.8%	18.0%	4.9%	26.7%
Basic	Count	172	533	73	778
	% of Total	4.3%	13.5%	1.8%	19.7%
Total	Count	656	2670	631	3957
	% of Total	16.6%	67.5%	15.9%	100.0%

Table 3.6 Discretization of $\ln(MW_{au})$ and pI_{est} variables within the nrPDB_v107. After detecting significant differences (p-value<0.01) with a KW test, a KS Test was performed pairwise for each MW x pI_{est} Bin to detect the individual differences for pH_{cryst} or Q_{cryst} (shaded).

	Small Acidic	Average Acidic	Large Acidic	Small Neutral	Average Neutral	Large Neutral	Small Basic	Average Basic	Large Basic
Small Acidic		**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**
Average Acidic	**0.001**		**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**
Large Acidic	**0.005**	0.719		**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**
Small Neutral	0.258	**0.005**	**0.003**		**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**
Average Neutral	**0.0001**	0.322	0.985	**0.0002**		**0.0001**	**0.0001**	**0.0001**	**0.0001**
Large Neutral	**0.0002**	**0.0003**	0.030	**0.0001**	**0.004**		**0.0001**	**0.0001**	**0.0001**
Small Basic	0.655	0.784	0.467	0.202	0.300	0.017		**0.0001**	**0.0001**
Average Basic	**0.003**	0.866	0.990	**0.003**	0.935	**0.006**	0.693		**0.0001**
Large Basic	0.034	0.115	0.507	**0.006**	0.461	0.896	0.170	0.385	

Note: The values presented are the p-values resulting from the KS Test.

3.2.2 Separating Proteins by their Estimated Titration Curve

After determining that the slopes of the estimated titration curves were different for proteins separated by their MW or pI_{est}, separating proteins by their titration curves was investigated. For this method, Self-Organizing Maps (SOMs) were used (Kohonen, 2001). For a more detailed explanation on SOMs, see Section 4.7.2.2. The input into the SOM algorithm was a one-dimensional vector for each protein representing the Q values over a given range of pH values. This vector was composed of the Q values every 0.2 pH units from 1.0-15.0. Although a pH value of 15.0 is not possible, it was included so that the estimated titration curves were relatively flat at the end of the curve. This was primarily due to the pK_a value of Arginine, 12.48.

Initial analysis on the estimated titration curve using SOMs basically separated the proteins by their MW. The proteins were separated primarily by the initial or final values of the titration curve. This led us to attempt to scale the titration curves, so each curve started at a Q of 1.0 and ended at a Q of 0.0.

Initially, the GSOM software was applied to the entire set of PDB proteins (PDB_v107) to cluster the estimated titration curves. The algorithm was able to place all the proteins into a group of 61 clusters (Figure 3.4). Proteins with an 'Acidic' pI_{est} ($pI_{est} \leq 6$) were clustered near the top of the SOM, while proteins with a 'Basic' pI_{est} ($pI_{est} \geq 8$) were clustered towards the bottom of the SOM. Examples of some mean estimated curves from the clusters in Figure 3.4 are shown in Figure 3.5a. The pH_{cryst} distributions of the 'Acidic' vs. 'Basic' proteins are shown in Figure 3.5b. When the 'Acidic' and 'Basic' proteins were examined from Figure 3.5b, significant differences in pH_{cryst} distributions were discovered. This observation could lead to an experimental protocol such as that shown in Figure 3.6.

Figure 3.4: The resulting GSOM derived from the data distributed into 61 clusters.

From the input of a target protein's sequence, the estimated titration curve and the scaled titration curve are calculated. The scaled titration curve would then be presented to the GSOM algorithm based on all proteins listed in the PDB_v107 data set. The SOM algorithm then places the test protein into the cluster with the best matching curves. This cluster or neighborhood of clusters could then be examined for the frequency of pH_{cryst} values and used to intelligently guide the researcher for selecting pH ranges for their test protein.

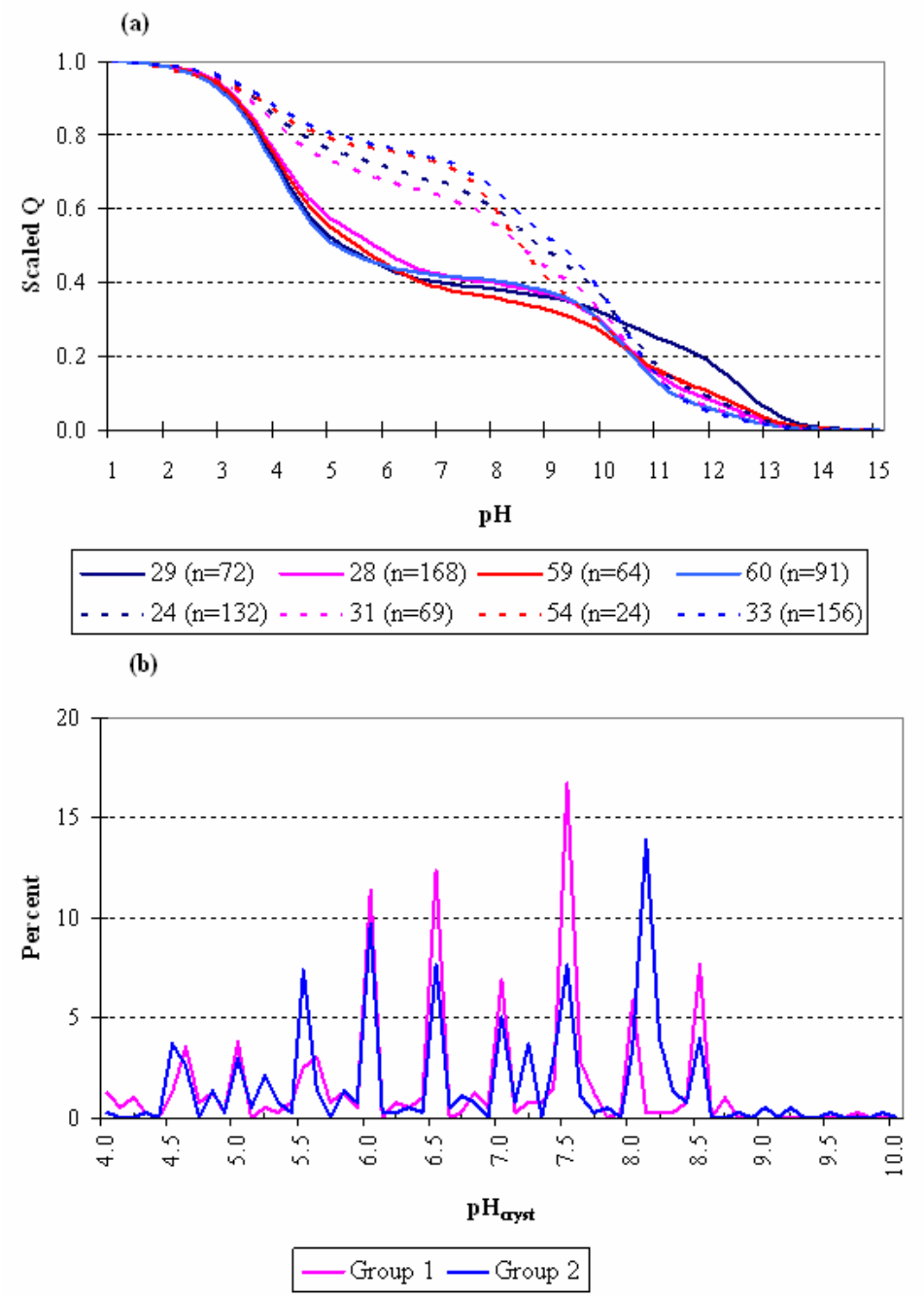


Figure 3.5: (a) The GSOM Clusters closer together in space (Figure 3.4) have more similar estimated titration curves. Clusters 28, 29, 59, and 60 are used for Group 1, while Clusters 24, 31, 33, and 54 are used for Group 2. (b) the pH_{cryst} distributions of Group 1 and Group 2.

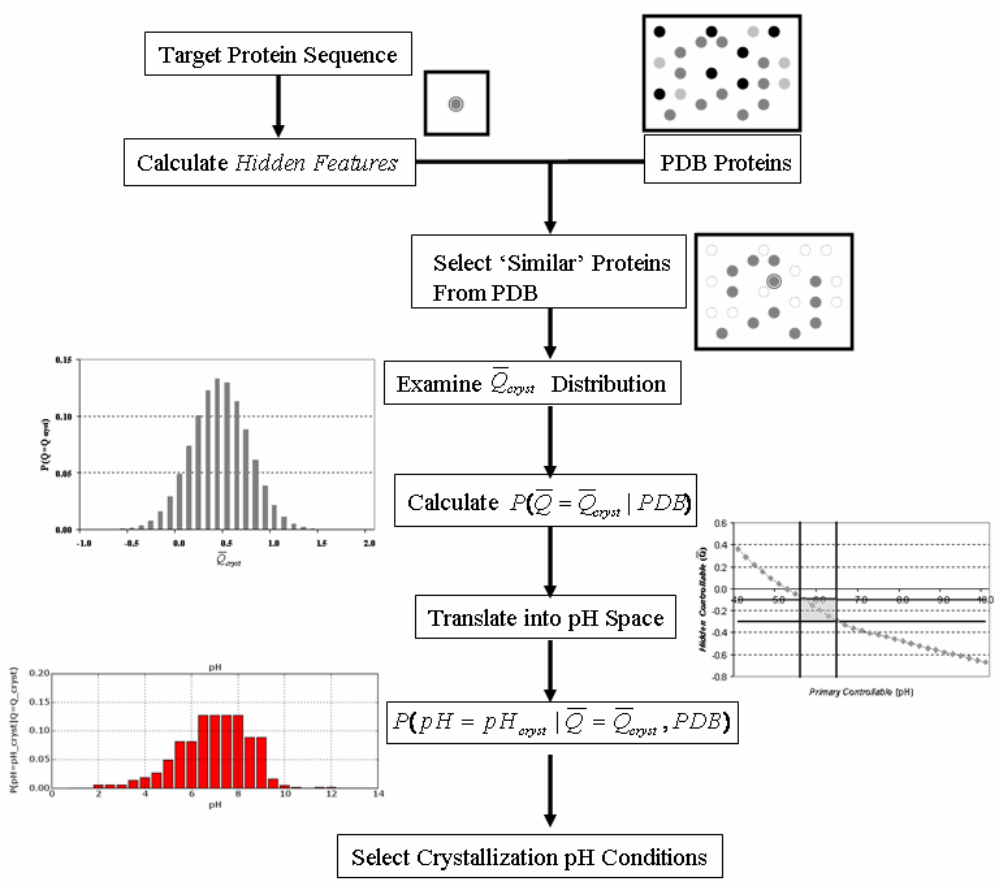


Figure 3.6 Expected flow of selecting solution pH conditions for a target protein.

So far, differences were found in experimental conditions for protein structures based upon their size (MW) and pI_{est}. Another method to group proteins (SOMs) based on the estimated titration curve was also explored. While exploratory, these novel findings warranted further analysis. These findings have been examined in subsequent chapters by attempting to account for the size of the protein by dividing the estimated net charge by either the molecular weight in kilodaltons or estimated solvent accessible surface area in square nanometers to obtain an estimated specific charge curve or estimated surface charge density curve, respectively. The translation from theory to the workbench will also be discussed.

3.3 CHAPTER SUMMARY

A retrospective study was undertaken as an attempt to explain the behavior of proteins undergoing crystallization attempts. The initial attempt focused on the pH_{cryst}, because it is the most widely reported experimental crystallization variable available in the Protein Data Bank (PDB). The pH_{cryst} values were obtained for proteins by extracting the information from each protein structure's mmCIF file. Two data sets with different levels of redundancy were created from the available data in version #107 of the PDB. The first data set consisted of all protein structures that had a pH_{cryst} with some added constraints, such as the length being greater than or equal to 20 amino acids and removing all membrane proteins. The PDB is known to contain a large amount of data on several well-studied proteins such as lysozyme. Therefore, the second data set removed all proteins with a sequence similarity greater than a BLAST p-value of 10^{-80}. Next, the distribution of protein features and crystallization variables was examined between each data set.

Statistically significant differences were observed for various *Given* protein *Features* and crystallization variables (*Observables*) between the two data sets. Several of these could be explained by the abundance of redundant structures within the database. Therefore, it was felt that the non-redundant data set would remove much of the selection bias.

After choosing the non-redundant data set, the correlations between both the protein *Features* (MW and pI_{est}) and the pH_{cryst} were examined. Similar to previous studies the pH_{cryst} was not correlated to the pI_{est} of the proteins ($r = 0.061$). Because the solution pH helps control the charge of the proteins, the estimated net charge (Q) of the protein at the pH_{cryst} (Q_{cryst}) was calculated and examined. The thought was that Q_{cryst} might be used as a proxy variable for the pH_{cryst}. However, the Q_{cryst} is obtained by calculating the estimated titration curve of the protein structure, which uses the amino acid sequence, the assumed pK_a values of the charged residues, and the pH_{cryst}. . Because the Q_{cryst} is calculated using these variables, it might be expected that there would also be a high correlation with these variables. The Q_{cryst} was found to have a positive linear relationship with the pI_{est} ($r = 0.661$; $p < 0.0001$; Figure 3.3d), which is also

obtained from the estimated titration curve. However, Q_{cryst} was found to have a negative linear correlation with pH_{cryst} (r = -0.475; p <0.0001) and a negative non-linear correlation with ln(MW) (r = -0.252; p <0.0001). From the estimated titration curve, various other shape features were obtained, such as the slope of the titration curve at both the pI_{est} and pH_{cryst}. Because the possible range of Q values are dependent upon the amino acid composition and the size of the protein, an attempt was made to account for these differences by scaling each curve so that the maximum Q value was 1.0 and the minimum Q value was 0.0. It was believed that this scaling would account for size differences. However, the results here were inconclusive, as clear MW effects were still observed with the slopes of the scaled curves.

After the intercorrelations of variables were examined, proteins were separated into subgroups by their MW_{au}, pI_{est}, or a combination of their MW_{au} x pI_{est}. When proteins were separated into three or five groups by their ln(MW) values, the proteins labeled 'Small' had a slight shift in their pH_{cryst} distributions to lower pH values (statistically significant). When the MW groups were further separated into a 'Very Small' group, this group was found to have a much lower shift in pH_{cryst} distributions than the 'Small' proteins. While only slight statistical differences in the pH_{cryst} distributions were seen in the ln(MW) groups, the Q_{cryst} distributions of each group had statistically significant shifts in their distributions. As proteins grew larger, the Q_{cryst} distribution became much more negative. Additionally, each group displayed a significant difference in their $diff_{lim}$ distributions and the slopes of the titration curve at both the pI_{est} and pH_{cryst}. However, significant differences were still observed with the slopes of the scaled titration curves. This indicated that scaling the titration curves did not remove all of the influence of molecular weight and hinted at a possible non-linear relationship.

When proteins were separated into groups based upon their pI_{est}, no differences were observed in the pH_{cryst} distributions between groups. However, when the Q_{cryst} distributions were examined, large differences were observed among all groups. This was also observed in the MW_{au} x pI_{est} groups. Groups with the same size did not display significantly different pH_{cryst} distributions, but every group had significantly different Q_{cryst} distributions. These findings confirmed previous analyses of failing to correlate a protein's pI_{est} with the pH_{cryst}, but

it did offer a possible explanation to why this link could not be found. Although proteins may crystallize under similar pH conditions, the underlying net charge of the proteins may be quite different. This is seen in the different shapes of estimated titration curves. While the pH_{cryst} cannot be predicted apriori, these results suggest that there are estimated net charge (Q) values that have a higher probability of resulting in the formation of crystals.

After determining that protein crystallization behavior might be influenced by the magnitude of the estimated titration curve, proteins were clustered by their scaled estimated titration curve. An unsupervised clustering algorithm, SOMs, was used to group proteins into 61 clusters by their estimated titration curve. 'Acidic' and 'Basic' proteins were at opposite ends of the SOM and had significantly different pH_{cryst} distributions. This indicated that rather than one point along the estimated titration curve, the pI_{est}, the titration curve itself might be used for predicting solution pH ranges that are more probable or less probable in growing a protein crystal.

In conclusion, the size of the protein clearly has a significant impact on the magnitude of the Q_{cryst}. When proteins were separated into groups on the basis of their MW_{au}, pI_{est}, or the estimated titration curve, significantly different Q_{cryst} distributions were observed. Based upon these observed differences, it was hypothesized that the Q_{cryst} distribution of previously crystallized proteins might be used to select pH ranges that have a higher probability of generating crystals suitable for diffraction studies.

4.0 PROTEIN SEQUENCE-PROPERTIES EVALUATION FRAMEWORK

When initial attempts are made at crystallizing a new target protein, researchers typically use commercially developed screens with little consideration towards the protein and its biophysical properties. While many researchers believe that there are clues present within a protein's sequence that can serve as a guide to selecting appropriate solution conditions for crystallization, there are few examples within the literature on how to test and use this knowledge. In this chapter, the development of the Protein Sequence-Properties Evaluation (PSPE) Framework was largely motivated by the rapid growth and evolution of the Protein Data Bank. The PSPE framework can be used for examining any crystallization variable of interest, such as the solution pH, and correlating the variable of interest with biophysical properties of successfully crystallized proteins. But rather than just a computational approach, searching for correlations without the generation of hypotheses, this framework merges the computational and experimental approaches, where a hypothesis is generated and then tested. This is done by conducting a retrospective evaluation of the experimental data, in this case, the successful solution conditions used to grow protein crystals, which were available from the Protein Data Bank (PDB). It is hypothesized that the Protein Sequence-Properties Evaluation Framework can be used to frame and test hypotheses in-silico about variables that are believed to be important in crystallization. If successful, these findings should increase our understanding of the crystallization process and provide a standard for developing more successful crystallization screens for other variables in addition to the solution pH.

4.1 PSPE FRAMEWORK

The Protein Sequence-Properties Evaluation (PSPE) framework developed in this dissertation is presented in Figure 4.1, as both (a) a general framework and (b) the specific framework used in this dissertation. While the PSPE framework was specifically applied to the area of protein crystallization, this framework could also be applied to other biological problems, where differences in protein behavior are observed, but not explained. An example of another future application in another unrelated field would be the prediction of antigenic activity of short peptide sequences. This would be useful for immunologist predicting the immune response to pathological organisms and their proteins. The identification of proteins that exhibit an immune response may lead to the development of treatments and vaccinations against the organisms, an important area in human health.

In this dissertation, the PSPE framework was used in an attempt to limit the solution pH ranges examined in initial crystallization screens. However, this framework could be used for determining differences in any of the crystallization experimental variables. Similar to the scientific method, the first step in the framework involves developing a testable hypothesis (Section 4.2). Rather than performing actual experiments and collecting data, a list of proteins that match any constraints are obtained from the Protein Data Bank (PDB; Section 4.3). The available data that can be obtained apriori (*Givens*) include the *Primary Feature* (amino acid sequence) and *Primary Observable* (reported pH of crystallization; pH_{cryst}; Section 4.4). After obtaining the reduced list of protein structures, various *Hidden Features* are calculated (Section 4.5). From the *Primary Feature* and other relevant information available within the database, a reduced list of non-redundant structures is created (Section 4.6). The distribution of these *Features* for subpopulations of proteins may be modeled in their ability to predict ranges of the *Observable* values (pH_{cryst}, Q_{cryst}, $\overline{Q}_{cryst}$, and σ_{cryst}; Sections 4.7 and 4.8). The model's abilities to predict a range for a *Controllable* by setting the values equal to the *Observables* is then tested on an independent test set (Section 4.9). These models may be applied in the future for determining more probable solution conditions that will result in crystallization for new target proteins.

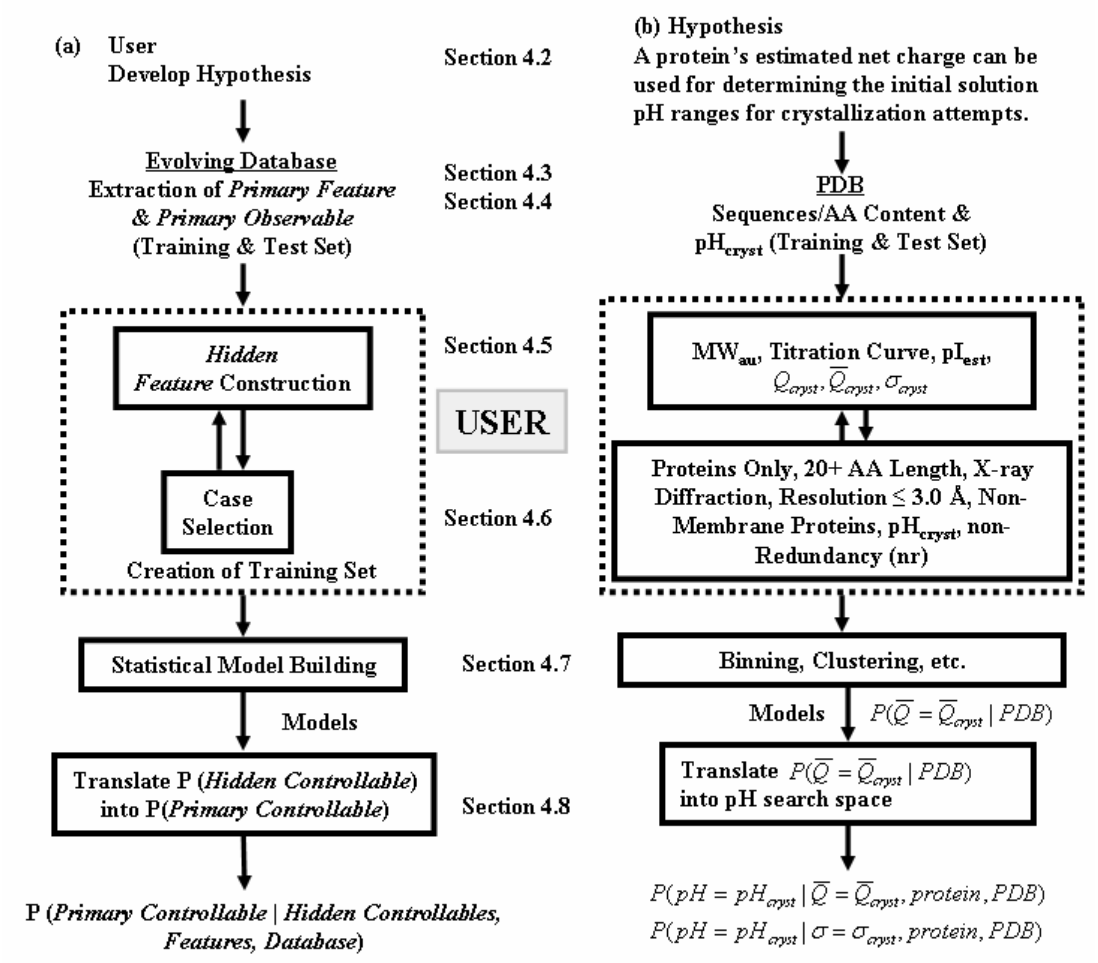


Figure 4.1 The (a) general and (b) specific applications of the Protein Sequence-Properties Evaluation (PSPE) framework as discussed in this dissertation.

4.2 DEVELOPMENT OF A HYPOTHESIS

Similar to the scientific method, the first step in the framework that is proposed is to identify a problem of interest and develop a testable hypothesis. For this project, hypotheses were framed based upon physical reasons for correlations between a protein's biophysical properties and crystallization *Observables*. The next step would be to perform experiments and collect the data. If the hypothesis is true, the correlations should be seen in the data. In the example application, a

retrospective study was conducted with no actual experiments being performed. The initial hypothesis developed in Chapter 3 was that the pH_{cryst} could be predicted using a protein's pI_{est}.

As mentioned earlier, this seemed logical and with more 3D protein structures available, this hypothesis was tested. Although the results showed no useful relationship between the two variables, another relationship between a *Hidden Observable*, the Q_{cryst}, and the pI_{est} proved promising. The hypothesis was that the Q_{cryst} could be used as a proxy variable for predicting the most probable pH_{cryst} ranges. This explanation also seemed plausible based on physical chemistry, because the solution pH largely exhibits its influence on the protein's ionizable residues. The Q_{cryst} values are not available in any database and had to be estimated using sequence and pH_{cryst} information. However, it should be mentioned that a negative result is still useful; *i.e.* the recognition that a variable has minimal utility in predicting ranges or probabilities of crystallization allows energy and attention to be focused elsewhere, where it may be more productively employed.

4.3 FEATURE EXTRACTION & EVOLVING DATABASE

The next step involved in the framework was to obtain a set of protein structures. Because the application involved using protein structures that had been successfully crystallized to select conditions for future proteins undergoing crystallization attempts, the PDB (Berman et al., 2000; Section 4.3.1) was used as the source of proteins. The PDB (http://www.rcsb.org/pdb/) is the repository for 3D biological macromolecular structures (proteins and nucleic acids). Because all the information required could not be found in the PDB, several related databases (nrPDB, Sections 4.3.2; and PDB-REPRDB, Section 4.3.3) were also used to augment the information for this dissertation. All these databases are continually evolving as new three-dimensional (3D) structures are deposited into the PDB. These new instances provide a rich source of test cases for models developed prior to their deposition (Section 4.9).

4.3.1 Protein Data Bank (PDB)

The Protein Data Bank (PDB) collects and organizes all information regarding the 3D structure of biological macromolecules. The PDB is updated on a weekly basis. During each update, structures may be added or removed. When structures are removed, they are generally replaced by a better example of the structure, i.e. better resolution. The list of structures to use as the training set was downloaded from the PDB on October 5th, 2004. A newer set of structures was obtained and downloaded a year later in November, 2005 to use as a test set (see Section 4.9 for more details).

The PDB contains many instances of redundant structures, where two or more sequences share a high level of similarity. This high level of redundancy has the potential to bias any analysis of the PDB. The most common example of redundancy is the case of mutations, where two sequences differ by a single amino acid. For example, a PDB search for 'T4 lysozyme' using the method of X-ray diffraction returned 408 structures (02/01/2005). The exact same search retrieved 419 structures nine months later (11/22/2005). Other examples of redundancy include the same protein unbound or bound to another protein or ligand, which creates a new PDB entry. For example, sperm whale myoglobin is present in the PDB in both the oxygenated (PDB ID: 1A6M) and deoxygenated (PDB ID: 1A6N) form. The only difference between these structures is a bound oxygen atom. These redundant structures may bias the analysis if they are present in a large amount; therefore, a non-redundant data set of structures was created.

4.3.2 Non-Redundant PDB (nrPDB)

In order to deal with the issue of redundancy in the Protein Data Bank (PDB), the National Center for Biotechnology Information (NCBI) created the non-redundant Protein Data Bank (nrPDB; http://www.ncbi.nlm.nih.gov/Structure/VAST/nrpdb.html). From this list, subgroups of PDB structures with varying degrees of non-redundancy could be selected (Table 4.1). The nrPDB, which is updated approximately once a month, operates at the chain level when considering redundancy. However, this dissertation focused at the structural level. Because each

structure may be composed of multiple chains, there needed to be a conversion to the structural level. For example, a structure can have different chains in different non-redundant groups.

Using the methods of Holm and Sander (1998), the nrPDB compares all chains within the PDB pairwise using the BLAST algorithm (Altschul et al., 1997). Using one of the four levels of redundancy, the chains are clustered by their sequence similarity and then ranked (Table 4.1). The rankings are based on structural quality using factors such as resolution, percentage of residues of unknown amino acid type, incomplete coordinate data, missing data, and incomplete side-chain coordinates. The nrPDB automatically removes chains less than 20 amino acid residues. In some cases, if a mutant structure has a higher ranking than does the native structure, the native structure may be manually placed higher if it is of comparable quality.

Table 4.1 Four levels of varying redundancy based on sequence similarity are available from the nrPDB.

BLAST p-value	Stringency
10e-7	Most
10e-40	↓
10e-80**	
100% sequence identity	Least

** Level of redundancy used in this dissertation

4.3.3 PDB-REPRDB

The Representative Protein chains from the PDB database (PDB-REPRDB; http://www.cbrc.jp/pdbreprdb-cgi/reprdb_menu.pl) was created to provide a list of representative protein chains from the PDB (Noguchi and Akiyama, 2003). The PDB-REPRDB is updated every 2-4 months. The 'best' representative structure are prioritized by the user based on 9 features, most of which are the same as the nrPDB, such as resolution, R-factor, number of chain breaks, and missing data. One of the key features used in this database was the ability to easily identify membrane proteins for removal from the training and test sets. The user can also filter out mutants, complexes, and fragments, which cannot be easily accomplished by the PDB (Section 4.3.1) or nrPDB (Section 4.3.2).

4.4 EXTRACTION OF PRIMARY FEATURES

The *Primary Feature* or protein *Given* (amino acid sequence and composition) was obtained from the PDB. The PDB provides a list of all sequences found in the structure's asymmetric unit at ftp://ftp.rcsb.org/pub/pdb/derived_data/pdb_seqres.txt. This list is in FASTA format (Figure 4.2) for all structures present within the PDB, regardless of method (X-ray, NMR, etc.) and type (protein or nucleic acid). Each sequence within the file has the PDB ID, chain id, molecular type, amino acid length, description, and the chain sequence reported. This file was parsed into structural units (summation of chains) using Biopython (Chapman and Chang, 2000). A Python (Version 2.3) script (Appendix A) was written to parse this file into individual PDB structures, grouping all chains into one sequence, while documenting all -NH3 and -COOH termini for specific *Hidden Feature* construction (estimated titration curve calculation). *Hidden Features* were constructed if the structures contained only proteins, so all structures with nucleic acids or -HET Groups (either protein-HET or nucleic-HET groups) were removed.

After obtaining the list of protein structures, knowledge of the *Primary Observable* (pH_{cryst}) had to be assessed from the PDB. If a protein did not have a pH_{cryst} listed it had to be removed from the data sets. The structures that had a pH_{cryst} (*Primary Observable*) could be queried at the PDB website using the 'SearchFields' option or examining each protein's macromolecular Crystallographic Information File (mmCIF) (exptl_crystal_grow_ph field). The pH_{cryst} values of structures were queried and downloaded from the PDB website on 11/29/04. Several structures had pH_{cryst} values that were invalid (i.e. greater than 14). These were manually examined in the mmCIF file and the structures were removed if their pH_{cryst} was listed as 0 or greater than 14. On 11/29/2004, there were 16,129 out of 24,118 (67%) PDB structures that had a pH_{cryst} value listed. More recently (04/05/2005), 20,468 / 30,205 (68%) PDB structures reported a pH_{cryst}. In order to calculate the *Hidden Observable* ($\overline{Q}_{cryst}$) only structures that had a valid pH_{cryst} could be used.

```
>1eil_A mol:protein length:292 2,3-Dihydroxybiphenyl 1,2-Dioxygenase
SIERLGYLGFAVKDVPAWDHFLTKSVGLMAAGSAGDAALYRADQRAWRIAVQPGELDDLAYAGLEVDDAAALERM
ADKLRQAGVAFTRGDEALMQQRKVMGLLCLQDPFGLPLEIYYGPAEIFHEPFLPSAPVSGFVTGDQGIGHFVRCV
PDTAKAMAFYTEVLGFVLSDIIDIQMGPETSVPAHFLHCNGRHHTIALAAFPIPKRIHHFMLQANTIDDVGYAFD
RLDAAGRITSLLGRHTNDQTLSFYADTPSPMIEVEFGWGPRTVDSSWTVARHSRTAMWGHKSVRGQR
>1ein_A mol:protein length:269 Lipase
EVSQDLFNQFNLFAQYSAAAYCGKNNDAPAGTNITCTGNACPEVEKADATFLYSFEDSGVGDVTGFLALDNTNKL
IVLSFRGSRSIENWIGNLNFDLKEINDICSGCRGHDGFTSSWRSVADTLRQKVEDAVREHPDYRVVFTGHSLGGA
LATVAGADLRGNGYDIDVFSYGAPRVGNRAFAEFLTVQTGGTLYRITHTNDIVPRLPPREFGYSHSSPEYWIKSG
TLVPVTRNDIVKIEGIDATGGNNQPNIPDIPAHLWYFGLIGTCL
>1ein_B mol:protein length:269 Lipase
EVSQDLFNQFNLFAQYSAAAYCGKNNDAPAGTNITCTGNACPEVEKADATFLYSFEDSGVGDVTGFLALDNTNKL
IVLSFRGSRSIENWIGNLNFDLKEINDICSGCRGHDGFTSSWRSVADTLRQKVEDAVREHPDYRVVFTGHSLGGA
LATVAGADLRGNGYDIDVFSYGAPRVGNRAFAEFLTVQTGGTLYRITHTNDIVPRLPPREFGYSHSSPEYWIKSG
TLVPVTRNDIVKIEGIDATGGNNQPNIPDIPAHLWYFGLIGTCL
>100d_A mol:nucleic length:10 DNA/RNA Chimeric Hybrid Duplex (5'-R(Cp*)-D(C
CCGGCGCCGG
>100d_B mol:nucleic length:10 DNA/RNA Chimeric Hybrid Duplex (5'-R(Cp*)-D(C
CCGGCGCCGG
```

Figure 4.2 An example of sequences in FASTA format in the pdb_seqres.txt file.

4.5 HIDDEN FEATURE AND CONTROLLABLE CONSTRUCTION

The *Primary Feature* (AA_{au}) and *Primary Observable* (pH_{cryst}), both *Givens*, were then used to calculate various *Hidden Features* and *Hidden Observables* that were not present within the original database (PDB). The *Hidden Features* were initially chosen based on their ability to be easily calculated from a protein's primary sequence and their possible significance for crystallization. These *Features* included the asymmetric unit molecular weight (MW_{au}; Section 4.5.1), estimated solvent accessible surface area (A_S; Section 4.5.2), pI_{est}, and the estimated titration curve (Section 4.5.3). Because the solution pH is known to affect the charged state of amino acid residues, various *Hidden Observables* were calculated based on the pH_{cryst}, including the Q_{cryst}, $\overline{Q}_{cryst}$, and σ_{cryst} (Section 4.5.4). The hypothesis was that these variables or combinations of these *Features* would be predictive of the *Primary* or *Hidden Observables* at which crystals were grown (Figure 4.3a and d). Note that one of the hazards of hidden variables

is that they introduce additional information and care must be taken to ensure that "discoveries" based on the hidden variables are not simply reflections of that additional information.

4.5.1 Molecular Weight

The protein's MW_{au} in kilodaltons (kDa) was calculated from the amino acid sequence of all asymmetric unit chains (AA_{au}) listed in the seqres.txt file. The MW values used for the amino acids are shown in Table 4.2. When calculating the MW_{au}, a water molecule (18.02 Daltons) had to be subtracted for every peptide bond. It is important to note that the protein's asymmetric unit may be composed of one or more biological units. The biological unit is the protein's functional biological form, which may consist of a single (monomer) or multiple chains (dimer, trimer, or greater), which may be identical (homo-) or different (hetero-) in amino acid sequence. However, the functional protein may or may not be known *apriori*. For this analysis, it was assumed that the asymmetric unit was known prior to any crystallization attempts.

Table 4.2 The amino acid MW and pK_a values used in this dissertation (Nelson and Cox, 2000).

			pK_a Value		
Abbreviation	Amino Acid	MW (Da)	-NH3 Term.	-COOH Term.	Side Chain
A (Ala)	Alanine	89.09	9.69	2.34	-
C (Cys)	Cysteine	121.15	10.28	1.96	8.18
D (Asp)	Aspartic Acid	133.10	9.60	1.88	3.65
E (Glu)	Glutamic Acid	147.13	9.67	2.19	4.25
F (Phe)	Phenylalanine	165.19	9.00	1.83	-
G (Gly)	Glycine	75.07	9.60	2.34	-
H (His)	Histidine	155.16	9.17	1.82	6.00
I (Ile)	Isoleucine	131.17	9.68	2.36	-
K (Lys)	Lysine	146.19	8.95	2.18	10.53
L (Leu)	Leucine	131.17	9.60	2.36	-
M (Met)	Methionine	149.21	9.21	2.28	-
N (Asn)	Asparagine	132.12	8.80	2.02	-
P (Pro)	Proline	115.13	10.96	1.99	-
Q (Gln)	Glutamine	146.15	9.13	2.17	-
R (Arg)	Arginine	174.20	9.04	2.17	12.48
S (Ser)	Serine	105.09	9.15	2.21	-
T (Thr)	Threonine	119.12	9.62	2.11	-
V (Val)	Valine	117.15	9.62	2.32	-
W (Trp)	Tryptophan	204.23	9.39	2.38	-
Y (Tyr)	Tyrosine	181.19	9.11	2.20	10.07

4.5.2 Estimated Solvent Accessible Surface Area and Surface Charge Density

Previous work has shown that a protein's solvent accessible surface area (A_S) was proportional to its molecular weight raised to the power of a constant, *b,* where *b* ranged between 0.73 and 0.76. This was demonstrated for both monomeric (Chothia, 1975; Janin, 1976; Teller, 1976; Miller et al., 1987a) and oligomeric proteins (Miller et al., 1987b; Janin et al., 1988). For a sphere, the surface area would be proportional to the density to the power of two-thirds (0.67). However, proteins are not spherically shaped, having a convoluted shape with many cavities. Therefore,

the exponent needs to be adjusted for the shape, with the exponent giving a measure of the shape complexity. To estimate the A_S several assumptions were made. First it was assumed that the protein's estimated mean surface charge density (σ_{cryst}) was relatively constant, being proportional to the Q_{cryst}/A_S. The finding by Barlow and Thornton (1986) that the surface charge density of proteins was relatively constant made this assumption reasonable. However, A_S had to first be approximated by the data.

To estimate A_S, it was assumed that the protein's σ_{cryst} was proportional to Q_{cryst}, $\sigma_{cryst} = (Q_{cryst})/(a * MW^b)$. In order to test this hypothesis, the MW was first plotted against the absolute values (abs) of Q_{cryst} for all proteins in the training set. Then, the best empirical least squares fit of $abs(Q_{cryst}) = a * (MW)^b$ was calculated using the power law. If the hypothesis was correct, this empirical fit should be similar to that estimated previously by Chothia, Miller, and Janin. The absolute value of Q_{cryst} was taken due to the large number of negative Q_{cryst} values. Additionally, it should not matter whether the charge is positive or negative, just that σ_{cryst} is proportional to the molecular surface area.

Next, the σ based upon the Q_{cryst} divided by our A_S estimation (σ_{ours}) above was plotted against the σ that was calculated using the A_S equation developed for monomers by Miller et al. (1987a), where $A_S=6.3*MW^{0.73}$ (σ_{miller}). The resulting linear fit of $\sigma_{miller} = m * \sigma_{ours} + c$ gives the scale factor for converting our estimate of A_S to more closely resemble the empirical fit of A_S previously developed by Miller et al. Although, many of the protein structures in the training set are not monomers, Miller et al.'s equation was used as an initial starting point for examining the effects of A_S on protein crystallization. Similar to Q_{cryst}, $\overline{Q}_{cryst}$ and σ_{cryst} were obtained from the $\overline{Q}$ and σ curves respectively, using the $\overline{Q} = \overline{Q}_{cryst}$ or $\sigma = \sigma_{cryst}$ values of the solution where the $pH = pH_{cryst}$.

4.5.3 Estimated Titration Curves and Estimated Isoelectric Point (pI_{est})

Each protein's estimated titration curve was calculated using the HHE (Equation 2.1), the asymmetric unit amino acid sequence (AA_{au}), the assumed pK_a values for the titratable amino acid residues, and the pH range from 1.0 to 14.0 every 0.1 pH unit where *n* was the number of each type of charged residue, either positive (Histidine, Lysine, and Arginine) or negative (Cysteine, Aspartic acid, Glutamic acid, and Tyrosine; Ries-Kautt and Ducruix, 1997; 1999). The point along the curve where the protein has zero net charge is the pI_{est}. This equation accounts for every C-terminus and N-terminus amino acid in the complex or chain. The pK_a values of the amino acids are the model pH values where the residues are 50% ionized. The pK_a values (Nelson and Cox, 2000) used in all calculations are shown in Table 4.2. It was previously demonstrated that pI_{est} values were generally in good agreement with experimentally measured values (Patrickios and Yamasaki, 1995).

The availability of a large amount of experimental data within the PDB warranted a reinvestigation of the link between the pI_{est} and pH_{cryst} (Chapter 5). Although experimental conditions used to grow the crystals are not required for depositing structural information into the PDB, the conditions are beginning to be reported more frequently. The most reported experimental condition used to grow the crystals is the pH_{cryst} (now available for 16,000+ structures). Because the solution pH controls the Q (Figure 4.3e), the Q_{cryst} distribution was also examined to determine if it could be used as a predictor of solution pH ranges leading to successful crystallization. The Q is a function of the amino acid composition, the pH of the solution, and the assumed pK_a values for the titratable amino acid residues in the protein. From examining estimated titration curves (Q plotted as a function of pH), it was observed that the Q could remain relatively constant over a wide range of pH values. Alternatively, the Q could change drastically over a small range of pH values, especially near the pK_a values of the charged residues.

Rather than one point along the estimated titration curve, the pI_{est}, it was hypothesized that the whole titration curve may be more predictive for initial experimental design and optimization (Chapter 6). Additionally, the location on the estimated titration curve at which

crystals appear should dictate a finer (areas with a steep slope) or broader search (areas with a relatively flat slope) of pH ranges.

4.5.4 Hidden Variables

From the protein's amino acid sequence (charged amino acid composition) and the solution pH, the protein's Q is estimated (*Hidden Controllable*). This is done by calculating the protein's estimated titration curve (Section 4.5.3). Then the Q_{cryst} (*Hidden Observable*) is obtained from the estimated titration curve by assigning the $Q_{cryst} = Q$ where the solution $pH = pH_{cryst}$. Because the Q is dependent upon its amino acid content and size, a larger protein can have a wider range of possible Q values due to the presence of more charged residues. Therefore, an attempt was made to adjust for variations in size, surface area, and Q by calculating either the estimated specific charge ($\overline{Q}$) or the estimated surface charge density (σ), which both are also considered as *Hidden Controllables*. The $\overline{Q}$ is the ratio of the Q to the protein mass, expressed here in units of electrons (*e*)/kDa. The σ is the ratio of the Q to the A_S of the protein; a convenient unit is millielectrons (m*e*)/nm^2 (10^{-3} electron charges per square nm). Although the estimation of surface area is difficult, σ facilitates comparisons with other biologically relevant macromolecules. An example of a $\overline{Q}$ curve is shown in Appendix C, Figure C.2b. The $\overline{Q}_{cryst}$ and σ_{cryst} (*Hidden Observables*) were obtained from the $\overline{Q}$ and σ curves the same way as previously mentioned for Q_{cryst}. One of the hazards of using hidden variables is that they introduce additional information and care must be taken to ensure that "discoveries" based on the hidden variables are not simply reflections of that additional information.

Barlow and Thornton (1986) have shown that the surface charge density varied little among proteins, where they defined the surface charge density as the number of charged residues per surface contact area in Å^2. For the current study, the A_S would have to be estimated without the solved 3D structure (Section 4.5.2). Additionally, the σ that was examined in this study was

defined as Q /A_S. It was hypothesized that the $\overline{Q}$ and σ of crystallized proteins ($\overline{Q}_{cryst}$ or σ_{cryst}) would be relatively constant.

Because the solution pH controls the protein's charge (Figure 4.3e), the Q_{cryst}, $\overline{Q}_{cryst}$, and σ_{cryst} were examined to determine if any of these *Hidden Observables* could be used as a predictor of the solution pH ranges leading to successful crystallization (Figure 4.3f). From the pH_{cryst} listed within the PDB, the Q_{cryst} could be estimated using the HHE (Equation 2.1). The pH controls the Q by controlling the amount of protons (H^+) available in the solution. A large amount of protons will cause Lysine, Arginine, and Histidine to become positively charged. Alternatively, a small amount of protons in solution will cause the deprotonation of Glutamic acid, Aspartic acid, Cysteine, and Tyrosine. Therefore, the $\overline{Q}_{cryst}$ was hypothesized to be a proxy variable for pH_{cryst} and was used as the outcome variable (Figure 4.3f).

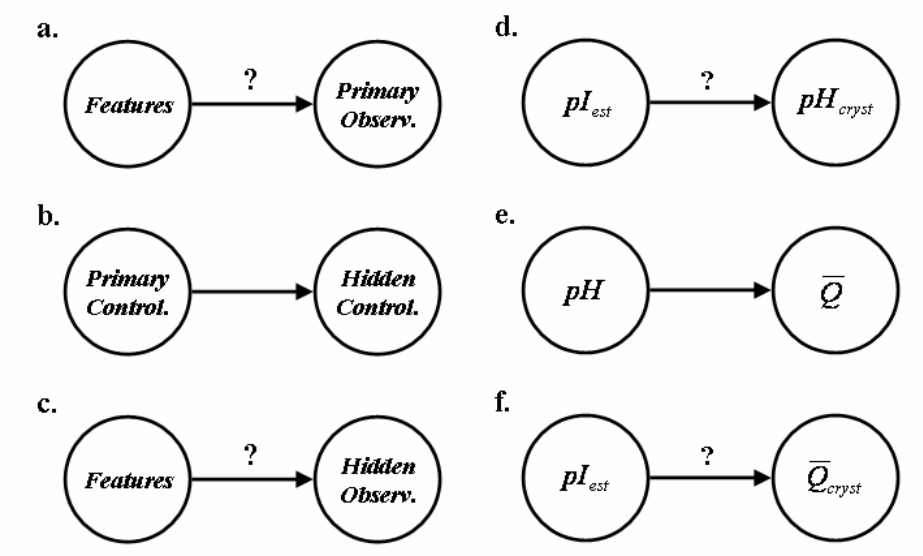


Figure 4.3 Can a protein's biophysical properties (*Features*) be used to predict crystallization conditions (*Observables*)?

4.6 CASE SELECTION

For the initial investigations, only non-redundant, non-membrane proteins were chosen and nucleic acids, membrane proteins, structures with poor resolution, and peptides were excluded. It was expected that nucleic acids would behave quite differently than would proteins for crystallization, because nucleic acids are more charged than proteins. Additionally, membrane proteins are a special case of proteins, which are generally difficult to crystallize, requiring special conditions. Finally, the PDB contains a large amount of redundancy. Two proteins that differ in a single amino acid are different entries. The same protein from two different organisms can also be present in the database. Identification of these instances and their removal should eliminate a potential source of bias.

4.6.1 Macromolecular Type and Method

The seqres.txt file identified whether the structure contained proteins and/or nucleic acids. This was found within the mol:*xxx* description field, where *xxx* is 'protein,' 'protein-het,' 'nucleic,' or 'nucleic-het.' The '-het' indicated the presence of an 'X' within the sequence, which codes for an unknown amino acid or nucleic acid. The unspecified group complicates calculation of the *Hidden Features*, so any structure with an 'X' within its sequence (protein-het or nucleic-het) was removed. Additionally, any structures with an amino acid length less than 20 were removed.

While the seqres.txt file does not contain the experimental method of structure determination, another file (entries.idx) found at http://pdbbeta.rcsb.org/robohelp/FTP_Server/ftp_derived_data_index.htm does. The tab-delimited entries.idx file contains a column for experiment type. Possible experiment types include X-ray Diffraction, NMR, Electron Microscopy, Synchrotron X-ray Diffraction, Neutron Diffraction, Fiber Diffraction, Fluorescence Transfer, and Theoretical Models. The number of structures within the PDB solved by each method is listed in Table 4.3. X-ray diffraction remains the current method of choice, and generally provides the most molecular detail. Due to possible differences between methods, structures not solved by X-ray diffraction were removed. An additional constraint on the resolution limit of diffraction ($diff_{lim}$) was imposed, a $diff_{lim} \leq 3.0$ Angstroms (Å). At a $diff_{lim}$

greater than 3.0 Å, detailed information about the 3D structure is lost. Therefore, a $diff_{lim}$ of 3.0 Å was set as the high threshold.

Table 4.3 Method of structure determination as listed in the PDB (11/14/2005)

Method	**Frequency**
X-ray Diffraction	28,535
NMR	4,919
Synchrotron Diffraction	32
Electron Microscopy	55
Neutron Diffraction	17
Fiber Diffraction	24
Theoretical Model	2
Fluorescence	1
Total	33,585

4.6.2 Membrane Proteins

Membrane proteins are a special case of proteins. They are inherently difficult to crystallize, because they require the presence of a detergent or lipid bilayer in order to achieve solubility. A protein has to have a certain level of solubility in order to crystallize. Therefore, all structures containing membrane proteins were removed. Membrane proteins were identified in three ways. The first method was to directly search the PDB website using the keyword "membrane protein", which retrieved 718 structures on 11/14/2005. The second method flagged 360 structures that were listed as being membrane proteins in the PDB-REPRDB (Noguchi and Akiyama, 2003; http://www.cbrc.jp/pdbreprdb-cgi/reprdb_menu.pl) on 11/16/05. The final method examined the 'header field' in the 'entries.idx' file available from the PDB. Using this file, membrane proteins were identified with a search for '*membrane*', which resulted in 405 structures being labeled as membrane proteins. A conservative approach was then used in which a structure was discarded

if any of the three methods identified the structure as having a membrane protein chain (862 structures). There appeared to be some time delay for entries among the differing sources. A Venn diagram showing the overlap between the three methods is shown in Figure 4.4 (updated November 16, 2005).

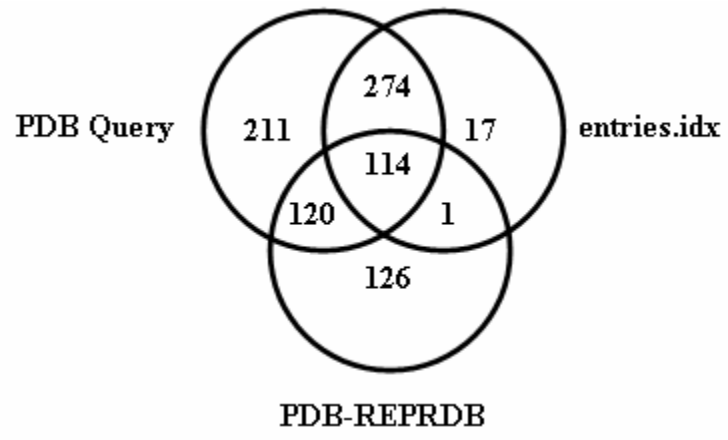


Figure 4.4 The Venn-diagram showing the frequency of proteins being identified as membrane proteins.

4.6.3 Redundancy

The PDB contains many instances where structures are different by a single amino acid mutation or by the source species. These multiple instances, which may number into the hundreds, were thought to be a potential source of bias for any analysis. For example, a PDB search for 'T4 lysozyme' using the method of X-ray diffraction returns 419 structures (11/22/2005). In order to account for similar sequences, a non-redundant (BLAST p-value of 10^{-80} similarity) data set was created. The list of non-redundant PDB entries was created using the PDB and the nrPDB (Holm and Sander, 1998) at the NCBI (http://www.ncbi.nlm.nih.gov/Structure/VAST/nrpdb.html). Chains with a low similarity score (BLAST p-value of 10^{-80}) were selected that had the highest structural quality as identified within the nrPDB database (Section 4.1.2) with a pH_{cryst}.

4.7 STATISTICAL MODEL BUILDING

First, the distributions and descriptive statistics of all variables were examined to determine normality. As appropriate, transformations were performed on the raw data in order to approximate normal distributions. Next, the intercorrelations between all *Features* and *Controllables* were calculated in order to detect multicollinearity among variables.

First, zero-order correlations (Spearman's rho) were examined between the *Given Features* (*Primary* and *Hidden*) and *Given Observables* (*Primary* and *Hidden*). *Features* that were significantly correlated with a *Given Observables* (pH_{cryst}, Q_{cryst}, $\overline{Q}_{cryst}$, or σ_{cryst}) might be able to predict that *Observable*. A significant correlation among *Observables* indicated a relationship of the variables that may allow the substitution of one of the correlated *Observables* for prediction purposes, i.e. proxy variables. Finding the *Features* that are predictive of a *Observable* may allow for predicting initial crystallization conditions for that *Observable*.

The initial hypothesis was that the estimated net charge of the protein can be used as a proxy variable for selecting the solution pH ranges for screen design. The solution pH has long been known to play a primary role for crystallization conditions. The solution pH controls the amount of protons present in solution, which may protonate or deprotonate the charged amino acids. Previous research has failed to identify a relationship between a *Feature* and the pH_{cryst} (*Observable*).

The second hypothesis was that there are groups of structures within the PDB, which display different *Primary* (pH_{cryst}) or *Hidden Observable* (Q_{cryst}, $\overline{Q}_{cryst}$, and σ_{cryst}) distributions that can be used for suggesting crystallization conditions on new test proteins. Various methods were examined for grouping structures by 'similarity,' including simple binning (Section 4.7.1) and unsupervised clustering methods (Section 4.7.2).

4.7.1 Creation of Groups by Binning

One method to group structures was by binning on the basis of a *Primary* (AA composition) or *Hidden Feature* (MW_{au} or pI_{est}; Chapter 6). The groups could be created by multiple methods,

such as dividing structures into an even number per group, for example, by using quartiles or deciles. Alternatively, the mean or median of a variable could be used to split structures into two groups, such as a median split. Groups could also be broken down by the distance from the mean using the standard deviation (SD). These methods might also be based on prior knowledge (supervised) to create separation points, if available.

For example, proteins were split into groups by their molecular weight using the distance from the mean value. Because the MW_{au} distribution is highly skewed, the MW_{au} values were transformed by taking the natural log (ln) of all MW_{au} values, which resulted in a relatively normal distribution. The mean±SD of the $\ln(MW_{au})$ distribution was then used to create three or five groups of proteins. For discretizing the three $\ln(MW_{au})$ groups, all proteins within 1 SD of the mean $\ln(MW_{au})$ were labeled 'Average.' Those proteins whose $\ln(MW_{au})$ was either greater than 1 SD below or above the mean were labeled as 'Small' and 'Large,' respectively. When proteins were further separated into five groups, the 'Small' and 'Large' groups were split into a 'Very Small' and 'Very Large' group, by considering whether the ln(MW) value was 2+ SD below or above the mean $\ln(MW_{au})$ value.

A similar method of discretization was used for the protein's pI_{est}. Proteins whose pI_{est} was less than or equal to 6.0 were labeled as 'Acidic.' Proteins with a pI_{est} greater than or equal to 8.0 were labeled as 'Basic,' while the remaining proteins were considered as 'Neutral,' $6.0 < pI_{est} < 8.0$ (Chapter 3). This discretization based on pI_{est} follows the description of Ries-Kautt and Ducruix (1999). After determining differences between the three pI_{est} bins (Chapter 3), the 'Acidic' and 'Basic' groups were further broken down into 'Very Acidic' and 'Very Basic' groups. Thus, the pI_{est} range of both the 'Acidic' and 'Basic' groups was reduced to $5.0 < pI_{est} \leq 6.0$ and $8.0 \leq pI_{est} < 9.0$, respectively. The 'Very Acidic' group had a $pI_{est} \leq 5.0$, while the 'Very Basic' group had a $pI_{est} \geq 9.0$. However, binning might not represent the optimal separation of proteins, as proteins with a 0.1 pI_{est} difference, such as 6.0 and 6.1, may be grouped with proteins that have a 0.9 lower pI_{est} (5.1) or a 1.8 higher pI_{est} (7.9). Therefore, an alternative strategy was also examined, unsupervised clustering.

4.7.2 Creation of Groups by Unsupervised Clustering

Other methods of grouping were also examined, including various unsupervised clustering algorithms (two-step clustering and self-organizing maps). These clustering methods were used to group structures by multiple *Features*, creating a feature vector (Chapter 6). After the unsupervised clusters were formed, Spearman's correlation values were examined among the variables to determine what *Features* were associated with the groupings. It was hypothesized that unsupervised clustering would result in a higher accuracy than binning.

4.7.2.1 Two-Step Clustering

Two-step clustering (2Step), as its name suggests, consists of two steps, pre-clustering and clustering (Chiu et al., 2001). An additional positive feature was that the two-step clustering algorithm could handle either continuous or categorical variables. The algorithm assumes that the continuous variables are normally distributed and independent. However, the algorithm is relatively robust to violations of normality and independence.

In the first step, the data are converted into a cluster feature tree. This involves an initial pass through the data to cluster similar items (x). Various statistics are used to represent the clusters, such as the number of items, the mean, and variance of each continuous feature. Starting from the root node, each item is passed to the child node with the minimum distance to the item. Once a leaf node is reached, the algorithm determines whether the item is within a threshold distance from the leaf node. Two distance measures can be used for each step, log-likelihood or Euclidean distance. If the distance between the leaf node and item x is below a threshold distance, x is placed within the leaf and the leaf's attributes are recalculated. If the x's distance from the leaf node is greater than the threshold, it forms a new leaf node. This results in a much smaller data set.

The second step uses agglomerative hierarchical cluster analysis with the log-likelihood function (Equation 4.1) to place items in the set number of clusters (J). For each item's (x) feature vector (θ_j), the probability of the item belonging to cluster C_j was determined, $P(x_i|\theta_j)$. The item is assigned to the cluster with the maximum likelihood contribution (Equation 4.2),

where l_{Cj} is the log-likelihood contribution from cluster C_j. $P(x_i|\theta_j)$ is the probability density function of x being in cluster C_j, and θ_j is the feature vector.

Equation 4.1 $$I = \sum_{j=1}^{J} \sum_{i \in I_j} \log P(x_i \mid \theta_i) = \sum_{j=1}^{J} I_{C_j}$$

Equation 4.2 $$I_{C_j} = \sum_{i \in I_j} \log P(x_i \mid \theta_j)$$

4.7.2.2 Self-Organizing Maps

Another unsupervised clustering algorithm, self-organizing maps (SOMs), was examined. SOMs have previously been used to solve many biological problems, such as clustering protein families from sequence (Ferran et al., 1994; Andrade et al., 1997), prediction of a proteins structural class from amino acid composition (Cai et al., 2000), prediction of genes (Mahony et al., 2003), and classification of crystallization drop images (Spraggon et al., 2002). A self-organizing map (SOM; Kohonen, 2001) is considered a 2D non-linear projection of high-dimensional data, composed of an X by Y matrix of neurons ($m_{x,y}$; Figure 4.5). The SOM compresses the multi-dimensional data into the 2D feature space, while preserving the spatial relationship between the data. Items in the dataset that are more similar are placed closer together in the 2D space, thus preserving the spatial relationship between items.

Each item (n) could be represented by a vector (v) of 1 to i features (f), which could be composed of a combination of *Primary* (amino acid composition) and/or *Hidden* (MW_{au}, pI_{est}, and estimated titration curve) *Features*. Each item vector ($v_n=[f_1, f_2,..., f_i]$), has its distance, usually Euclidean, measured to each $m_{x,y}$'s feature vector ($m_{x,y}=[f_1, f_2,..., f_i]$. The neuron with the minimum distance (Equation 4.3) to the input vector is the Best-Matching Unit (BMU), c. Next, the BMU's feature vector is recalculated to more closely resemble the input vector. Additionally, the BMU's neighboring neuron's feature vectors may also be adjusted to a lesser degree. Each item vector is presented to the SOM algorithm over a given number of iterations. The $m_{x,y}$ feature vectors are then adjusted by their contents with a decreasing weight after each iteration.

Equation 4.3 $$c = \arg\min_i \{\|v_i - m_{x,y}\|\}$$

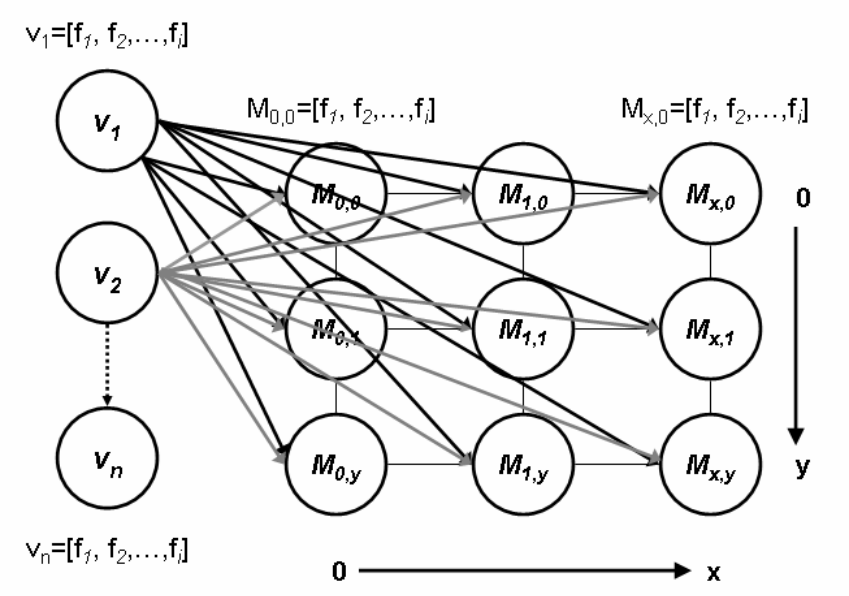


Figure 4.5 The SOM algorithm presents each feature vector, *v*, composed of *i* features (*f*) to each neuron's feature vector, $m_{x,y}$.

4.7.3 Comparison of groups within each clustering technique

After splitting all training set protein structures into groups, statistical differences in the *Hidden Features*, such as the pI_{est} or MW_{au}, and *Observables* (pH_{cryst}, Q_{cryst}, $\overline{Q}_{cryst}$, and σ_{cryst}) distributions were examined, using nonparametric tests (Section 4.10). If significant differences in *Observable* distributions were observed among groups, a model was developed to predict the *Observable* given the protein *Features*. Significant differences or correlations (Spearman rank) in *Hidden Features* among groups indicated that these *Hidden Features* might be used to predict any *Primary* or *Hidden Observable* that was significantly different between groups.

4.7.4 Modeling with Gaussians

The ability to model the *Primary* (pH_{cryst}) or *Hidden Observable* (Q_{cryst}, $\overline{Q}_{cryst}$, or σ_{cryst}) distributions with Gaussians was also examined for each grouping method. Using the observed

mean, Gaussians (Equation 4.4) were fit to (a) all structures *Observable* distributions and then (b) each group (*i*) individually, where y_i = observed value within group *i*; μ_i = mean of group *i*; σ_i = standard deviation (SD) of group *i*. The best-fit Gaussian was determined by minimizing the residual sum of square (RSS) between the observed and the predicted values (Equation 4.5). The value of the SD was adjusted in small increments (0.01 units) in order to determine the best-fit SD (SD_{bf}) with the minimum RSS.

Equation 4.4 $$f(y_i) = \frac{1}{\sigma_i \sqrt{2\pi}} e^{\frac{-(y_i - \mu_i)^2}{2\sigma_i^2}}$$

Equation 4.5 $$RSS = \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2$$

After determining the best-fit Gaussian, the number of proteins within 1 SD_{bf} for both the training and test sets was calculated. In order to compare all groups within a given grouping method, a common SD for all groups was chosen based on examining the SD_{bf} of all groups (rounding the decimal off to the tenths).

4.7.5 Comparing Binning to Clustering

The ability of each method to suggest ranges of the output variable (*Primary* or *Hidden Observable*) on an independent test set was used to compare methods (see Section 4.8) using predictive accuracy as a performance measure.

4.7.6 Determining the optimal number of clusters

After comparing the unsupervised clustering techniques to binning in the previous section, the number of clusters was allowed to vary. Rather than arbitrarily setting the number of clusters, the question was asked, "How many clusters are there within the data?" Both 2Step clustering and SOMs were able to answer this question, because the algorithms have established methods to predict the 'correct' number of clusters.

As mentioned in Section 4.7.2.1, the 2Step algorithm can predict the number of clusters in a dataset. The algorithm computes either the Schwarz's Bayesian Criterion (BIC; Schwarz, 1978) or Akaike Information Criterion (AIC; Akaike, 1983) to choose the number of clusters. A large drop off in the information criterion tells the algorithm what the 'correct' number of clusters should be.

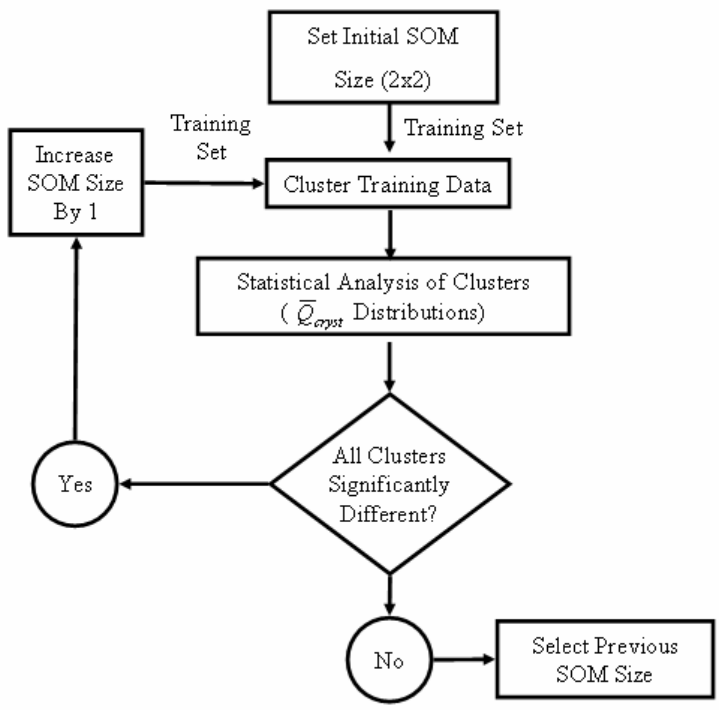


Figure 4.6 The Supervised SOM algorithm for determining the dimensions of the best-fit self-organizing map.

Alternatively, there are several SOM algorithms that have the ability to determine the 'correct' number of clusters, including the GSOM algorithm, which was used in the Preliminary Results (Section 3.2.2). Additionally, the Supervised SOM algorithm (Figure 4.6) was developed to determine the optimum SOM dimensions (i.e. number of clusters). Initially a 2x2 SOM was initialized with the maximum number of clusters set to 16. Sixteen would be the maximum number of groups if the structures were binned by their pI_{est} every 0.5 units from 4.0-12.0. Starting with an initial 2x2 SOM, each dimension of the SOM was increased one at a time starting with the x-dimension, i.e. the next SOM investigated would be 3x2. It was observed that if the x-dimension increased and the y-dimension stayed at 2, a one-dimensional SOM resulted,

with no structures in the 2nd y-dimension. Therefore, the x-dimension of the SOM was allowed to increase above 8, where an 8x2 would theoretically allow for sixteen groups. With the y-dimension staying at 2, the x-dimension was allowed to increase to a maximum value of 16. The stopping criterion was finding clusters that did not differ significantly from each other in their $\overline{Q}_{cryst}$ distributions.

4.7.7 Comparing Among All Grouping Methods

The ability of all methods to predict either the *Primary* (pH_{cryst}) or *Hidden* ($\overline{Q}_{cryst}$ or σ_{cryst}) *Observable* on an independent test set was used to rank the different methods. The number of test set structures whose *Observable* value was within a predicted range (accuracy) was calculated for each group in both the training and test sets.

4.8 TRANSLATION OF HIDDEN CONTROLLABLE INTO PRIMARY CONTROLLABLE SEARCH SPACE

If the protein structures in the test set follow a similar pattern observed in the training set, calculating a protein's estimated titration curve can theoretically be used for selecting pH ranges for crystallization attempts. Each protein exhibits a unique titration curve based upon its charged amino acid composition and MW (Appendix C; Figure C.1). It should be noted that proteins with different sequences, but the same charged amino acid compositions and similar molecular weights, may produce the same results. Some predictions may be of little use, due to relatively long flat regions along the titration curve in the most probable $\overline{Q}_{cryst}$ ranges. However, some predictions will narrow the pH search space to a narrow pH range (≤ 1 pH unit). Three methods are discussed in this paper for estimating the $P(pH = pH_{cryst} \mid \overline{Q} = \overline{Q}_{cryst}, PDB)$ over a pH range for the test set: (1) calculating the middle 50% confidence interval, (2) calculating a probability distribution, and (3) the Charge Range Test.

4.8.1 Confidence Interval Calculation

For the first method, a confidence interval (CI) was calculated from an *Observable* (pH_{cryst}, $\overline{Q}_{cryst}$, or σ_{cryst}) distribution of each cluster for the middle 50% (CI_{50}) of proteins. For example, the quartiles of the distribution were calculated for each group. The middle two quartiles (>25% and <75%) were used as the middle 50% range. These calculations form a CI for the predictions. The CI_{50} bracketed a specific *Controllable* range for each group where there would be an estimated 50/50 chance that similar new protein structures would have a *Observable* value in this range (Figure 4.7a). This range can then be applied to target proteins to select an initial *Controllable* range for crystallization attempts (Figure 4.7b).

The test set was used to determine if proteins not seen in the training set would follow a similar pattern. If the test set was representative of similar structures, 50% of them should fall in the CI_{50}. If less than 50% of the test proteins fall in the CI_{50} of the training set structures, this method would not be useful for prediction. The CI can be adjusted to make the distributions tighter (decrease the CI) or broader (increase the CI) depending upon the desires of the individual researcher. For example, the CI_{67} could be used to represent all structures within 1 SD of the mean *Observable* value.

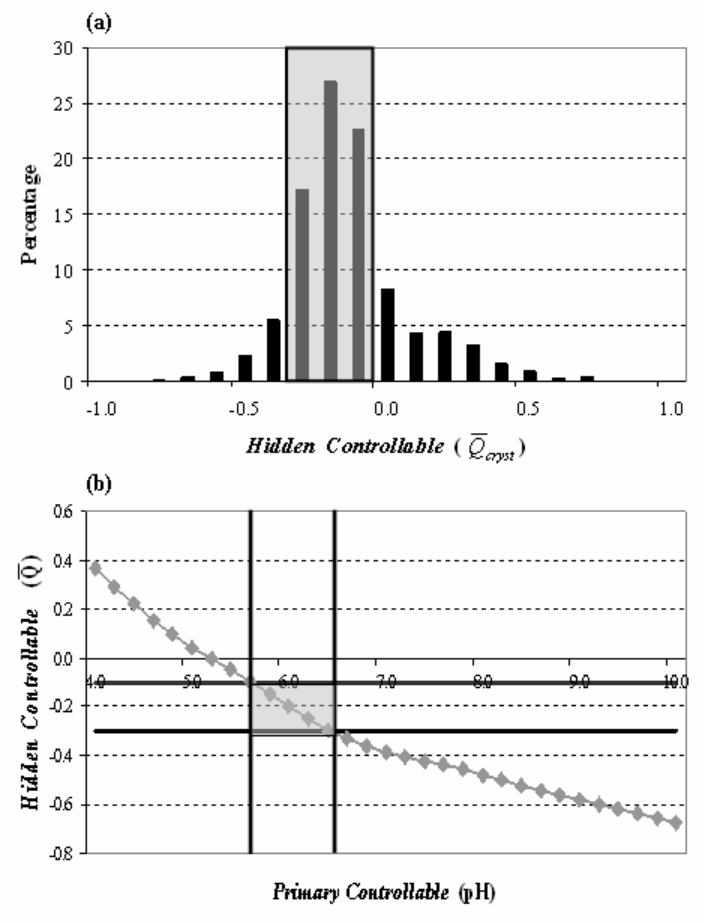


Figure 4.7 From a group's (a) distribution of a *Hidden Controllable*, the CI_{50} is calculated, and (b) applied to a test protein's $\overline{Q}$ curve to bracket a *Primary Controllable* range.

4.8.2 Probability distribution calculation

Based upon the amino acid sequences of the proteins in the training set (*Given Features*) and their pH_{cryst} (*Given Observable*), the $\overline{Q}_{cryst}$ frequencies (*Given Hidden Observables*) can be calculated and use to estimate the $P(\overline{Q} = \overline{Q}_{cryst} \mid data)$ (Figure 4.8a). These probabilities were calculated every 0.1 unit between -2.0 and +2.0 *e*/kDa by dividing the frequency of $\overline{Q}_{cryst}$ occurrence within each group by the number of proteins in that group. The range from -2.0 to +2.0 was chosen based on the observed $\overline{Q}_{cryst}$ distributions for proteins in the training set with

few proteins have a $\overline{Q}_{cryst}$ above or below this range. The probabilities for $\overline{Q} = \overline{Q}_{cryst}$ were then directly mapped onto the pH space, guiding selection of pH ranges for a test protein sequence (Figure 4.8c). The probability of success for a given $\overline{Q}$ value was calculated from the training dataset for (1) all proteins and (2) for each group (Figure 4.8a).

An estimated specific charge curve ($\overline{Q}$ vs. pH) can be calculated for any given protein sequence (Figure 4.8b). Every pH value has a $\overline{Q}$ value associated with it for a given protein. However, the same $\overline{Q}$ value can be associated with many pH values as seen from the titration curve in Figure 4.8b. The conditional probabilities are then calculated that a particular pH value over the selected range is the pH_{cryst} given the database (PDB) and that the $\overline{Q}$ is the $\overline{Q}_{cryst}$, $P(pH = pH_{cryst} \mid \overline{Q} = \overline{Q}_{cryst}, PDB)$. These pseudo-probabilities were calculated for every 0.5 pH units from 4.0-10.0. This narrow range was chosen because few structures have a pH_{cryst} outside this range (< 2%). The 0.5 steps in pH prediction were based upon the saw-tooth pattern in the pH_{cryst} distribution. This resulted in thirteen calculated probabilities for $P(pH = pH_{cryst} \mid \overline{Q} = \overline{Q}_{cryst}, PDB)$. Because long stretches of the titration curve can have similar $\overline{Q}$ values, the probabilities often added up to more than 100%. The probabilities were standardized to 100% by dividing each probability value by the total relative probability for that protein.

The most probable regions of the pH space (*Controllable*) could then be observed and a decision made as to where to set the pH values for the initial crystallization screens. Because proteins denature at the extreme pH ranges and the fact that few proteins crystallized at these ranges, probabilities were calculated over the pH range of 4.0 to 10.0 in 0.5 unit increments. The increments can be made coarser or finer. Alternatively, there could be a threshold, where only pH values that have a greater probability than the threshold are searched. The threshold can be raised or lowered accordingly. For this study, a threshold of 10% was used for high probability, 8-10% as low probability, and less than 8% as no better than a random pH selection. If one of the thirteen pH_{cryst} probability values were chosen randomly, a success rate of 7.7% would theoretically be observed. The other threshold value of 10% was arbitrarily chosen based on inspection of many probability distributions. If a 10% threshold is applied to the distribution in

Figure 4.8c, an initial pH range of 6.0-7.5 is suggested for crystallization attempts. The pH values of <5.0 or >7.5 have a probability no better than random (.06) and comprise 38% of the commercial screens with a listed pH. This would account for a significant reduction (62%) in experimental conditions. Additionally, other informative probability distributions could be obtained if we conditioned on more than just the $\overline{Q}_{cryst}$, for example, by using information related to the slope of the titration curve at the pH_{cryst}.[1]

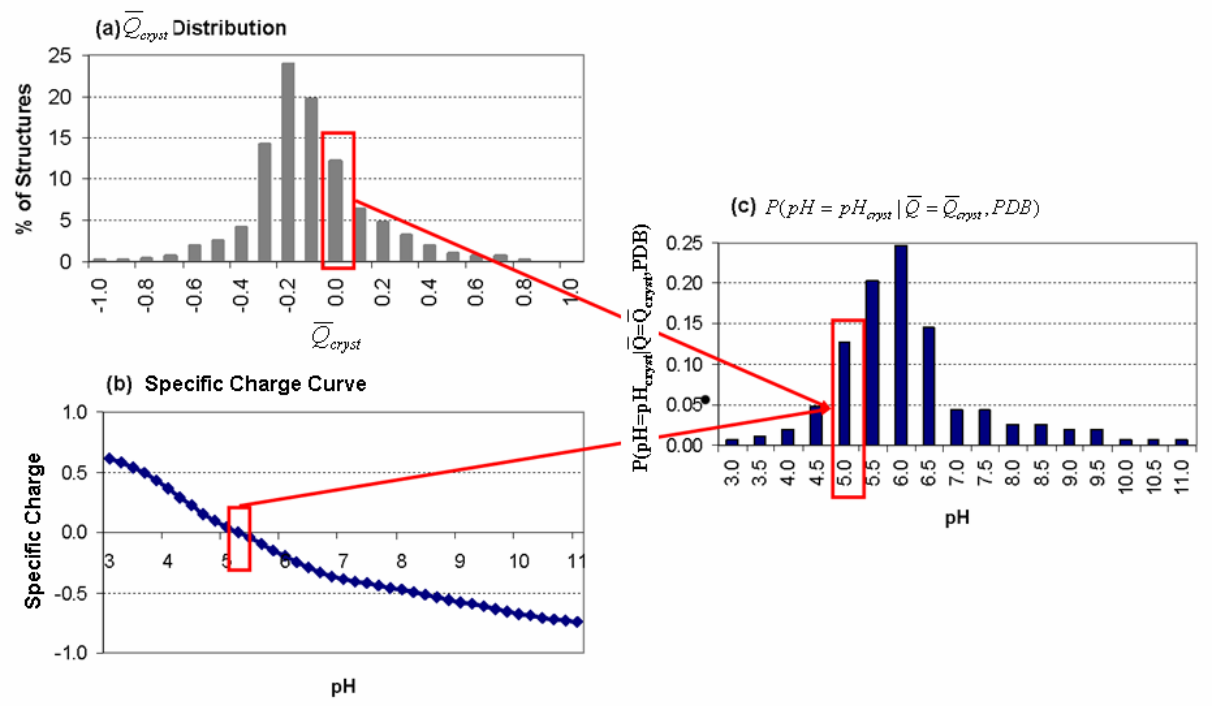


Figure 4.8 Creating a pH_{cryst} probability distribution from (a) a $\overline{Q}_{cryst}$ distribution and (b) a test protein's $\overline{Q}$ curve to obtain (c) $P(pH = pH_{cryst} | \overline{Q} = \overline{Q}_{cryst}, PDB)$.

[1] An alternative method for generating probabilities is to calculate the derivative of the estimated titration curve and then take the product of its absolute value and its frequency. This calculation would result in a different shape of the probability distribution and would alter all of the findings of the subsequent analyses.

4.8.3 Charge Range Test

Because the CI_{50} range can be quite different between groups, a third method was developed, the Charge Range Test. This method simply uses the $\overline{Q}_{cryst}$ or σ_{cryst} values (*Hidden Observables*) for all training set proteins and examines the percentage of proteins within a given range of the mean (Figure 4.9). The ranges examined were every ±0.1 $\overline{Q}$ or ±10 σ values up to the calculated standard deviation (SD). The frequency counts were converted to percentages and used as pseudo-probabilities. The observed percentage indicated the probability that a new protein structure would crystallize within the given $\overline{Q}$ or σ range (*Controllable*). This allowed for a more fair comparison among grouping methods, which may have widely different CI_{50} ranges.

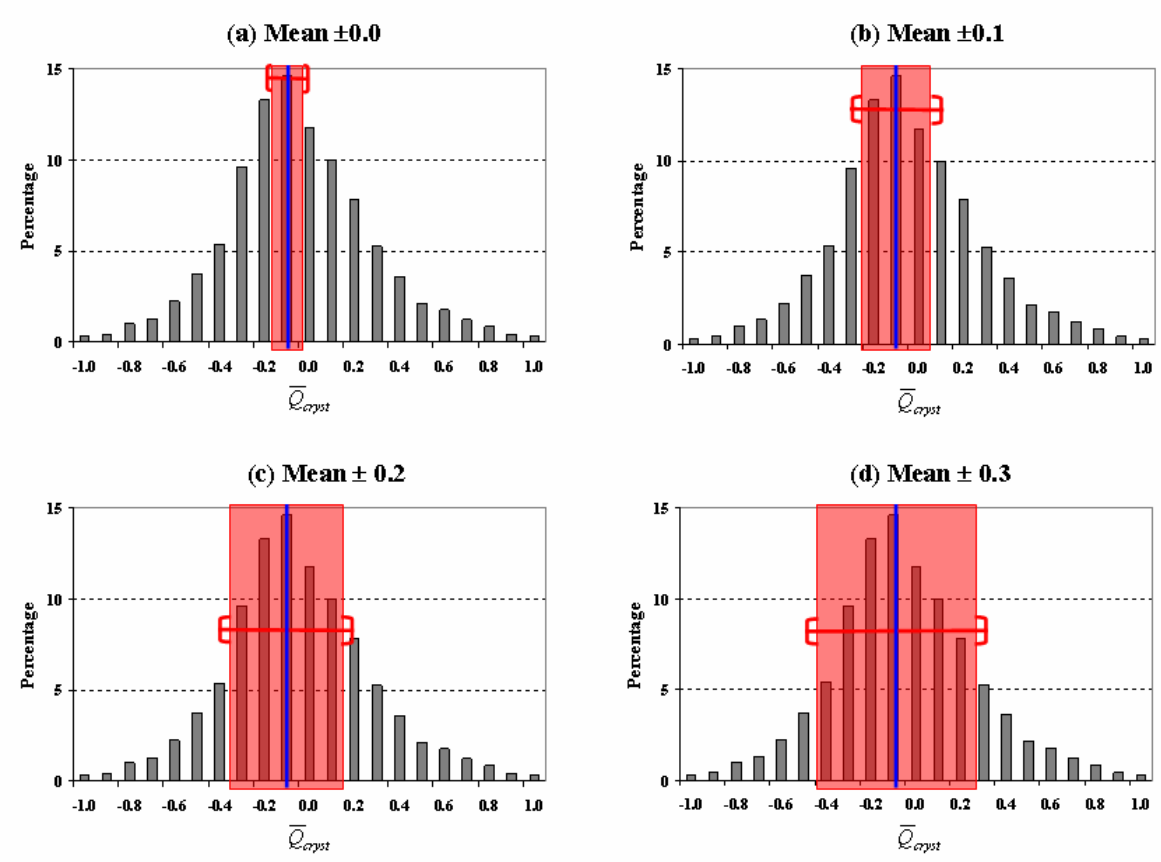


Figure 4.9 The Charge Range Test calculates the percentage of proteins within a given interval of the groups mean $\overline{Q}_{cryst}$ value, (a) Mean±0.0, (b) Mean±0.1, (c) Mean±0.2, and Mean±0.3.

4.9 CREATION AND USE OF TEST SET

The PDB is updated weekly with new structures. A newer nrPDB version (11/07/2005) was used to create a test set. Using the same framework (Figure 4.1b), the *Primary Feature* and *Observables* were obtained from the PDB (Sections 4.3-4.4). *Hidden Features* were created as in Section 4.5 and cases selected (Section 4.6) to create a non-redundant test set of 1,246 proteins that had a low sequence similarity, BLAST p-value < 10^{-80}, to the proteins in the original training set. The test set proteins were then used in the statistical models developed from the training set in Section 4.7 to predict a $\overline{Q}$ or σ range (*Hidden Controllable*). These ranges could then be translated back into the *Primary Controllable* search space (Section 4.8). The ability of the original training set to predict the *Observables*, *Primary* (pH_{cryst}) or *Hidden* ($\overline{Q}_{cryst}$ or σ_{cryst}), for these new proteins was then examined as validation for the methods of analysis. When comparing two different methods, the method that predicted the test set proteins more accurately was the better model.

4.10 STATISTICAL ANALYSIS

After groups of structures were formed, statistical measures were performed in order to determine if there were any significant differences ($p<0.01$) between the *Primary* (pH_{cryst}) and *Hidden* ($\overline{Q}_{cryst}$) *Observables* among the groups (Conover, 1999). When there were more than two groups, a Kruskal-Wallis test (a nonparametric ANOVA) was first performed to determine whether there were any significant differences in the location or shape of the distributions. For the Kruskal-Wallis (KW) test, the individual observations were first ranked from 1 to N, where N was the total number of observations. The ranks were then summed, R_i, for each group, *i*. The test statistic, T_{KW}, was computed from Equation 4.6, where the sample variance (S^2) was computed from Equation 4.7. The number of groups was represented by *k*. The chi-square distribution was then used for the determination of the null distribution of T_{KW}. The null

hypothesis was that all groups' distributions were identical with a given significance level $\alpha<0.01$. If T_{KW} was greater than the 1-α null distribution value, the null hypothesis was rejected.

Equation 4.6 $$T = \frac{1}{S^2}\left(\sum_{i=1}^{k} \frac{R_i^2}{n_i} - \frac{N(N+1)^2}{4} \right)$$, where

Equation 4.7 $$S^2 = \frac{1}{N-1}\left(\sum_{\substack{all \\ ranks}} R\left(X_{ij}\right)^2 - N\frac{(N+1)^2}{4} \right)$$

If significant differences ($p<0.01$) were found between the distributions, a Kolmogorov-Smirnov (KS) test was performed pair-wise to locate the individual differences in the location (means or median) and/or shape of the distributions (variance) between groups. The two-sided KS-test compared the vertical distance between the two empirical distributions. The test statistic, T_{KS}, was the largest absolute difference between the two distributions, S_1 and S_2 (Equation 4.8). Similar to the KW tests, the null hypothesis was that the two groups' distributions were identical at an α of 0.01. If T_{KS} was greater than the 1-α null distribution value, the null hypothesis was rejected (Conover, 1999). A nonparametric test was chosen because of the large differences in size and variance between the groups and independence from the assumption of normality. SPSS (Chicago, IL) version 13.0 was used for all analysis.

Equation 4.8 $$T_{KS} = \sup_x \left| S_1(x) - S_2(x) \right|$$

4.11 CHAPTER SUMMARY

The PSPE framework presented here was applied to the PDB to gain insight into the solution pH requirement for protein crystallization; however, it could be used for other analyses of biological sequences, structures, or folds. Often a certain representative subset of sequences or structures is desired for analysis, such as one member of a certain protein family. Biological databases often contain multiple instances of these biological molecules. There are many reasons for this, such as differences in the source organism or biological mutants. While these instances are very important, usually they are not needed for a specific analysis. A researcher may often just require one representative sequence with special constraints. The analysis presented here was a

good example of just needing one representative sequence/protein, while requiring the protein to have a pH_{cryst}. The PSPE general framework could be applied to the whole protein universe (UniProt; Apweiler et al., 2004), protein families (Pfam, Bateman et al., 2004), or protein folds (SCOP, Murzin et al., 1995; CATH, Orengo et al., 1997). Additionally, the framework could be applied to the prediction of signal peptides (Bendtsen, 2004), disordered peptides (Jones and Ward, 2003; Oldfield et al., 2005), or even prediction of Major Histocompatability Complex (MHC) binding peptides to T-cell epitopes (Zhu et al., 2006, Guan et al., 2006}.

While the pH_{cryst} was the most reported crystallization parameter reported in the PDB, there are many other variables to which this method could be applied (see Table 1.1 in Chapter 1). For more potential future applications in the area of protein crystallization, see Chapter 9. In the next chapter the PSPE Framework is applied to test the hypothesis that a protein's $\overline{Q}$ or σ values can be used to select solution pH ranges that have a higher probability of resulting in the formation of a well-ordered crystal.

5.0 A SPECIFIC APPLICATION OF THE PROTEIN SEQUENCE-PROPERTIES EVALUATION (PSPE) FRAMEWORK

In the previous chapter, the Protein Sequence-Properties Evaluation (PSPE) framework was developed to analyze differences in crystallization behavior of proteins. Meanwhile, the initial analysis of the Protein Data Bank (PDB) presented in Chapter 3 demonstrated a weak link between the pH_{cryst} and the pI_{est}. However, the estimated net charge at the pH_{cryst} (Q_{cryst}) demonstrated some potential use as an *Observable*, being highly correlated to both a *Given Feature* (pI_{est}) and a *Primary Observable* (pH_{cryst}). Additionally, the Q_{cryst} distributions displayed more statistically significant differences between MW_{au} or pI_{est} groups than did the pH_{cryst} distributions. Therefore, it was hypothesized that the specific charge ($\overline{Q}$) and estimated surface charge density (σ) of previously crystallized proteins could be used to suggest solution pH ranges with a higher probability of success (i.e. crystals) for new target proteins.

The previous results also indicated MW_{au} and pI_{est} effects on crystallization conditions. In this chapter, using a newer version of the PDB (October 20004) with more training set cases and a larger independent test set (November 2005), attempts were made to account for the MW effects in the calculation of Q_{cryst}, by calculating the $\overline{Q}$ at the pH_{cryst} ($\overline{Q}_{cryst}$). The $\overline{Q}_{cryst}$ was computed by dividing Q_{cryst} by the MW_{au} to obtain the $\overline{Q}_{cryst}$ values in electrons/kilodalton (*e*/kDa). Similarly, the estimated average surface charge density (σ_{cryst}) was calculated by dividing Q_{cryst} by the estimated solvent accessible surface area (A_S) to obtain the σ_{cryst} values in millielectrons/square nanometer (me/nm^2).

The hypothesis examined in this chapter was that the $\overline{Q}_{cryst}$ and σ_{cryst} can be used to guide researchers to intelligently select the initial pH ranges for screening attempts. Instead of using the Q_{cryst}, the $\overline{Q}_{cryst}$ and σ_{cryst} were calculated in an attempt to account for size effects.

Because both of these variables are largely controlled by the solution pH, any models developed could be used to determine the most probable pH_{cryst} ranges for initial screen design.

5.1 METHODS

The training set was composed from PDB structures downloaded on October 5th, 2004, while the test set was downloaded on November 7, 2005. The *Primary Feature*, amino acid sequence, was obtained from the seqres.txt file as described in Section 4.4. The pH_{cryst} values (*Primary Observable*) were queried and downloaded from the PDB in November 2004 (Section 4.4).

Next, several *Hidden Features* and *Observables* were calculated as described in Section 4.5, such as the asymmetric unit molecular weight (MW_{au}) (Section 4.5.1), A_S (Section 4.5.2), and the estimated titration curve (Section 4.5.3). From the estimated titration curve and the pH_{cryst}, both the pI_{est} and Q_{cryst} were obtained. By dividing each Q value along the estimated titration curve by the MW_{au}, the $\overline{Q}$ was calculated. From this $\overline{Q}$ curve, the $\overline{Q}_{cryst}$ was obtained similarly to Q_{cryst} (Section 4.5.4). For the σ curve, the correct equation for A_S had to be determined, as described in Section 4.5.2. Once the A_S values are calculated, each Q value along the estimated titration curve was divided by the A_S to obtain an σ value in m*e*/nm^2. The σ value along the curve where the $pH = pH_{cryst}$ becomes the σ_{cryst} (Section 4.5.4).

Individual cases were selected with the applied filters and constraints as described in Section 4.6. The final training set consisted of 4,114 non-redundant protein structures ($nrPDB_{10.04.05}$). The final test set used for this dissertation was created as described in Section 4.9. The final test set contained 1,246 non-redundant protein structures that had a low sequence similarity to the training set proteins.

5.2 RESULTS (DEVELOP STATISTICAL MODELS)

5.2.1 Calculation of Accessible Surface Area (A_S) and Average Surface Charge Density (σ)

Because a protein's accessible surface area was found to be proportional to its molecular weight (Chothia, 1975; Janin, 1976; Teller, 1976; Miller et al., 1987ab; Janin et al., 1988), attempts were made to adjust for variations in size (standardize), surface area, and net charge by dividing the estimated net charge, Q, by either the protein's MW_{au} or A_S to obtain the $\overline{Q}$ or σ, respectively. To estimate the A_S, several assumptions were made. First, the protein's σ_{cryst} was assumed to be relatively constant, being proportional to the Q_{cryst}/A_S. The finding by Barlow and Thornton (1986) that the surface charge density of proteins was relatively constant made this assumption reasonable. However, the A_S had to first be approximated by the data. In order to do this, a least squares fit of the absolute value (abs) of Q_{cryst} = a*(MW)^b was performed, which resulted in Equation 5.1. The correlation of the abs(Q_{cryst}) and MW_{au} was r = 0.664 (Pearson) or r = 0.525 (Spearman's rho), with both being highly significant, p<0.0001.

Equation 5.1 $$abs(Q_{cryst}) = 1.3476 * (MW_{au})^{0.806}$$

The scatter-plot of the ln(MW_{au}) by ln(abs(Q_{cryst})) is shown in Figure 5.1. From the linear regression line (r^2 = 0.264; p<0.0001), Equation 5.2 was devised. Removing the natural logs resulted in Equation 5.3. This was basically the same result as the least squares fit in Equation 5.1. Next, the σ using Equation 5.3 for the A_S (σ_{ours}) was plotted against the σ_{miller}, which uses the equation for the A_S determined by Miller et al. (1987a). This was done in order to determine the scale factor to more accurately represent A_S using the actual values (Figure 5.2). From the graph, the linear regression equation was calculated to be Equation 5.4. From this equation, a scale factor of 2.108 was used to convert our estimation of the A_S ($A_{S.ours}$) more closely to the observed and estimated A_S values calculated for monomers by Miller et al.

(1987a), $A_{S.miller}$. The plot of the $A_{S.ours}$ versus $A_{S.miller}$ is shown in Figure 5.3. Thus, our calculation of the A_S for very large proteins may be slightly off. Although many of the protein structures were not monomers, this was felt to be a close approximation. After determining the appropriate equation for the calculation of the σ_{cryst}, the analysis of crystallization variables continued.

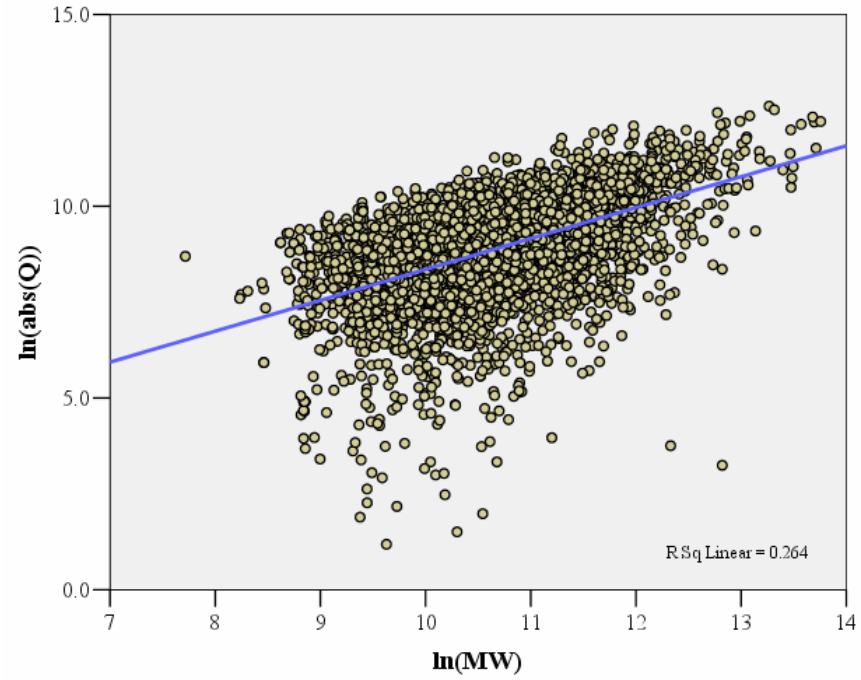


Figure 5.1 Plot of the ln(MW) against the ln(abs(Q_{cryst})).

Equation 5.2 $\ln(abs(Q_{cryst})) = 0.806 * \ln(MW_{au}) + 0.298$

Equation 5.3 $abs(Q_{cryst}) = 1.3472 * (MW_{au})^{0.806}$

Equation 5.4 $\sigma_{ours} = 2.108 * \sigma_{miller} + 1.810$

Thus, a $\overline{Q}_{cryst}$ of 0.1 *e*/kDa corresponds to 1 (positive) charge per 5,240-6,231 Å^2 for a monomeric protein. Although the surface area involved in lattice contacts can be quite variable, the 'typical' lattice contact in a protein crystal buries 570 Å^2 (Janin and Rodier, 1995) or

approximately 4.2% of the surface (Valdar and Thornton, 2001), which are both much smaller values.

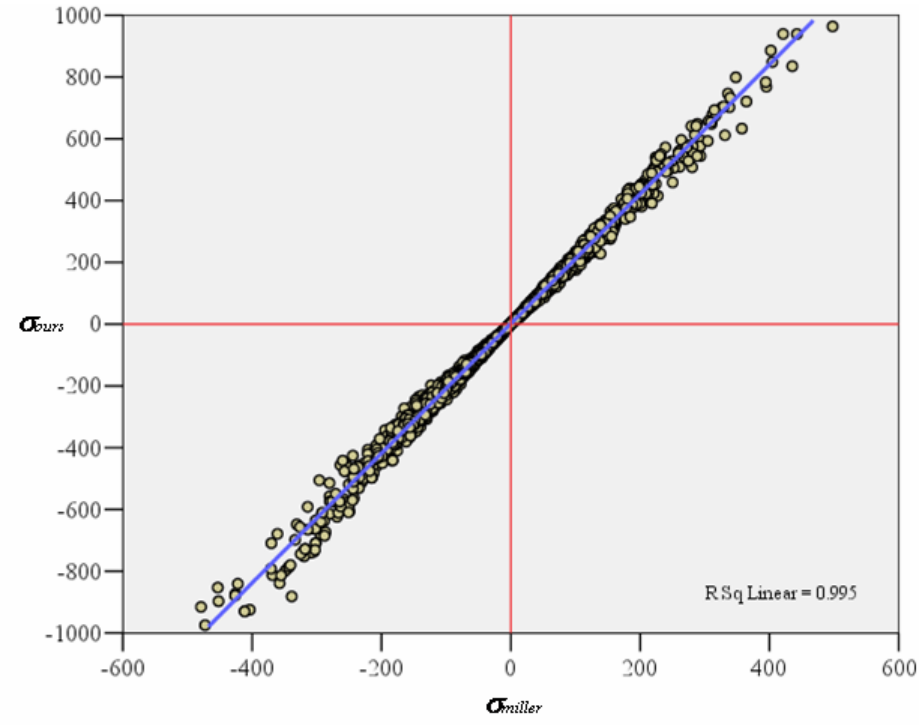


Figure 5.2 Plotting the σ_{miller} versus σ_{oursr} to determine the scale factor.

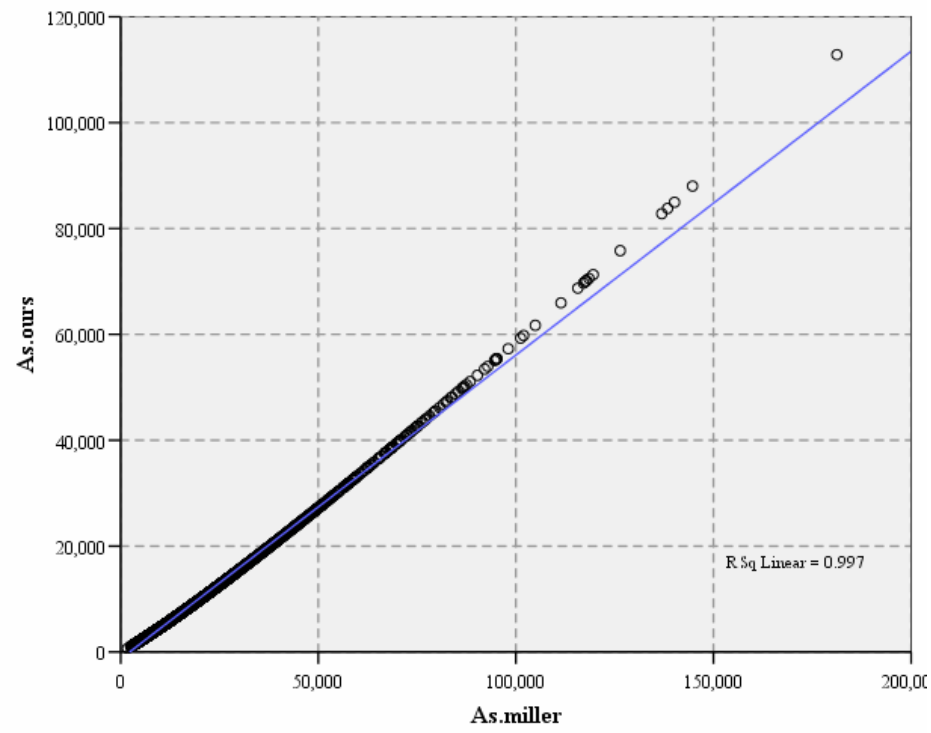


Figure 5.3 Our estimation of the A_S ($A_{S.ours}$) plotted against that of Miller et al. (1987a), $A_{S.millers}$.

5.2.2 Analysis of Variables

The analyses originally focused on the pH_{cryst} (*Primary Observable*) because it was the most widely reported experimental condition in the PDB and it is considered to be one of the most important crystallization variables (McPherson, 1999). The training set's pH_{cryst} frequencies (n = 4,114) and percentages are shown in Table 5.1 with the pH_{cryst} values rounded off to the tenths. Ninety-seven percent of the training set proteins had a pH_{cryst} within a pH range of 4.0-9.0. Most of the proteins, 98.5%, fell within reported buffer pH (4.0-9.5) of the crystallization screens listed in Appendix B. If this range is expanded another 0.5 pH units in each direction, pH 3.5-10.0, 99.3% of the training set proteins fell within this range. The histogram of the pH_{cryst} values used in this study is shown in Figure 5.4a, which was very similar to that found earlier in Chapter 3, using a smaller data set. From the figure, a saw-tooth pattern was observed at approximately every 0.5 pH unit from pH 4.5-9.0. Two additional spikes were also present at

pH values of 4.6 and 5.6. When these values were compared to the reported buffer pH values listed in 18 different screens from four commercial companies, an overlap was observed (Appendix B).

Table 5.1 The Training Set's pH_{cryst} values.

pH	Frequency	Percent
1.5	1	0.02
2.0	2	0.05
2.4	1	0.02
2.5	2	0.05
3.0	7	0.17
3.2	2	0.05
3.3	2	0.05
3.4	2	0.05
3.5	7	0.17
3.6	1	0.02
3.7	1	0.02
3.8	8	0.19
3.9	4	0.10
4.0	45	1.09
4.1	2	0.05
4.2	40	0.97
4.3	11	0.27
4.4	13	0.32
4.5	92	2.24
4.6	191	4.64
4.7	18	0.44
4.8	33	0.80
4.9	9	0.22
5.0	141	3.43
5.1	16	0.39
5.2	38	0.92
5.3	27	0.66

pH	Frequency	Percent
5.4	29	0.70
5.5	145	3.52
5.6	157	3.82
5.7	19	0.46
5.8	50	1.22
5.9	12	0.29
6.0	255	6.20
6.1	19	0.46
6.2	37	0.90
6.3	32	0.78
6.4	40	0.97
6.5	493	11.98
6.6	17	0.41
6.7	32	0.78
6.8	69	1.68
6.9	18	0.44
7.0	394	9.58
7.1	17	0.41
7.2	58	1.41
7.3	34	0.83
7.4	76	1.85
7.5	526	12.79
7.6	50	1.22
7.7	13	0.32
7.8	47	1.14
7.9	13	0.32
8.0	280	6.81

pH	Frequency	Percent
8.1	12	0.29
8.2	29	0.70
8.3	19	0.46
8.4	18	0.44
8.5	256	6.22
8.6	9	0.22
8.7	5	0.12
8.8	9	0.22
8.9	3	0.07
9.0	54	1.31
9.1	4	0.10
9.2	7	0.17
9.3	2	0.05
9.4	3	0.07
9.5	15	0.36
9.6	2	0.05
9.7	2	0.05
9.8	1	0.02
10.0	5	0.12
10.1	1	0.02
10.2	1	0.02
10.5	4	0.10
10.6	2	0.05
10.7	2	0.05
11.2	1	0.02
Total	4114	100.00

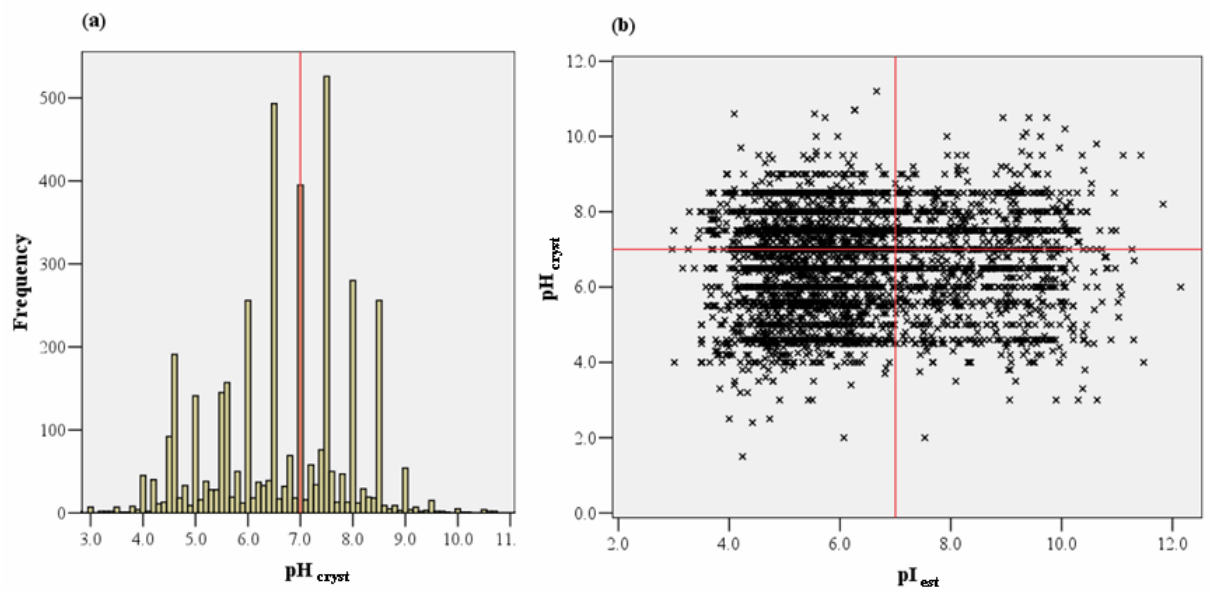


Figure 5.4 (a) The pH_{cryst} distribution of the training set proteins. (b) The scatter plot of the pI_{est} vs. pH_{cryst}.

This observed overlap has several interpretations, two of which are discussed here. The first interpretation would suggest that researchers may not be optimizing the crystals produced from the initial screens and that many crystals deposited within the PDB may be further improved (i.e., better resolution) by finely exploring the solution pH. The sparse matrix screens used by the commercial companies were originally designed to determine the initial conditions for nucleation, which would then be explored in more detail by further experiments (optimization). For example, the researcher may want to examine the pH every 0.05-0.1 units when optimizing conditions as suggested by McPherson (1995). A second possible interpretation would be that crystallization is relatively insensitive to pH. If the protein's ability to crystallize is relatively insensitive to a pH change of ±0.3 units, then there is no reason not to round to the nearest 0.5 pH unit. However, the nature of the pH influence on crystallization may be protein-specific. Some proteins crystallize over a wide range of pH (1-2 pH units); others crystallize under a narrow range (Cox and Weber, 1988). When initial crystals ('hits') are found, the slope of the estimated titration curve in the vicinity of the 'hit' should be examined. If the curve is steep, then a fine pH search may be mandatory. This would be especially important around the pK_a values of the charged amino acids (Table 4.2), where a small change in solution pH can have a large effect on a protein's charge, solubility, and ability to crystallize.

Next, the Spearman's correlations between the pH_{cryst} and the other variables were examined. While statistically significant ($p < 0.001$; Table 5.2), there was little correlation ($r = 0.059$) between the pI_{est} and pH_{cryst} (Figure 5.4b). This low correlation was similar to those previously reported by Page et al. (2003), $r < 0.01$, and Kantardjieff and Rupp (2004), $r < 0.10$. Due to the large number of proteins (4,114), even small correlations can become statistically significant. Similar statistically significant, but low correlations (Table 5.2) with the pH_{cryst} were also observed with the MW_{au} ($r = 0.088$; $p < 0.001$), A_S ($r = 0.088$; $p < 0.001$), and $diff_{lim}$ ($r = 0.036$; $p < 0.022$). As a control, a random number with a mean of 0.0 and standard deviation of 1.0 was generated for each protein and used for comparison of correlations. No correlation was observed between pH_{cryst} and the random number ($r = 0.013$; $p < 0.393$). While this lack of correlation with the random number was reassuring, we still believe that there is little information that can be exploited from the low, but statistically significant correlation (pH_{cryst} vs. pI_{est}) in the sense of guiding future crystallization efforts.

Due to the lack of correlation between the pH_{cryst} and pI_{est}, the analysis was expanded to include the protein's estimated net charge (Q_{cryst}), estimated specific charge ($\overline{Q}_{cryst}$), and estimated average surface charge density (σ_{cryst}) at the pH_{cryst}. The distributions of the Q_{cryst}, $\overline{Q}_{cryst}$, and σ_{cryst} for all training set proteins were all relatively normal and are shown in Figure 5.5b-d. The Q_{cryst} distribution was slightly negative with a mean of -4.9 and a standard deviation of 25.2. This slightly negative value may be due to the preponderance of proteins with a $pI_{est} < 7.0$ (Figure 5.5a). With a mean pH_{cryst} of 6.6, most of these proteins would be negatively charged to varying degrees. The protein's Q_{cryst} was also highly correlated with the pI_{est} ($r = 0.659$; $p < 0.001$) and pH_{cryst} ($r = -0.482$; $p < 0.001$; Table 5.2). However, it should be noted that the Q_{cryst} is calculated as a function of the pH_{cryst}, so this high correlation might be expected. The protein's MW_{au} had a stronger association with the Q_{cryst} ($r = -0.251$; $p < 0.001$) than did the pH_{cryst} ($r = 0.088$; $p < 0.001$). There was also a significant but weak correlation between the Q_{cryst} and $diff_{lim}$ ($r = -0.085$; $p < 0.001$). As the proteins increased in size, the Q_{cryst} and $diff_{lim}$ also increased. This was expected, because as the size increases, the number of charged amino

acids should increase, unless there is a significant change in the amino acid composition. With a greater number of charged amino acids, the protein can have a greater range of possible Q values. The correlation between the molecular weight and $diff_{lim}$ ($r = 0.323$; $p < 0.001$) is well known, with smaller proteins generally displaying better resolutions. These correlations were similar to those observed in the earlier analysis in Chapter 3 (Table 3.2).

Accounting for the size of the protein by dividing Q_{cryst} by MW_{au}, resulted in a weaker relationship between the $\overline{Q}_{cryst}$ (e/kDa) and MW_{au} ($r = -0.095$; $p < 0.001$) and $diff_{lim}$ ($r = -0.024$; $p < 0.118$; Table 5.2). Figure 5.5c showed that the protein's $\overline{Q}_{cryst}$ distribution was concentrated around a mean of zero (-0.1±0.4), the isoelectric point. Similar to the Q_{cryst}, a stronger correlation was also found between the pI_{est} and $\overline{Q}_{cryst}$ ($r = 0.746$; $p < 0.001$) than was found between the pI_{est} and pH_{cryst} ($r = 0.059$). This was not surprising given that a protein's charge can remain relatively constant over a wide range of pH values. Additionally, the $\overline{Q}_{cryst}$ was highly correlated with the pH_{cryst} ($r = -0.475$; $p < 0.001$).

Because of the known relationship between a protein's molecular weight and its accessible surface area, the estimated average surface charge density in me/nm^2 (σ_{cryst}) was also examined. The strength of other variable's correlations with the σ_{cryst} appeared to be in between those with the Q_{cryst} and $\overline{Q}_{cryst}$. For example, the correlations of the σ_{cryst} with pI_{est} ($r=0.737$; $p<0.001$) and MW_{au} ($r = -0.134$; $p < 0.001$) were between the correlations of the Q_{cryst} and $\overline{Q}_{cryst}$ with those same variables (Table 5.2). This finding was also observed for the correlation among crystallization parameters examined. The σ_{cryst} was more correlated with the Q_{cryst} ($r = 0.944$) than $\overline{Q}_{cryst}$ ($r = -0.912$). The correlation between the σ_{cryst} and $\overline{Q}_{cryst}$ was also very high ($r = 0.995$). This indicated that the $\overline{Q}_{cryst}$ was a good approximation of the mean surface charge density, σ_{cryst}. With only slight differences in the strength of their correlations, attempts were made to predict both a $\overline{Q}_{cryst}$ and σ_{cryst} interval for test set proteins.

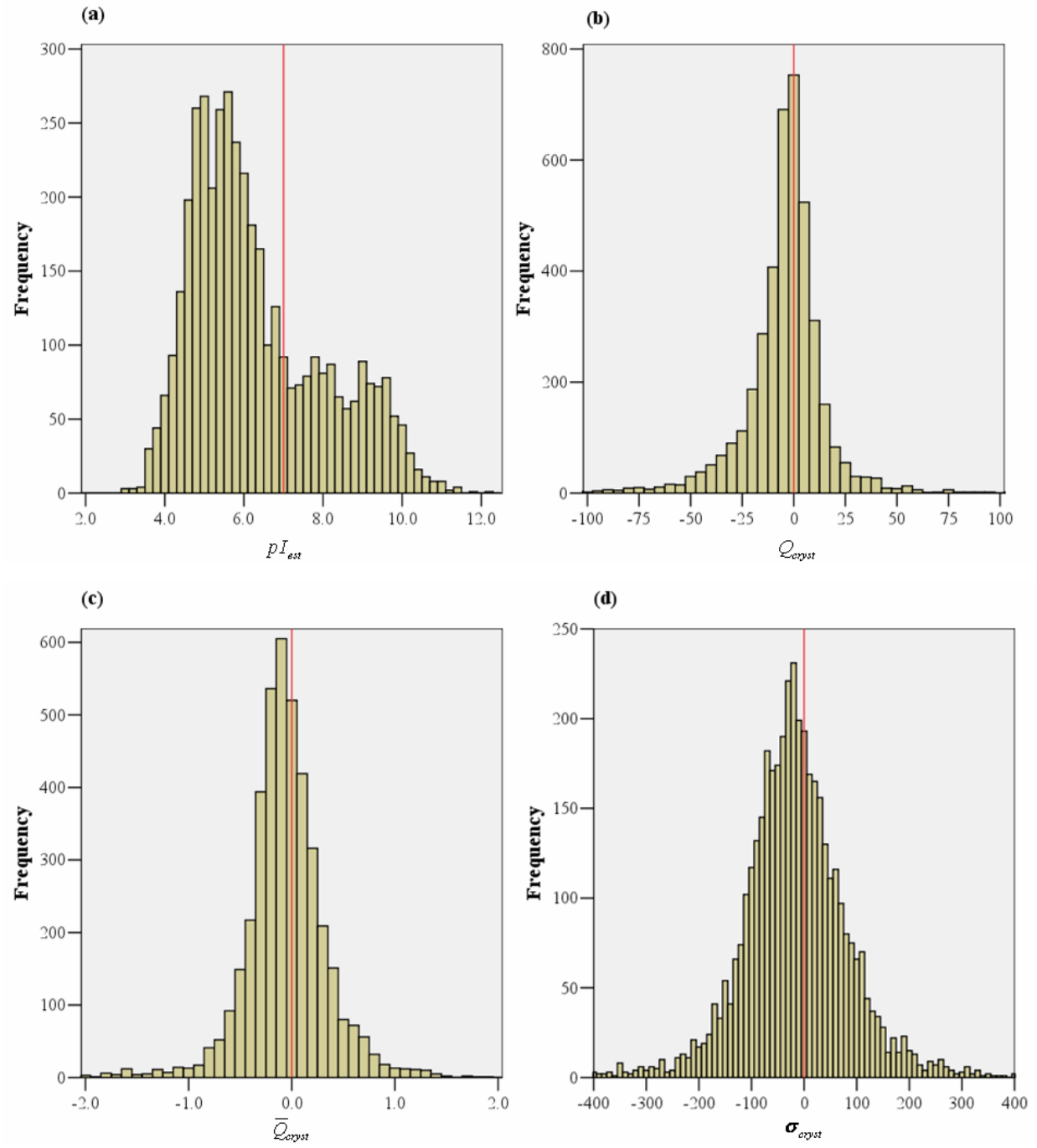


Figure 5.5 The distributions of (a) the pI_{est}, (b) Q_{cryst}, (c) $\overline{Q}_{cryst}$, and (d) σ_{cryst}.

These results have several possible explanations, two of which are discussed here. One is that these correlations might be expected, because the estimated net charge variables (pI_{est}, Q_{cryst}, $\overline{Q}_{cryst}$, and σ_{cryst}) are all calculated as a function of the solution pH, so the correlation to pH_{cryst}

might be expected. A second interpretation would be that while the pI_{est} and MW_{au} cannot predict the pH_{cryst} ranges for crystallization attempts, they might be able suggest values for $\overline{Q}_{cryst}$ or σ_{cryst} (proxy variables for the pH_{cryst}), which are determined by the amino acid composition and the pH of the mother liquor.

Table 5.2 Spearman's rho correlations among the training set's *Features* and *Controllables*.

Variable	pI_{est}	MW_{au}	A_S	pH_{cryst}	Q_{cryst}	$\overline{Q}_{cryst}$	σ_{cryst}	$diff_{lim}$	Random
pI_{est}	1.000	-	-	-	-	-	-	-	-
MW_{au}	-0.062**	1.000	-	-	-	-	-	-	-
A_S	-0.062**	1.000	1.000	-	-	-	-	-	-
pH_{cryst}	0.059**	0.088**	0.088**	1.000	-	-	-	-	-
Q_{cryst}	0.659**	-0.251**	-0.251**	-0.482**	1.000	-	-	-	-
$\overline{Q}_{cryst}$	0.746**	-0.095**	-0.095**	-0.475**	0.912**	1.000	-	-	-
σ_{cryst}	0.737**	-0.134**	-0.134**	-0.483**	0.944**	0.995**	1.000	-	-
$diff_{lim}$	0.006	0.323**	0.323**	0.036	-0.085**	-0.024	-0.038	1.000	-
Random	0.006	0.004	0.004	0.013	-0.005	-0.006	-0.006	0.025	1.000

** p-value < 0.001

The $\overline{Q}_{cryst}$ and σ_{cryst} distributions indicate that proteins appear to crystallize at low values of $\overline{Q}$ and σ. One problem with this observation is that "low" is a relative term with no frame of reference. One frame of reference is to compare the known σ values for nucleic acids and phospholipid bilayers, whose σ values are at least an order of magnitude greater than that of proteins. However, there are no pH values at which proteins are as highly charged as nucleic acids. Because nucleic acids readily crystallize, a high σ is not a barrier for crystallization. Another frame of reference is in relation to the mean $\overline{Q}$ or σ values at a "physiological pH" of 7.4 ($\overline{Q}_{pH=7.4}$). The mean pH_{cryst} is 6.6, which is 0.8 pH units less than "physiological pH" (Table 5.3). Therefore, the $\overline{Q}$ and σ values at the pH_{cryst} would be expected to be lower, i.e. less negative. Indeed they are, with a mean $\overline{Q}_{cryst}$ of -0.06 *e*/kDa, which is approximately three-

fold lower than the $\overline{Q}_{pH=7.4}$ mean (-0.17 *e*/kDa). The distributions (Figure 5.6) are significantly different as judged by a Student's t-test ($p<0.0001$). However, it should be noted that many proteins function in physiological compartments where the pH is significantly different. A more serious problem is that the standard deviations of the two distributions are more than twice the shift between them. The shift is statistically significant because of the sample size, but cannot be used to make meaningful predictions about specific proteins

Table 5.3 Comparing the $\overline{Q}_{cryst}$ and $\overline{Q}_{pH=7.4}$ values.

	Mean	SD	SE
$\overline{Q}_{cryst}$	-0.06	0.39	0.01
$\overline{Q}_{pH=7.4}$	-0.17	0.34	0.01

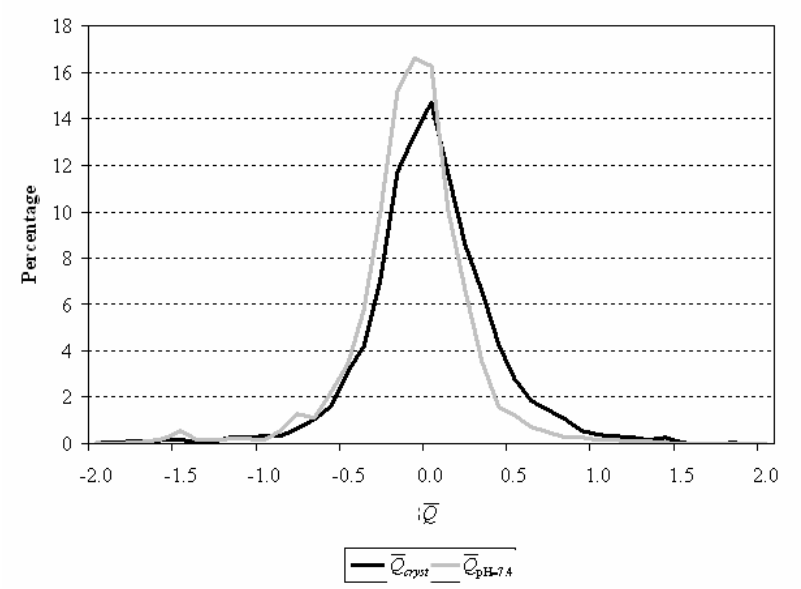


Figure 5.6 Comparing the $\overline{Q}_{cryst}$ and $\overline{Q}_{pH=7.4}$ distributions.

5.2.3 Analysis of the Test Set

Attempts were made to verify any findings on an independent test set of 1,246 proteins from more recent entries in the PDB (November 2004 thru November 2005). In order to remove potential bias, these proteins were filtered to maintain a low sequence similarity (BLAST p-value

$<10^{-80}$) to the proteins in the training set. First, the variable distributions of proteins between the training and test sets were compared (Table 5.4). The proteins were similar in all aspects, except molecular weight, between the two data sets. The proteins in the test set had a significantly larger MW_{au} distribution (65.0±69.0 kDa) than did those proteins in the training set (61.2±74.9 kDa). A similar effect was observed with A_S, which was estimated using the MW_{au}. Given the advances in technology, it might not be too surprising that the newer proteins were slightly larger than those present in the training set, which consisted of previously crystallized proteins from the inception of the PDB. After concluding that the two data sets were similar enough, especially in the crystallization *Observables*, methods were developed using the training set proteins to predict the $pH = pH_{cryst}$ ranges for the test set proteins, the probability that the *Primary Controllable* (solution pH) would equal the *Primary Observable* (pH_{cryst}).

Table 5.4 The mean and SD for all examined variables.

Data Set	n		pI_{est} [b]	MW_{au} [b]	A_S [b]	pH_{cryst} [a]	Q_{cryst} [b]	$\overline{Q}_{cryst}$ [b]	σ_{cryst} [b]	$diff_{lim}$ [a]	Random Number
Train	4,114	Mean	6.34	61.2	90.9	6.64	-4.88	-0.06	-17.2	2.02	0.02
		SD	1.69	74.9	80.6	1.30	25.19	0.39	101.5	0.44	1.00
Test	1,246	Mean	6.36	65.0	96.4	6.65	-2.89	-0.03	-10.3	2.03	0.02
		SD	1.63	69.0	76.2	1.36	28.19	0.36	98.9	0.42	0.97
KS		Z	1.07	2.78	2.78	0.71	2.03	1.16	1.14	0.89	0.490
Test		p<	0.206	0.0001	0.0001	0.689	0.001	0.138	0.151	0.410	0.972

[a] Extracted from the PDB

[b] Calculated using sequence information

5.2.4 Developing and Testing Models on the Independent Test Set

Each protein exhibits a unique titration curve based upon its amino acid sequence; therefore, these curves may result in protein-specific predictions. Some proteins are likely to be less affected by pH due to relatively long flat regions along the estimated titration curve in the most probable $\overline{Q}_{cryst}$ ranges. Even the identification of such proteins should aid in screen design. Alternatively, some predictions will narrow the pH search space to a narrow pH range (< 1 pH

unit). The three methods discussed in Section 4.8 for estimating the $P(pH = pH_{cryst} \mid \overline{Q} = \overline{Q}_{cryst}, PDB)$ or $P(pH = pH_{cryst} \mid \sigma = \sigma_{cryst}, PDB)$ over a pH range for the test set are examined here.

5.2.4.1 Confidence Interval

First, the fifty percent confidence intervals (CI_{50}) were calculated (Section 4.8.1) for all crystallization *Observables* examined, the pH_{cryst}, Q_{cryst}, $\overline{Q}_{cryst}$, and σ_{cryst}, for the training set proteins (Table 5.5). The CI_{50} for the pH_{cryst} suggested that a pH range of 5.6 to 7.5 should be searched for all proteins. Fifty-seven percent of the proteins in the training set were observed to crystallize within this pH range, while forty-nine percent of the test set proteins crystallized within this range. More than 50% of the training set proteins were present in the CI_{50} due to the erratic pH_{cryst} distribution. Approximately 46% of the commercial screens examined in Appendix B had a buffer pH value within this range (5.6-7.5). However, it should be noted that an additional 15% of the screens did not mention the solution pH. Searching pH values only in this range (5.6-7.5) would result in a narrow pH sparse matrix screen, offering little insight into the crystallization process and more than likely miss many of the proteins that are difficult to crystallize.

While the pH_{cryst} distribution offered little direct insight into the crystallization process, the solution pH exhibits its effects on the protein by controlling the protonation and deprotonation of the charged amino acid residues. Therefore, the estimated net charge of crystallized proteins was also examined. As mentioned previously, all the estimated net charge variable distributions (Q_{cryst}, $\overline{Q}_{cryst}$, and σ_{cryst}) were centered near zero, the isoelectric point (Figure 5.5b-d). To determine which variable performed best, the frequency of the test set proteins within the training set protein's CI_{50} was examined. While 50% of the training set proteins had a Q_{cryst} between –11.82 and +4.73, only 44.9% of the test set proteins Q_{cryst} values fell within this range. After attempting to account for the size of the protein by dividing the estimated net charge by the MW_{au}, 48.2% of the test set proteins had a $\overline{Q}_{cryst}$ value between the training sets CI_{50}, –0.25 and +0.14 *e*/kDa. Although there was a small difference between the

training (50.6%) and the test set's $\overline{Q}_{cryst}$ CI_{50} results (1.8%), it was felt that this was not significant given the sample size and round off effects. A slightly lower percentage of the test set proteins, 46.6%, had a σ_{cryst} that fell within the CI_{50} of the training set, -0.73 to +0.38 m*e*/nm^2. This slight decrease in the test set compared to the $\overline{Q}_{cryst}$ CI_{50} results may be due to errors in the calculation of A_S, which may increase in larger proteins. Unless, a more accurate calculation of A_S can be easily performed, it is thought that the $\overline{Q}_{cryst}$ values are more likely to yield better results. These results suggest that if a new protein was crystallized, there would be an approximately 50% chance that the pH_{cryst} would result in the protein having a $\overline{Q}_{cryst}$ value between –0.25 and +0.14 *e*/kDa. Assuming that the Q-related variables (*Hidden Observables*) can be used as proxy variables for the pH_{cryst} (*Primary Observable*), examining the net charge variables allows for a more detailed insight into the crystallization process than simply examining the solution pH values.

Table 5.5 The CI_{50} values and ranges for the *Observables*.

Variable	Low Value	High Value	Range	% of Training Set	% of Test Set	% Difference
pH_{cryst}	5.6	7.5	1.9	57.2	49.1	8.1
Q_{cryst}	-11.82	4.73	16.54	50.0	44.9	5.1
$\overline{Q}_{cryst}$	-0.25	0.14	0.39	50.0	48.2	1.8
σ_{cryst}	-73.07	37.96	111.03	50.0	46.6	3.4

5.2.4.2 Probability Distribution

The generation of probability distributions was discussed in Section 4.8.2. For this method a probability distribution was estimated based upon the relative frequency of proteins with a particular value for a *Given Observable*. This can be performed for any of the *Observables* examined, the pH_{cryst}, Q_{cryst}, $\overline{Q}_{cryst}$, or σ_{cryst}. As a reminder, any of the *Hidden Observables* can then be theoretically translated back into the pH search space (*Primary Controllable*). From the

probability distribution, the researcher can choose any threshold for selecting pH values to search for crystallization.

pH_{cryst}

If a 10% threshold is applied to the pH_{cryst} frequency distribution in Figure 5.7a, an initial pH range of 6.5-7.5 is suggested for all crystallization attempts. This would eliminate 44% of all screening conditions, not including those conditions where no pH value is listed. This would allow for a significant reduction in experimental conditions. However, this approach would likely miss the crystallization conditions of many proteins and is therefore not recommended.

Q_{cryst}

The Q_{cryst} relative frequency distribution was calculated every 5 *e* units from -60 *e* to +60 *e* (25 values). Assuming every Q value has an equal probability, a 'Random' value of 4% would result. Based upon the Q_{cryst} relative frequency distribution, a threshold of 10% was initially selected (Figure 5.7b). This would suggest a Q range of -5 *e* to +5 *e* for all proteins. Each protein's estimated titration curve is then examined to determine what pH values result in a Q of -5 *e* to +5 *e*.

$\overline{Q}_{cryst}$

The $\overline{Q}$ probabilities were calculated every 0.1 unit between -2.0 and +2.0 *e*/kDa by dividing the relative frequency of $\overline{Q}_{cryst}$ by the number of proteins in that $\overline{Q}$ group. The range from -2.0 to +2.0 *e*/kDa was chosen based on the observed $\overline{Q}_{cryst}$ relative frequency distributions for structures in the training set (Figure 5.7c). For this study, a threshold of 10% was used for high probability, 6-10% as average probability, and less than 6% as no better than a random pH selection. The 10% threshold value was arbitrarily chosen based on inspection of many probability distributions.

σ_{cryst}

The σ range chosen was -200 to +200 every 10 m*e*/nm^2 increments, which resulted in 41 values. Assuming all the σ values have an equal probability of succeeding in growing a crystal, a 0.024 (1/41) probability was used as 'Random.' There were no σ values that are greater than

10%. In fact, the highest relative probability of the 41 values was 5.6% (Figure 5.7d). Therefore, a cut off of 4% was chosen as the threshold. This would translate into any pH value that resulted in the protein having a σ value of -70 to +20 me/nm^2.

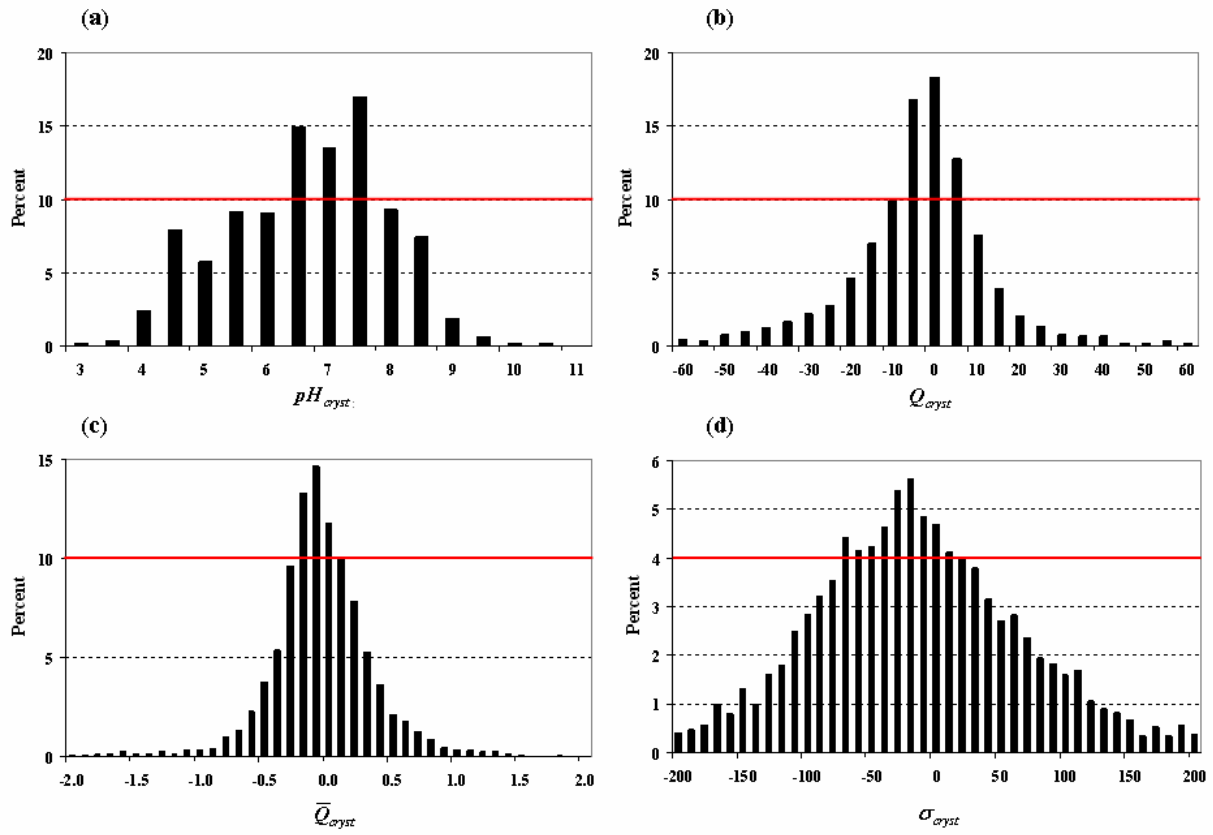


Figure 5.7 *Observable* distributions with arbitrary threshold probabilities.

For any of the *Observable* distributions, the researcher is free to choose any relative probability threshold cutoff. This information can then be translated back into the pH search space (*Primary Controllable*) by matching the Q values along the estimated titration curve (*Hidden Controllable*) with the relative probability distributions in Figure 5.6b as discussed in Section 4.8.2. Similarly, the $\overline{Q}$ or σ values along the specific charge or surface charge density curves can inherit the probabilities of the previously crystallized proteins in Figures 5.7c and 5.7d. The pH values with the highest probabilities can then be selected for the initial screen design.

5.2.4.3 Charge Range Test

Due to potential differences in CI_{50} ranges among different methods of grouping proteins by similarity, a third method was developed to compare models, the Charge Range Test (Section 4.8.3). This method examines the percentage of proteins that had crystallized within a given $\overline{Q}$ or σ range. The initial ranges were chosen based upon the group's mean and standard deviation (SD). Then the percentage of proteins within a given range of the mean value was calculated for both the training and test sets. For example, the $\overline{Q}$ values examined were ±0.1 to ±0.3 *e*/kDa of the mean and ±10 to ±100 m*e*/nm^2 for σ values. The resulting model with the most test set proteins within a given *Controllable* range has the highest accuracy. Thus, the *Observable* variable with highest accuracy should be used in future analyses. The results are shown in Table 5.6.

From this table, it can be observed that approximately 40% of the training set proteins crystallized within ±0.1 of the mean value. A slightly lower percentage of test set proteins crystallized within the same range. When a slightly larger range (±0.2) was used, an increase of ~20% was observed in both the training and test sets. Finally, when an $\overline{Q}$ range of ±0.3 was used, 72% of the tests set proteins were captured within this range, an increase of about 12% over the ±0.2 results.

When the σ values were examined, a much smaller percentage of proteins were found within the lower ranges examined, < 40%. Once the Mean±40 was reached, similar percentages of proteins to the $\overline{Q}$ ranges were observed. It should be noted that the σ SD for all training set proteins was ±102. An increase of 5-10% is observed for every ±10 m*e*/nm^2 , with a gradual reduction in the number of proteins captured within the ranges. Because of possible errors in the estimation of σ, it was felt that all future Charge Range Test results would focus on using the specific charge ($\overline{Q}$) as the measure of charge. Once $\overline{Q}$ values are determined, these values can be converted to σ by estimating the A_S as was done here.

Table 5.6 The Baseline Charge Range Test results for the $\overline{Q}_{cryst}$ and σ_{cryst}.

$\overline{Q}$			σ		
$\overline{Q}$ Range	Training Set	Test Set	σ Range	Training Set	Test Set
Mean ±0.1	40.4	37.5	Mean ±10	11.1	9.5
Mean ±0.2	60.1	59.6	Mean ±20	20.6	17.7
Mean ±0.3	73.1	71.8	Mean ±30	29.7	27.0
Mean ±0.4	81.8	81.1	Mean ±40	38.1	35.8
			Mean ±50	45.9	41.4
			Mean ±60	53.7	50.0
			Mean ±70	59.6	58.7
			Mean ±80	65.6	64.6
			Mean ±90	71.0	68.3
			Mean ±100	75.0	72.6

5.3 CHAPTER SUMMARY

In this chapter, a specific application of the PSPE framework (Chapter 4) was presented and applied to the Protein Data Bank (PDB). This application was meant to gain insight into the solution pH, the estimated specific charge ($\overline{Q}$), and the estimated average surface charge density (σ) required for the crystallization of proteins. For this analysis, the *Primary Observable* (pH_{cryst}) was obtained from the PDB and the *Hidden Observables* ($\overline{Q}_{cryst}$ and σ_{cryst}) were calculated using the *Primary Observable* and *Primary Feature* (amino acid sequence) for a non-redundant set of PDB proteins. While the pH_{cryst} was not statistically correlated to any protein *Feature*, the *Hidden Observables* were. However, uncertainty remains, because the *Hidden Observables* ($\overline{Q}_{cryst}$ and σ_{cryst}) are calculated from both protein *Features* and the *Primary Observable* (pH_{cryst}). Therefore, the high correlations among variables may be expected. Thus,

the hypothesis that the Protein Sequence-Properties Evaluation Framework can be used to frame and test hypotheses in-silico about variables that are believed to be important in crystallization appears inconclusive.

Based upon the observation that the Q_{cryst} appeared to be a possible proxy variable for the pH_{cryst} (Chapter 3), it was hypothesized that a protein's $\overline{Q}_{cryst}$ or σ_{cryst} values could be used as proxy variables for the pH_{cryst} and thus used for identifying solution pH ranges (*Primary Controllable*) with a higher probability of generating crystals. Regardless, a key observation was that the $\overline{Q}$ and σ distributions were low centered on zero. Also, there was a statistically significant, but weak correlation between the pI_{est} and pH_{cryst}, which was similar to previous studies. However, much stronger correlations were observed between the other *Hidden Observables* and the *Features*. Of particular interest were the *Feature's* correlations with the $\overline{Q}_{cryst}$ and σ_{cryst}. These two *Hidden Observables* were much more correlated to the pI_{est}, another charge derived variable, while removing much of the correlation to the MW_{au} that the Q_{cryst} displayed. This observation has two possible interpretations. The first is that although many statistically significant correlations among the Q-related quantities were noted, no evidence could be developed to suggest they were anything other than those expected from the additional information introduced with the hidden variables (*Features* and *Observables*). The second interpretation would be that the pI_{est} might suggest estimated charge values, whether the Q_{cryst}, $\overline{Q}_{cryst}$, or σ_{cryst}, that have a higher probability in generating crystals, which can then be translated back into pH space. However, the second interpretation cannot be proven at this time.

Based on the second hypothesis, three methods were examined to use previously crystallized proteins to suggest a pH_{cryst} range for target proteins using the target protein's $\overline{Q}$ curve. All methods were tested on an independent test set of 1,246 proteins. The first method calculated the middle 50% confidence interval (CI_{50}) of all *Observables* (pH_{cryst}, Q_{cryst}, $\overline{Q}_{cryst}$, and σ_{cryst}) and examined the frequency of the test set proteins that fell within that range. In theory, 50% of the test proteins should also fall within that range. Indeed, all of the *Observables* capture approximately 50% of the test cases. The smallest difference observed between the number of proteins within the *Observable's* CI_{50} range for the training and test sets of proteins

was with $\overline{Q}_{cryst}$ followed by σ_{cryst}. Therefore, it was felt that these estimates of $\overline{Q}$ and σ could be used to estimate the probability of success over the pH ranges.

The next method used the relative frequencies of the *Observables* to calculate a pseudo-probability distribution. A researcher can then choose a desired threshold for selecting initial solution pH ranges. While the pH_{cryst} probabilities based upon the relative frequencies result in the same initial pH values for all proteins, the values of the *Hidden Observables* translate into more specific pH conditions. This is accomplished using the protein's $\overline{Q}$ or σ curve and matching the probabilities to the pH values that result in the particular $\overline{Q} = \overline{Q}_{cryst}$ or $\sigma = \sigma_{cryst}$ value along the curve. Although the same $\overline{Q}$ and σ probability thresholds are used for each protein, the translation of these *Hidden Observables* into the pH search space (*Primary Controllable*) may result in quite different pH ranges for each protein.

For the final method, the Charge Range Test was devised to be a hybrid of the previous two methods. The researcher simply uses the mean $\overline{Q}_{cryst}$ value ±0.1, ±0.2, or ±0.3 *e*/kDa to select the $\overline{Q}_{cryst}$ ranges for initial crystallization attempts. This method allows for a more fair comparison among methods, which may have quite variable CI_{50} ranges. Similar to the previous method, a relative probability could be estimated for a protein crystallizing in this range based upon the relative frequency distribution of the proteins for that particular $\overline{Q} = \overline{Q}_{cryst}$ or $\sigma = \sigma_{cryst}$ range.

6.0 GROUPING PROTEINS BY SIMILARITY

In Chapter 4, the Protein Sequence-Properties Evaluation (PSPE) Framework was developed to examine differences in protein sequence derived properties (*Primary* or *Hidden Features*) or experimental variables (*Primary* or *Hidden Observables*) between two or more groups of proteins. The PSPE Framework was implemented in Chapter 5 to determine the relative importance of the estimated specific charge ($\overline{Q}_{cryst}$) and estimated average surface charge density (σ_{cryst}) in protein crystallization. Two possible interpretations were derived. The first interpretation was that no direct evidence was obtained to suggest the observed correlations among these variables was anything other than those expected, because these *Hidden Observables* ($\overline{Q}_{cryst}$ and σ_{cryst}) were calculated from the protein *Features and* the *Primary Observable* (pH_{cryst}). The second interpretation was that the *Hidden Observables* ($\overline{Q}_{cryst}$ and σ_{cryst}) may the potential to be utilized for initial crystallization screen design, particularly in the selection of the solution pH ranges (*Primary Controllable*) with a higher likelihood of success (crystals). This alternative hypothesis is used for this chapter. Therefore, the $\overline{Q}_{cryst}$ or σ_{cryst} became the *Observables* of interest and other protein *Features* were examined in their ability to predict these two variables. In this chapter, several methods, including binning and unsupervised clustering, were used to group proteins by similarity. It was hypothesized that there are groups of 'similar' proteins that crystallize under similar conditions. Similarity was determined by using the values of a single variable (binning by MW_{au}, or, pI_{est}) or a *Feature* vector ($\overline{Q}$ or σ curves).

Binning by the *Hidden Features* was the simplest approach to grouping proteins by similarity and was examined first, (Section 6.1). After grouping protein structures, differences in distributions were examined for the *Observables*, pH_{cryst}, Q_{cryst}, $\overline{Q}_{cryst}$, and σ_{cryst}. If significant

differences were observed between the distributions of experimental variables, the groups were further examined for pairwise differences. Attempts were then made to predict the *Observable* ranges from the test set, using the CI_{50} Test (Section 6.1.2) and the Charge Range Test (Section 6.1.3). After binning was used to separate proteins into groups by a *Hidden Feature* (MW_{au} or pI_{est}), two unsupervised clustering techniques were then examined to determine their ability to separate the proteins into groups by using a *Feature* vector, the $\overline{Q}$ or σ curves (Section 6.2). Finally, the distributions of the *Hidden Observables* appeared relatively normal in Chapter 5, so attempts were made in Section 6.3 to model the $\overline{Q}$ distributions with Gaussians.

6.1 BINNING

Binning involved selecting a single protein sequence *Feature* and determining cutoff points for group separation. Two main features frequently used to describe proteins are the molecular weight (MW) and pI_{est}. Additionally, these *Features* are thought to play an important role in the crystallization process. Therefore, initial attempts focused on separating proteins based upon their MW_{au} or pI_{est}. The number of cut points could be arbitrary (MW_{au}) or based upon some limited domain knowledge (pI_{est}). As a control, the proteins were also randomly grouped.

The first binning technique used the protein's size, MW_{au}, to group proteins. To determine the cutoff points for group separation, the distribution was first examined. The distribution of MW_{au} was positively skewed with no obvious points of separation (Figure 6.1a). The MW_{au} values were therefore transformed by taking the natural log (ln) of the MW_{au} values, $\ln(MW_{au})$, in order to make the distribution appear more symmetric (Figure 6.1b). Similar to the Preliminary Results in Section 3.2.1.1, the proteins within one standard deviation (SD) of the mean $\ln(MW_{au})$ were labeled as 'Average' (n = 2758). All proteins less than 1 SD below the mean $\ln(MW_{au})$ were labeled 'Small' (n = 689), while proteins greater than 1 SD above the mean were labeled 'Large' (n = 667).

Another *Feature* typically used to describe a protein is its pI_{est}, the solution pH where a protein exhibits an Q of zero. As mentioned previously, researchers had speculated a high

correlation between a protein's pI_{est} and its pH_{cryst}. It was expected that proteins would have a higher probability of crystallizing at or near their pI_{est} for reasons already discussed.

The pI_{est} distribution for all proteins is shown in Figure 6.1c. Similar to other studies, which examined the proteomes of various organisms from Archaea bacteria, Bacteria, and Eukaryotes (Urquhart et al., 1998; Van Bogelen et al., 1999; Schwartz et al. 2001), a bimodal distribution was observed for pI_{est}, where a much greater percentage of 'Acidic' ($pI \leq 6$) proteins than either 'Basic' ($pI \geq 8$) or 'Neutral' ($6 < pI < 8$) proteins was observed. In order to examine the effects of pI_{est} differences among proteins, the pI_{est} was first discretized. Unlike molecular weight, some knowledge could be applied to the separation of proteins by pI_{est}. Ries-Kautt and Ducruix's (1999) descriptions of 'Acidic' ($pI < 6$), 'Basic' ($pI > 8$), and 'Neutral' ($6 < pI < 8$) were initially used in the Chapter 3. However, due to the large amount of proteins and the significant differences found earlier, a finer sampling of pI_{est} was used. Proteins with a $pI_{est} \leq 5$ were labeled as 'Very Acidic' (n = 976), $5 < pI_{est} \leq 6$ labeled as 'Acidic' (n = 1188), $8 \leq pI_{est} < 9$ as 'Basic' (n = 564), and those $pI_{est} \geq 9$ as 'Very Basic' (n = 448). All remaining proteins with a pI_{est} between 6 and 8 were labeled as 'Neutral' (n = 938).

The third and final binning technique examined random groups. For this method, all the proteins in the training set were randomly assigned to one of five groups. To do this, each protein was assigned a random number from a distribution with a mean of 0.0 and a standard deviation of one. The proteins were then rank ordered by the random number and divided into five equal groups.

Differences in the distribution of *Features* (pI_{est}, MW_{au}, and A_S), *Observables* (pH_{cryst}, Q_{cryst}, $\overline{Q}_{cryst}$, σ_{cryst}, and $diff_{lim}$), and the random variable were examined within each binning method, using a Kruskal-Wallis (KW) test with an α of 0.01 (Section 4.10). The hypothesis was that proteins with different MW or pI_{est} would behave differently in crystallization attempts. In particular, larger proteins would crystallize over a broader range of charge, because of the greater range of possible Q values. It was also hypothesized that binning by pI_{est}, would increase the accuracy of the modeling methods on the test set, because each group would have its own $\overline{Q}_{cryst}$ or σ_{cryst} range.

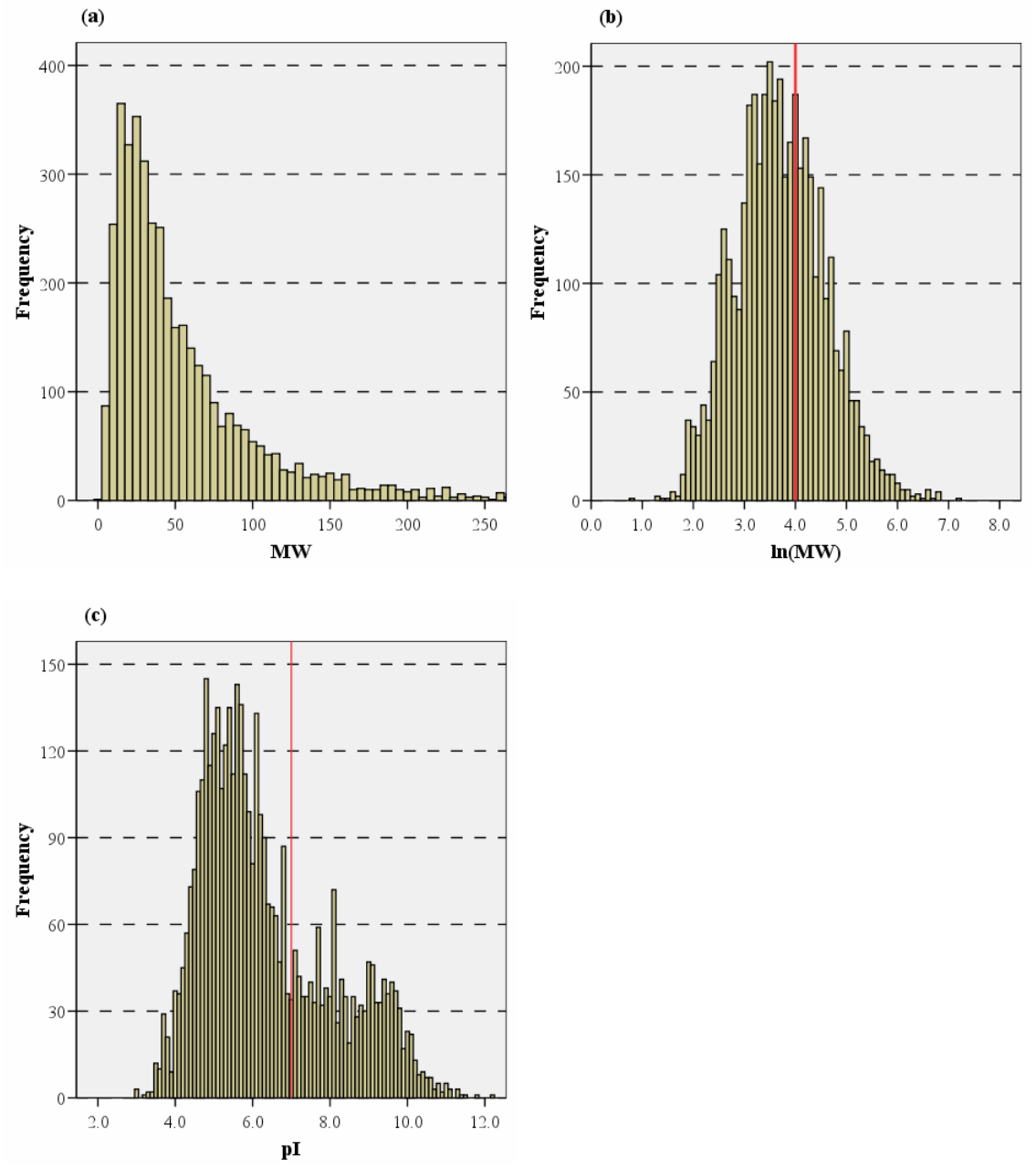


Figure 6.1 The Baseline distributions of the (a) MW_{au}, (b) $\ln(MW_{au})$, and (c) pI_{est}.

6.1.1 Variable Distributions

6.1.1.1 Molecular Weight Bins

A KW test was used to determine if the three ln(MW_{au}) groups had significantly different (p <0.01) distributions of *Observables* (pH_{cryst}, Q_{cryst}, $\overline{Q}_{cryst}$, σ_{cryst}, and $diff_{lim}$) and *Features*, pI_{est} or ln(MW_{au}). No significant difference was observed for either the pI_{est} or random number distributions (Table 6.1; Figure 6.2a). However, significantly different distributions were observed for the ln(MW_{au}), $diff_{lim}$, pH_{cryst}, Q_{cryst}, $\overline{Q}_{cryst}$, and σ_{cryst} (Figure 6.2b-g). To find out where the individual differences were in the distributions, a Kolmogorov-Smirnov (KS) test was performed on the groups pairwise using an α of 0.01. The 'Small' proteins had a significantly lower pH_{cryst} distribution, than did the 'Average' or 'Large' groups. In particular, there were a greater percentage of 'Small' proteins that crystallized with a pH between 4.0-4.5. The larger two groups of proteins had a greater percentage of proteins that crystallized between a pH of 7.0-8.0.

As the proteins grew larger, their Q_{cryst} distribution grew wider, but was still centered on a Q of zero electron units of charge (e). The 'Small' MW_{au} proteins had a Q_{cryst} distribution that was more tightly centered on zero (Figure 6.2c). The mean Q_{cryst} for 'Small' proteins was slightly negative, -0.4±6.8 electron units of charge (e). The 'Average' proteins were slightly more negative than the 'Small' proteins at -3.1±15.6 e, while the 'Large' proteins were much more negatively charged, -17.0±51.8 e (Table 6.1). This was not surprising given that larger proteins can have more charged amino acids, thus giving them a greater possible charge range. Therefore, attempts were made to take account for the size of the proteins by dividing the Q_{cryst} by either the protein's MW_{au} or A_S to obtain $\overline{Q}_{cryst}$ and σ_{cryst}, respectively.

Previous research has shown that a protein's surface area is proportional to its MW (Chothia, 1975; Teller, 1976; Miller et al., 1987ab; Janin et al., 1988). While there may be more charged residues on the surface of larger proteins, they should have approximately the same average surface charge density as smaller proteins. While having widely different Q_{cryst} distributions, the $\overline{Q}_{cryst}$ (Figure 6.2b) distributions were much more similar to each other, varying in degrees of kurtosis. However, all three ln(MW_{au}) groups were still significantly different from

each other, although less so. This was also demonstrated by all size groups displaying a similar mean $\overline{Q}_{cryst}$ and a low but still significant Spearman's correlation between the MW_{au} and $\overline{Q}_{cryst}$, r = -0.095, p <0.0001. The σ_{cryst} distributions were more different than the $\overline{Q}_{cryst}$ distributions (Figure 6.2g), but still much less different than the Q_{cryst} distributions. This may be due to differences in the calculation of A_S for 'Small' versus 'Large' proteins. It was expected that there would be larger errors in the estimation of A_S for 'Large' proteins, because of more complex shapes. Therefore, it was felt that the $\overline{Q}_{cryst}$ may result in a better estimation of surface charge density.

Additional significant differences were observed in the $\ln(MW_{au})$ and $diff_{lim}$ distributions (Figure 6.2e-f). The 'Small' proteins had a significantly better (i.e. lower) $diff_{lim}$ than did both the 'Average' and 'Large' proteins. Similarly, the 'Average' proteins had a better $diff_{lim}$ than did the 'Large' proteins. This was also expected, because there was a moderate correlation between the MW_{au} and $diff_{lim}$ in the training set, r = 0.323, p <0.0001. The $\ln(MW_{au})$ was used to separate proteins; therefore, all groups had significantly different values.

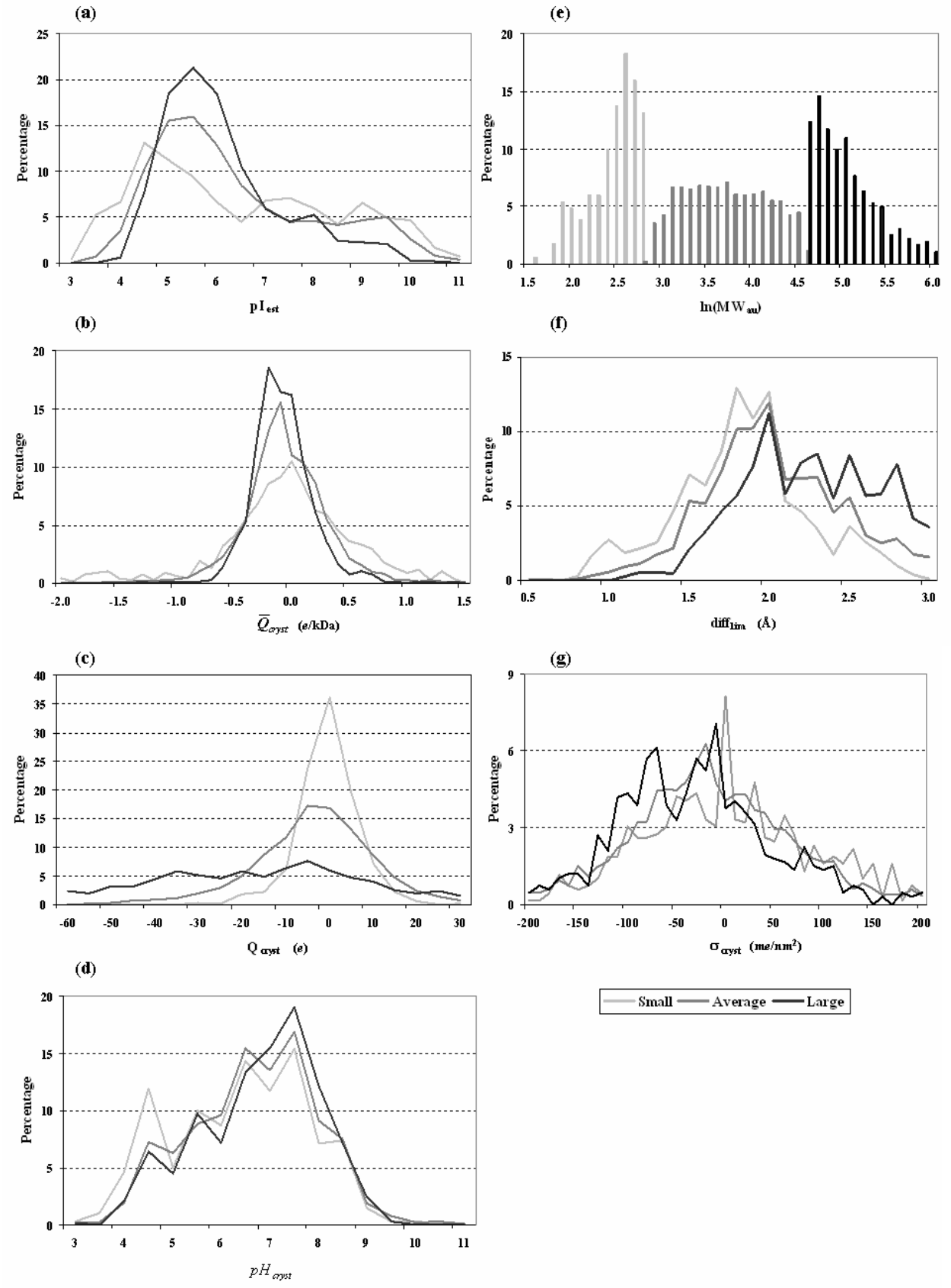

(a)
Percentage
pI est
(b)
Percentage
$\overline{Q}_{cryst}$ (e/kDa)
(c)
Percentage
Q cryst (e)
(d)
Percentage
pH_{cryst}
(e)
Percentage
ln(MW su)
(f)
Percentage
diff lim (Å)
(g)
Percentage
σ cryst (me/nm²)
Small
Average
Large

Figure 6.2: The distribution of (a) pI_{est}, (b) $\overline{Q}_{cryst}$, (c) Q_{cryst}, (d) pH_{cryst}, (e) $\ln(MW_{au})$, (f) $diff_{lim,}$ and (g) σ_{cryst} for the MW_{au} Bins.

Table 6.1 The mean and SD of the training set proteins separated by their (a) $\ln(MW_{au})$, (b) pI_{est}, or (c) random group numbers.

(a)

$\ln(MW_{au})$ Group	n		*Features*		*Observables*					Random Number
			pI_{est} [b]	$\ln(MW_{au})$ [b]	σ_{cryst}	$\overline{Q}_{cryst}$ [b]	Q_{cryst} [b]	pH_{cryst} [a]	$diff_{lim}$ [a]	
Small	689	Mean	6.5 [A]	2.45 [A]	-9.1 [A]	0.0 [A]	-0.4 [A]	6.4 [A]	1.83 [A]	0.01
		SD	2.0	0.30	125.2	0.6	6.8	1.4	0.43	0.98
Average	2,758	Mean	6.4 [A]	3.69 [B]	-15.9 [B]	-0.1 [B]	-3.1 [B]	6.7 [B]	2.02 [B]	0.02
		SD	1.7	0.46	96.1	0.4	15.6	1.3	0.42	1.00
Large	667	Mean	6.1 [A]	5.07 [C]	-30.6 [C]	-0.1 [C]	-17.0 [C]	6.8 [B]	2.25 [C]	0.04
		SD	1.3	0.45	94.4	0.3	51.8	1.2	0.42	0.99
KW		x^2	3.56	2837.0	36.1	13.1	167.8	28.7	290.7	0.43
Test		$p<$	0.169	0.0001	0.0001	0.001	0.0001	0.0001	0.0001	0.805
Total	4,114	Mean	6.3	3.71	-17.2	-0.1	-4.9	6.6	2.02	0.02
		SD	1.7	0.87	101.5	0.4	25.2	1.3	0.44	1.00

(b)

pI_{est} Bin	n		*Features*		*Observables*					Random Number
			pI_{est} [b]	MW_{au} [b]	σ_{cryst}	$\overline{Q}_{cryst}$ [b]	Q_{cryst} [b]	pH_{cryst} [a]	$diff_{lim}$ [a]	
Very Acidic	976	Mean	4.5 [A]	56.3 [A]	-108.0 [A]	-0.4 [A]	-19.3 [A]	6.5 [A]	2.0 [A]	0.04
		SD	0.4	83.5	99.8	0.4	29.3	1.4	0.4	1.02
Acidic	1,188	Mean	5.5 [B]	74.2 [B]	-39.8 [B]	-0.1 [B]	-10.3 [B]	6.7 [B]	2.1 [B]	0.00
		SD	0.3	80.2	63.5	0.2	22.5	1.2	0.5	1.00
Neutral	938	Mean	6.6 [C]	661 [C]	5.7 [C]	0.0 [C]	0.3 [C]	6.7 [B]	2.1 [BC]	0.00
		SD	0.4	72.2	61.9	0.2	20.0	1.2	0.4	0.97
Basic	564	Mean	8.2 [D]	52.9 [A]	48.3 [D]	0.2 [D]	7.8 [D]	6.7 [B]	2.0 [AC]	0.06
		SD	0.4	64.7	60.8	0.3	14.4	1.2	0.4	0.98
Very Basic	448	Mean	9.7 [E]	37.8 [D]	110.5 [E]	0.4 [E]	14.7 [E]	6.7 [AB]	2.0 [AC]	0.05
		SD	0.5	44.1	86.4	0.3	19.3	1.4	0.4	1.02
KW Test [c]		x^2	3894	236.7	2128	2170	1784	20.0	27.6	2.0
		$p<$	0.0001	0.0001	0.0001	0.0001	0.0001	0.0005	0.0001	0.744

(c)

Random Group	n		*Features*		*Observables*					Random Number
			pI_{est} [b]	MW_{au} [b]	σ_{cryst}	$\overline{Q}_{cryst}$ [b]	Q_{cryst} [b]	pH_{cryst} [a]	$diff_{lim}$ [a]	
RG 1	823	Mean	6.3	57.4	-15.69	-0.05	-3.9	6.6	1.99	-1.38 [A]
		SD	1.7	65.8	106.19	0.41	22.7	1.3	0.43	0.46
RG 2	823	Mean	6.3	66.0	-17.4	-0.06	-5.7	6.6	2.04	-0.51 [B]
		SD	1.7	87.8	102.43	0.39	30.1	1.3	0.46	0.17
RG 3	823	Mean	6.3	60.7	-17.14	-0.06	-4.8	6.6	2.01	0.03 [C]
		SD	1.6	74.4	97.601	0.38	23.4	1.3	0.44	0.14
RG 4	823	Mean	6.4	58.8	-19.12	-0.06	-5.4	6.7	2.02	0.54 [D]
		SD	1.7	71.8	101.11	0.40	23.3	1.3	0.44	0.17
RG 5	822	Mean	6.4	63.3	-16.41	-0.06	-4.6	6.6	2.05	1.42 [E]
		SD	1.7	72.9	100.21	0.38	25.7	1.3	0.44	0.47
KW	Test [c]	x^2	0.5	10.9	1.4	1.2	2.7	3.1	7.6	3947
		$p<$	0.970	0.028	0.853	0.877	0.602	0.545	0.106	0.0001

[a] Extracted from the PDB

[b] Calculated using asymmetric unit sequence information

[c] Degrees of freedom of 4

Note: Groups labeled with different letters (A, B, C, D, or E) have significantly different distributions ($p < 0.01$) as determined by a KS Test.

6.1.1.2 pI Bins

After the proteins were separated by their pI_{est}, their group frequencies were compared to those of the ln(MW) bins (Table 6.2). While all pI_{est} bins had approximately 67±3% in the ln(MW)='Average' group, the distribution of 'Large' and 'Small' proteins was not equally divided among the five pI_{est} bins. The proteins at either end of the pI_{est} range ('Very Acidic,' 'Basic,' and 'Very Basic') had a much greater percentage of 'Small' proteins than 'Large' proteins. The 'Neutral' proteins had a slightly greater percentage of 'Large' proteins, while the 'Acidic' proteins had a much greater percentage of 'Large' proteins. These differences may be due to the technology bias in isolating and purifying slightly acidic and neutral proteins (using more standardized methods).

Table 6.2 Cross-tabulation of the proteins separated using either binning by the pI_{est} or $\ln(MW_{au})$.

	$\ln(MW_{au})$ Bin				%
pI_{est} Bin	Small	Average	Large	Total	Total
Very Acidic	218 (22.3)	635 (65.1)	123 (12.6)	976 (100)	23.7
Acidic	114 (9.5)	815 (68.6)	259 (21.8)	1188 (100)	28.9
Neutral	130 (13.9)	627 (66.8)	181 (19.3)	938 (100)	22.8
Basic	118 (20.9)	368 (65.2)	78 (13.8)	564 (100)	13.7
Very Basic	109 (24.3)	313 (69.9)	26 (5.8)	448 (100)	10.9
Total	689	2758	667	4114	100.0
% Total	16.7	67.0	16.2	100.0	

Note: The percentage of proteins in each row is shown in parentheses.

Next, differences in the distribution of the *Observables* (pH_{cryst}, Q_{cryst}, $\overline{Q}_{cryst}$, and σ_{cryst}) were examined among the pI_{est} groups, using a KW test with an α of 0.01. When significant differences were found, a KS test was performed pairwise to determine where the individual differences occurred between groups. Similar to the $\ln(MW_{au})$ bins, there were some group differences in the pH_{cryst} distributions. The 'Very Acidic' proteins had a greater percentage of proteins that crystallized at a pH 4.0-4.5 than the other pI_{est} groups (Figure 6.3a). The remaining groups had no significant differences in their pH_{cryst} distributions. The pH_{cryst} distributions among the pI_{est} bins were spread wide over the pH range, generally between pH of 4 and 10 with varying degrees of smoothness. This broad range corresponded to the recommended range to search as suggested by other researchers (Gilliland, 1997; Farr et al., 1998; Kantardjieff and Rupp, 2004). All distributions had an average of 6.5-6.8, while exhibiting their peaks between a pH_{cryst} of 6.5-7.5.

The distribution of the Q_{cryst} among the five pI_{est} groups is shown in Figure 6.3c. The KW test indicated that there were statistically significant differences among the populations. The KS test demonstrated that all groups had significantly different pairwise Q_{cryst} distributions (Table 6.3). The 'Neutral' group of proteins was centered on a Q_{cryst} of zero, the pI_{est}. As the protein's pI_{est} deviated from 'Neutral', the distributions shifted. When the pI_{est} decreased, the

Q_{cryst} distribution became more negative, and when the pI_{est} increased, the Q_{cryst} distribution became more positive.

When the proteins were binned by their pI_{est}, a KW test indicated differences in $\overline{Q}_{cryst}$ distributions. The pairwise KS test demonstrated that all five $\overline{Q}_{cryst}$ distributions were significantly different from each other (Figure 6.3b). This was expected because the Q_{cryst} and $\overline{Q}_{cryst}$ variables were highly correlated, r = 0.912, p <0.0001. The probability of successful crystallization was higher for an 'Acidic' protein when the $\overline{Q}_{cryst}$ was slightly negative. The $\overline{Q}_{cryst}$ distribution shifted towards a more negative value for 'Very Acidic' proteins. 'Basic' proteins were more likely to crystallize when the $\overline{Q}_{cryst}$ was slightly positive, with 'Very Basic' proteins having a higher mean $\overline{Q}_{cryst}$. The 'Neutral' protein's $\overline{Q}_{cryst}$ distribution was centered on zero, the pI_{est}. Similar results were observed for the σ_{cryst} (Figure 6.3g). Again, because the proteins were separated by a Q-related variable, the pI_{est}, it might be expected that the distributions of the other Q-related *Observables* (Q_{cryst}, $\overline{Q}_{cryst}$, and σ_{cryst}) would therefore also be significantly different.

After demonstrating that binning proteins by their pI_{est} resulted in significantly different Q_{cryst}, $\overline{Q}_{cryst}$, and σ_{cryst} distributions, the distribution of the $diff_{lim}$ and MW_{au} were examined. The KW test indicated that there were significant differences in both distributions. Few pairwise differences were found in the distribution of the $diff_{lim}$ (Figure 6.3f). The 'Very Basic', 'Basic', and 'Very Acidic' proteins had $diff_{lim}$ distributions that were slightly lower than the 'Acidic' group. Additionally, the 'Very Basic' proteins were significantly smaller, averaging <40 kDa, than were the other four groups, which averaged 50+ kDa (Figure 6.4e). The 'Acidic' proteins were significantly larger than the other groups were. Additionally, the 'Neutral' proteins were the next largest group and significantly larger than the remaining three groups. In general, all of these distributions were very rough with several sharp points.

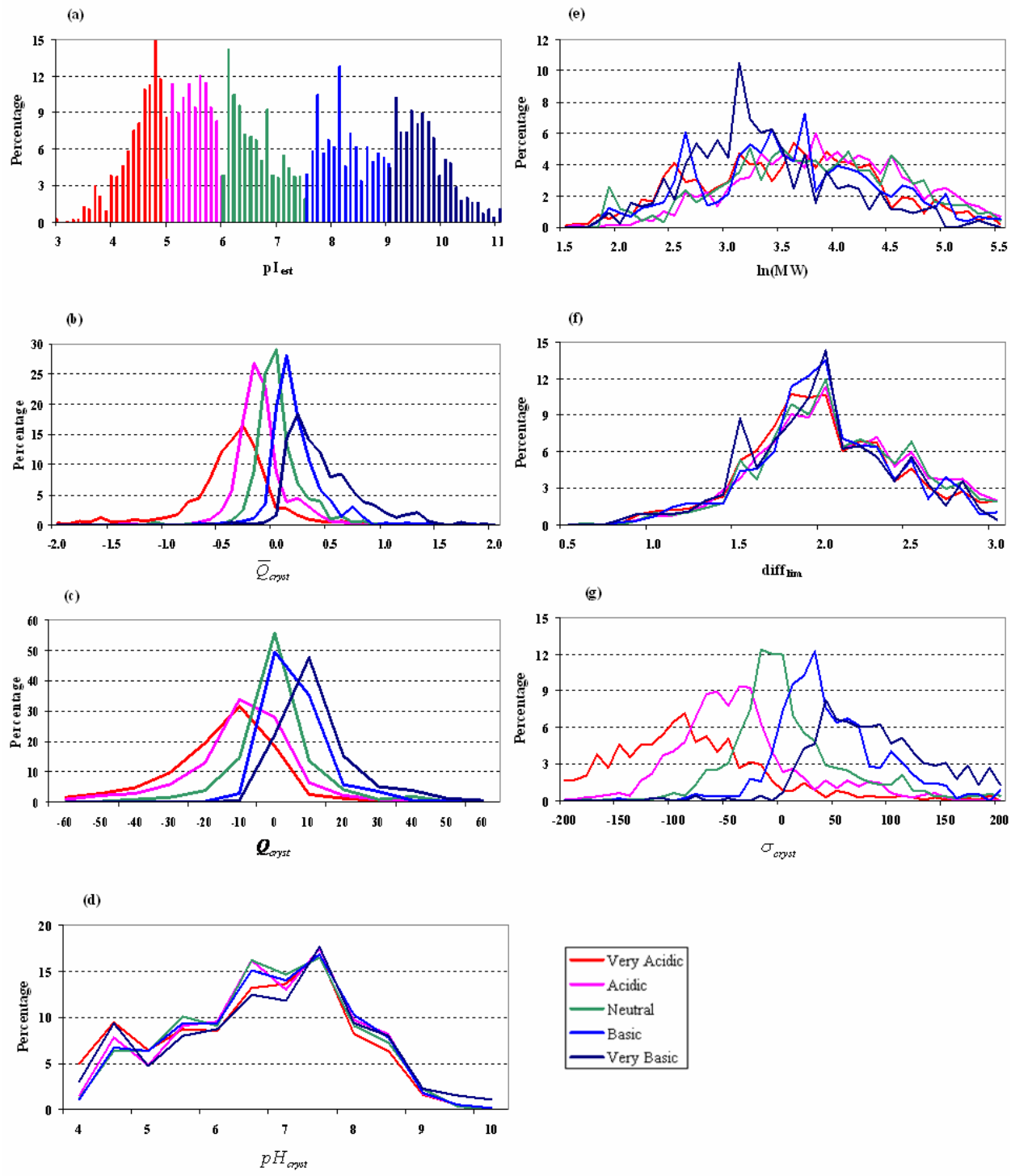


Figure 6.3 The distributions of the (a) pI_{est}, (b) $\overline{Q}_{cryst}$, (c) Q_{cryst}, (d) pH_{cryst}, (e) $\ln(MW_{au})$, (f) $diff_{lim}$, and (g) σ_{cryst} for the proteins binned by their pI_{est}.

6.1.1.3 Random Bins

As seen in Table 6.6, no variable besides the one used to form the groups (random number) had significantly different distributions as determined by a KW test. The distributions of the MW_{au} and $\ln(MW_{au})$ were approaching statistical significance, which partially justifies the use of an α of 0.01. After examining the distribution of variables among the five random groups, the ability of each group to predict the $\overline{Q}_{cryst}$ was examined.

After examining differences in protein *Features* and *Observables* between the $\ln(MW_{au})$, pI_{est}, and random groups, the ability of these groups to predict a range of $\overline{Q}$ values for crystallization was tested on the same independent test set used in Chapter 5. The same methods used in Chapter 5 were used here, the CI_{50} test (Section 6.1.2) and the Charge Range Test (Section 6.1.3). These results were then compared to the results in Chapter 5, where no subgroups were formed (Baseline), to determine whether separating proteins by molecular weight had a positive effect, i.e. increased predictive power.

6.1.2 CI_{50} Test

The $\overline{Q}$ interval that captured the middle 50% (CI_{50}) of the proteins in each $\ln(MW_{au})$ group is shown in Table 6.3. Approximately, 0.3-0.6 $\overline{Q}$ units were needed to capture the middle 50% of the proteins. The 'Large' and 'Average' proteins had tighter $\overline{Q}_{cryst}$ distributions (Figure 6.2b), thus displaying a smaller $\overline{Q}_{cryst}$ range to capture the middle 50%. The CI_{50} range for the 'Average' proteins was the same as Baseline in Chapter 5, -0.25 to 0.14 $\overline{Q}$ units. The overall accuracy as judged by the amount of test set proteins within the CI_{50}, was slightly lower than when modeled on all proteins, 46.6% versus 48.2%. The 'Small' test set proteins had an accuracy of 48.9% on the test set proteins, while the 'Large' proteins decreased to 45.5%. These findings indicated little to no improvement in $\overline{Q}_{cryst}$ predictions when separating proteins on the basis of their MW_{au}. However, it remained unclear whether molecular weight along with other variables could improve prediction.

Although the CI_{50} analysis gave an indication on how well the test set fits any model developed on the training set, it failed in not taking into account the different $\overline{Q}$ ranges required to obtain the CI_{50}. For example, the CI_{50} ranges for the MW_{au} bins (Table 6.3) were quite different among groups. The CI_{50} range for the 'Small' proteins (0.61) was 2x that of the 'Large' proteins (0.28). However, the number of proteins within the CI_{50} of the 'Small' proteins (49%) was greater than those in the 'Large' proteins, 45.5%. These results were misleading, giving the appearance that the 'Small' proteins captured the information better. Therefore, another procedure (Charge Range Test) was developed in which the number of proteins within a $\overline{Q}$ range of the group mean was examined in Section 6.1.3.

The results of the CI_{50} Test for proteins grouped by their pI_{est} are shown in Table 6.3. Approximately 50% of the training set proteins were found in the CI_{50} of both, the binning by pI_{est} and the Baseline group. However, the CI_{50} $\overline{Q} = \overline{Q}_{cryst}$ range, 0.34, was 0.13-0.15 larger than the 'Acidic', 'Neutral', and 'Basic' groups. When the test set proteins were examined, the 'Very Acidic' and 'Very Basic' bins had more proteins within the CI_{50} range than the other groups. However, when the CI_{50} ranges were examined, this test again proved misleading as a much smaller CI_{50} range was used for the 'Acidic,' 'Neutral,' and 'Basic' bins. Additionally, based on the cumulative results, it appeared that $\ln(MW_{au})$ binning outperformed pI_{est} binning, 46.6% versus 43.9%. Again, when the CI_{50} ranges are compared for the $\ln(MW_{au})$ and pI_{est} bins, there was often a larger range for the $\ln(MW_{au})$ bins. For these reasons, the CI_{50} test would not be examined in further clustering methods (Section 6.2).

The CI_{50} Test results from randomly grouping the proteins are shown in Table 6.3. All five groups had a similar low, high, and $\overline{Q}$ range, which was very similar to using all proteins as one group (Baseline; Chapter 5). Approximately 50% of the training set proteins were found in each group's CI_{50}, as expected. The amount of test set proteins within each random group's CI_{50} was slightly variable, ranging from 46-52%. Using the random groups also gave an indication of the error on the test set. This method was a pseudo-five fold cross-validation, which gave an error of ±2.2%. With a larger sample size, all groups may be less variable and more distributed around 50%. The cumulative average for the random groups, 48.6% was again very similar to that obtained with the Baseline group, 47.8%, and well within the 2.2% error identified with the random groups.

Table 6.3 The CI_{50} $\overline{Q}$ ranges for the protein separated by their MW_{au}, pI_{est}, and random number values along with the CI_{50} results.

MW_{au} Cluster	Training Set CI_{50} Range			% Proteins in CI_{50}	
	Low $\overline{Q}$	High $\overline{Q}$	$\overline{Q}$ Range	Training Set	Test Set
Small	-0.32	+0.29	0.61	50.1	48.9
Average	-0.25	+0.14	0.39	50.0	46.6
Large	-0.24	+0.04	0.28	50.2	45.5
MW Cum %				**50.1**	**46.6**
Very Acidic	-0.55	-0.21	0.34	50.0	52.5
Acidic	-0.25	-0.06	0.19	50.0	42.0
Neutral	-0.07	0.13	0.20	50.1	34.6
Basic	0.06	0.27	0.21	50.0	45.6
Very Basic	0.18	0.59	0.41	50.0	52.7
pI Sum Total				**50.0**	**43.9**
RG 1	-0.26	+0.16	0.42	50.2	51.8
RG 2	-0.24	+0.13	0.37	50.2	46.3
RG 3	-0.24	+0.15	0.39	50.2	48.1
RG 4	-0.26	+0.13	0.39	50.2	47.1
RG 5	-0.26	+0.13	0.40	50.1	49.7
RG Sum Total				**50.2**	**48.6±2.2**
Baseline	-0.25	+0.14	0.39	**50.0**	**47.8**

6.1.3 Charge Range Test

For this method, a $\overline{Q}$ range was selected around the group's mean $\overline{Q}_{cryst}$ value for the training set proteins and the amount of proteins within the given range was then examined on both the training and test set data. The widths of the ranges chosen were Mean±0.1, Mean±0.2, and Mean±0.3 based on the results in Section 5.2.4.3. Using all proteins as one group, these ranges

were able to capture approximately 40%, 60%, and 70% of both the training and test set proteins. By calculating the percentage of proteins within a given range, this method allowed for easy comparison between groups or methods of separating proteins.

When the overall Charge Range Test results of the $\ln(MW_{au})$ binning were compared to the results of Baseline in Chapter 5, no differences in the $\overline{Q}$ results were observed (Table 6.4). However, when each group was examined individually, some differences became more apparent (Figure 6.4). These differences seemed more in line with the actual group $\overline{Q}_{cryst}$ distributions (Figure 6.2b) than with the CI_{50} results. The group of 'Large' proteins performed the best on both the training and test sets. This was not to surprising given that the 'Large' proteins had the tightest $\overline{Q}_{cryst}$ distribution (Figure 6.2b). When a Mean±0.1 range was applied, 48.6% of the 'Large' proteins in the test set fell within the range. This was approximately 11% better than Baseline (37.5%). While the 'Large' proteins performed better, the 'Small' proteins performed much worse. At the same Mean±0.1 interval, only 28.3% of the 'Small' proteins were captured. While having a greater range of possible net charge values (Q_{cryst}; Figure 6.2c), the 'Large' proteins actually had a tighter estimated specific charge distribution ($\overline{Q}_{cryst}$). This observation was the opposite of the hypothesis that the large proteins would have a greater range of possible charge values for crystallization. Thus it appears that 'Small' proteins can be crystallized over a greater $\overline{Q}$ range than can the 'Large' proteins, that were more constrained to lower $\overline{Q}$ values, near the pI_{est}. A possible reason for this may be the possible electrostatic repulsion between large proteins when the $\overline{Q}$ or σ is higher. Another possible explanation would be that other molecules present in the solution, such as salts, might aid in dissipating/shielding the charges of the 'Small' molecules. However, this could not be examined with the current data set due to the unavailability of this information for many of the proteins in the PDB and would offer a possible future endeavor.

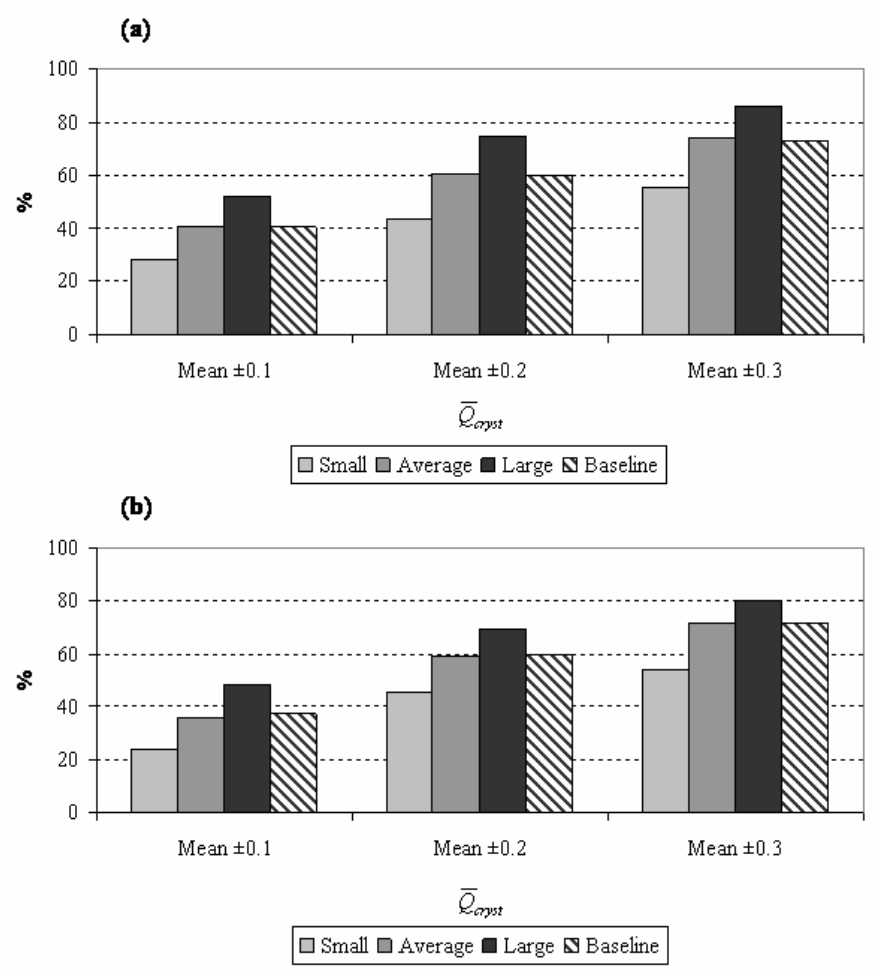


Figure 6.4 (a) The Charge Range Test results (a) between methods (cumulative percent) or (b) within grouping method, Baseline and ln(MW_{au}) binning for proteins in the test set.

The results of the Charge Range Test for the pI_{est} bins is shown in Table 6.5. When the amount of proteins whose $\overline{Q}_{cryst}$ fell within a given $\overline{Q}$ range were examined and compared to those obtained with Baseline, the pI_{est} binning method appeared to perform much better, contrary to the CI_{50} results. At every $\overline{Q}$ range for both data sets, binning by pI_{est} increased the cumulative percentage of proteins captured within the $\overline{Q}$ range by 12-15% (Figure 6.5a).

When each pI_{est} group was examined individually, some groups were more effective than others in capturing the $\overline{Q}_{cryst}$ of the test set (Figure 6.5b). Again, the effectiveness mimicked the calculated $\overline{Q}_{cryst}$ distributions of each group in Figure 6.3b. When examining the CI_{50} results, the

'Very Acidic' proteins in the test set were captured 58% of the time, which was a very good result. However, the CI_{50} $\overline{Q}$ range for 'Very Acidic' proteins was much larger than most of the other pI_{est} bins (Table 6.4). The percentages of proteins within ±0.1 to ±0.2 $\overline{Q}$ units of the mean $\overline{Q}_{cryst}$ value were 10-25% better than no binning for all groups except the 'Very Basic' group. The 'Very Basic' proteins had a lower accuracy than did the Baseline group until the Mean±0.3 $\overline{Q}$ unit width, where they were more accurate than the Baseline group.

Table 6.4 The Charge Range Test results for the proteins separated by their $\ln(MW_{au})$ and compared to Baseline.

Training Set N=4,114	Small N=689	Average N=2,758	Large N=667	Cum% MW_{au} Bins	Baseline
Mean ±0.1	28.3	40.4	52.3	40.3	40.4
Mean ±0.2	43.4	60.8	74.7	60.1	60.1
Mean ±0.3	55.7	74.2	86.2	73.1	73.1
Test Set N=1,246	Small N=139	Average N=887	Large N=220	Cum% MW_{au} Bins	Baseline
Mean ±0.1	23.7	35.9	48.6	36.8	37.5
Mean ±0.2	45.3	59.5	69.5	59.7	59.6
Mean ±0.3	54.0	71.8	80.0	71.3	71.8

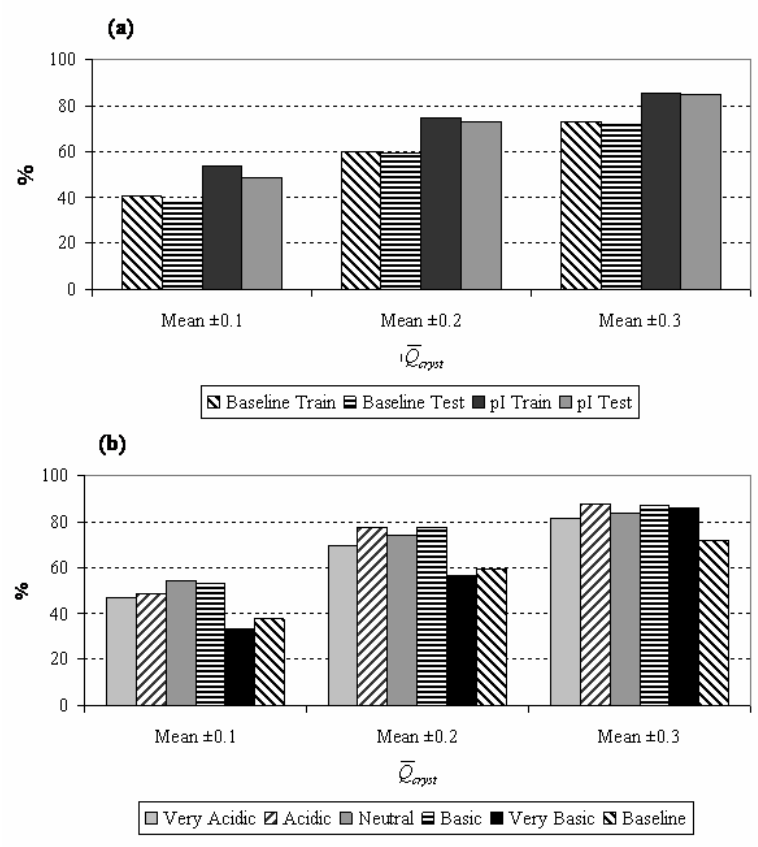


Figure 6.5 (a) The Charge Range Test results for the proteins separated by their pI_{est} or Baseline. (b) The within pI_{est} bin comparison to the Baseline group using the test set proteins.

The Charge Range Test was designed to provide a more fair comparison of clusters than the CI_{50} test, which does not take into account the variable $\overline{Q}_{cryst}$ CI_{50} ranges. The frequency of proteins within a given $\overline{Q}$ range for random groups (RG) of proteins was also examined. The Charge Range Test result for the training set proteins in the random clusters is shown in Figure 6.6 and Table 6.6. Although there was some variation among the percentages of training set proteins within each $\overline{Q}$ interval, all random groups were within 5% of each other at all levels. When the cumulative percentage of the training set proteins within each interval was compared to the Baseline group, they were exactly the same. As demonstrated in Table 6.3, each RG and the Baseline group had the same $\overline{Q}_{cryst}$ mean and standard deviation, -0.1±0.4 *e*/kDa. Therefore it was no surprise that the test set proteins Charge Range Test results for the random groups and the Baseline group were the same. This was similar to a five-fold cross-validation.

Table 6.5 The Charge Range Test results for the proteins separated by their pI_{est} and compared to Baseline.

Training Set	Very Acidic	Acidic	Neutral	Basic	Very Basic	Cum. % pI Bins	Baseline
Mean ±0.1	42.6	58.2	67.5	56.7	34.2	**53.8**	**40.4**
Mean ±0.2	63.5	79.4	83.3	82.1	60.9	**74.9**	**60.1**
Mean ±0.3	75.9	89.6	90.0	89.7	81.0	**85.5**	**73.1**
Test Set	Very Acidic	Acidic	Neutral	Basic	Very Basic	Cum. % pI Bins	Baseline
Mean ±0.1	47.1	48.9	54.6	53.2	33.3	**49.0**	**37.5**
Mean ±0.2	69.6	77.5	74.3	77.8	56.6	**73.0**	**59.6**
Mean ±0.3	81.3	87.7	83.5	87.1	86.0	**85.1**	**71.8**

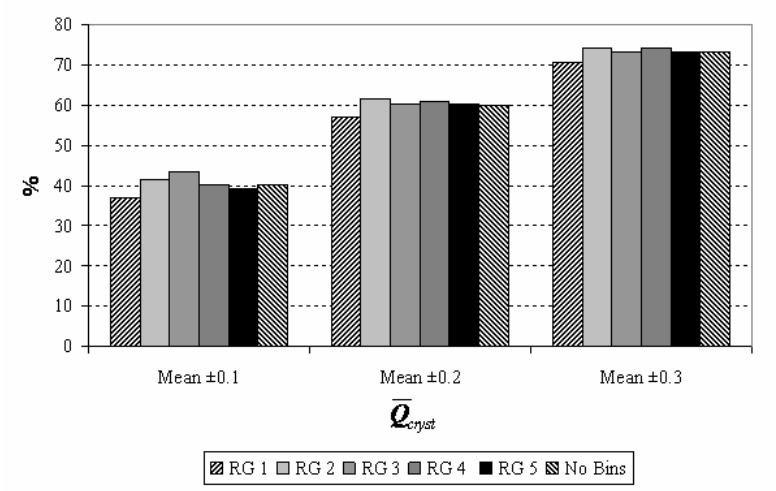


Figure 6.6 Comparing Baseline and random clusters to bracket the $\overline{Q}_{cryst}$ value of both the training and test sets.

Table 6.6 The Charge Range Test results for the proteins randomly grouped and compared to Baseline.

Training Set	RG 1	RG 2	RG 3	RG 4	RG 5	Cum. % Random Group*	Baseline
Mean ±0.1	37.1	41.6	43.6	40.2	39.4	40.4±2.4	40.4
Mean ±0.2	57.1	61.6	60.4	61.1	60.5	60.1±1.8	60.1
Mean ±0.3	70.6	74.2	73.1	74.2	73.2	73.1±1.5	73.1
Test Set	RG 1	RG 2	RG 3	RG 4	RG 5	Cum. % Random Group	Baseline
Mean ±0.1	37.5	37.5	37.5	37.5	37.5	37.5	37.5
Mean ±0.2	59.6	59.6	59.6	59.6	59.6	59.6	59.6
Mean ±0.3	71.8	71.8	71.8	71.8	71.8	71.8	71.8

* mean±SD

6.1.4 Binning Summary

As a reminder, the results presented in this section assume that the Q-related *Observables* (Q_{cryst}, $\overline{Q}_{cryst}$, and σ_{cryst}) can be used as proxy variables for the pH_{cryst}. When previously crystallized protein structures were binned into groups based on their MW_{au} or pI_{est}, statistically significant differences were observed in their *Observable* variable distributions (pH_{cryst}, Q_{cryst}, $\overline{Q}_{cryst}$, and σ_{cryst}). Using either method of binning, few pairwise differences were observed in the pH_{cryst} distributions. However, when the distributions of the Q_{cryst}, $\overline{Q}_{cryst}$, and σ_{cryst} were examined, using either method of binning, statistically significant differences were found between all groups pairwise. When proteins were randomly grouped into five clusters, no differences were found between group distributions for any *Feature*, or *Observable*. After examining for differences in variables within binning techniques, the ability of the individual bins to predict a $\overline{Q}$ interval that captured the $\overline{Q}_{cryst}$ of an independent test set was examined.

Two methods of comparing the results were examined and compared, calculating a 50% confidence interval (CI_{50}) and the Charge Range Test. Although the CI_{50} demonstrated some value in comparing the proteins within the grouping method, comparison among different

methods proved more difficult, because of varying $\overline{Q}$ CI_{50} intervals for each group. This led to the development of the Charge Range Test, which calculated the frequency (percentage) of proteins within a given $\overline{Q}$ interval around the group's mean $\overline{Q}_{cryst}$. Using the same $\overline{Q}$ interval to compare methods allowed for a more fair comparison among the different methods of grouping protein structures. Therefore, all future methods would only examine the Charge Range Test values.

Binning by the MW_{au} had little advantage over using all proteins as one group in the overall Charge Range Test. However, when the MW_{au} groups were examined individually, the 'Large' proteins were more tightly distributed around a $\overline{Q}_{cryst}$ of zero. This resulted in a better Charge Range Test, more proteins within a given $\overline{Q}$ interval, than the other two ln(MW_{au}) Bins ('Small' and 'Average') and the Baseline group. One interpretation would be that negative electrostatic forces are more detrimental for larger proteins than smaller proteins. This may simply be due to more charged amino acids. 'Small' proteins were found to crystallize over a larger $\overline{Q}$ range than did either the 'Average' or 'Large' proteins. This might be due to shielding of the charges by other solution parameters, such as salts. As such, the Charge Range Test wasn't as accurate for this group ('Small') when compared to the other two groups.

Although no overall improvement in the Charge Range Test was found when binning by ln(MW_{au}), the same was not true for separating proteins into pI_{est} Bins. Binning by pI_{est} had a significant improvement (12-15%) on both the between method (pI_{est} vs. Baseline) and within method results. Each pI_{est} Bin, except the 'Very Basic' Bin, had a significant improvement over the Baseline group. This may be explained by the high correlation between a protein's pI_{est} and the $\overline{Q}_{cryst}$. Therefore, when predicting a protein's estimated specific charge, proteins should be separated into pI_{est} Bins, rather than using all proteins as one group.

When proteins were grouped into five random groups, there were some differences observed in the individual results for the training set proteins. However, each random group had the same mean and standard deviation, which led to the same Charge Range Test results for all groups on the test set proteins. The cumulative percentage of $\overline{Q}_{cryst}$ values within a given $\overline{Q}$ interval was the same as that observed for using all proteins as one group, Baseline. This result ended up being a pseudo-five-fold cross-validation, which gave an estimated error for the Charge

Range Test of 2.2%. While the number of clusters was set rather arbitrarily in each binning method, there was no reason to believe that these methods were the optimal method of grouping proteins by similarity. Therefore, other unsupervised methods of grouping proteins by similarity were also explored in the next section.

6.2 UNSUPERVISED CLUSTERING

In the previous section, successfully crystallized proteins were grouped by either their MW_{au} or pI_{est}. Although significant differences were found in these group's distributions of *Observables*, the pH_{cryst}, Q_{cryst}, $\overline{Q}_{cryst}$, and σ_{cryst}, the number of groups actually present in the data was unknown. The number of pI_{est} groups was initially chosen based upon some limited domain knowledge (Ries-Kautt and Ducruix, 1999), while the number of $\ln(MW_{au})$ groups was chosen rather arbitrarily.

In this section, rather than arbitrarily setting cutoff points for $\ln(MW_{au})$ or using limited domain knowledge (pI_{est}), two unsupervised clustering algorithms were used to group proteins by the similarity of their estimated specific charge curves ($\overline{Q}$ plotted against pH), where similarity is judged by the $\overline{Q}$ values over a given range and interval of pH. Instead of one point along the $\overline{Q}$ curve to group proteins (binning by pI_{est}), the ability of a larger section of the $\overline{Q}$ curve to separate proteins was examined. This section builds upon the preliminary results in Chapter 3, Section 3.2.2, where proteins were grouped by their scaled estimated titration curve using the GSOM algorithm. However, instead of using the whole estimated titration curve or the scaled titration curve (pH 1.0-14.0), a portion (pH 4.0-10.0) of the $\overline{Q}$ curve was used. It was hypothesized that using the $\overline{Q}$ curve would result in better separation of groups than using one point on the curve, binning by pI_{est}. With more knowledge available, the unsupervised clustering algorithms should more accurately group proteins than did binning as determined by the Charge Range Test. A one-dimensional vector of $\overline{Q}$ values every 0.2 pH units 4.0-10.0 (31 values) was used as the input for each clustering algorithm. This range was chosen, because few proteins (~1.3% of the training set) were crystallized outside of this pH range, as demonstrated by the

pH_{cryst} distributions in Chapters 3 & 5. Also, pH values above and below this range may have denaturing effects on the native protein. It was hypothesized that there would be statistically significant differences in the distribution of protein *Features* and *Observables* between the different clusters. Initially, a subset of five was the initial arbitrary number chosen. There could be more or less number of groups, but it was hypothesized that there would be a greater number of subgroups than five.

Two unsupervised clustering algorithms (two-step clustering and self-organizing maps) were used to compare the binning techniques in the previous section (Section 6.1) with an unsupervised approach. The first method, two-step clustering, has the ability to determine the 'optimum' number of clusters by examining the information criteria of the clusters as they are formed (for more details see Section 4.7.2.1). First, the number of groups was set to five for each unsupervised method to compare with results from binning by pI_{est}. Then the 2Step algorithm was allowed to determine the optimal number of clusters. Similarly, the initial application of the self-organizing map (SOM) algorithm used five clusters. Then the Supervised SOM algorithm was used to determine the number of clusters (Section 4.7.2.2).

Initially, each clustering method was compared to binning by pI_{est} by examining the pI_{est} distribution of the clusters to determine how the unsupervised clustering algorithms separated the proteins into different clusters. Then various protein *Features* and *Observable* distributions were examined for differences within the clustering method and between clustering methods. The ability to separate proteins into the maximum number of groups with different statistically significant *Observable* distributions was the goal of using the unsupervised clustering techniques. Once this could be done, modeling the distributions with Gaussians was examined for predictive purposes and to determine which of the methods worked best in separating the proteins (Section 6.3).

6.2.1 Two-Step Clustering

The two-step clustering algorithm (2Step) of Chui et al. (2001) was available in SPSS 14.0/Clementine 8.0. To compare the two-step clustering algorithm to the initial supervised clustering method (Binning by pI_{est}), the number of clusters was initially fixed at five (Section 6.2.1.1). Next, the 2Step clustering algorithm was allowed to choose the 'optimum' number of

clusters (Section 6.2.1.2). It was not known apriori how many clusters would be found present in the data, but it was hypothesized that the number would be larger than five due to the large number of protein structures examined.

6.2.1.1 Two-Step with 5 Clusters

The distribution of proteins within the 5 two-step ($2Step_5$) clusters was compared to the frequency distribution of proteins within the pI_{est} Bins (Section 6.1.1.2) in Table 6.7. The mean $\overline{Q}$ curves for each cluster are shown in Figure 6.7. The 2Step algorithm grouped the 'Very Acidic' and 'Acidic' proteins in $2Step_5$ Clusters 2 and 5, while $2Step_5$ Cluster 1 contained most of the proteins with a $6.0 \leq pI_{est} \leq 9.0$ (46% of the data; Table 6.7; Figures 6.8a). Two of the clusters (1 and 2) contained ~80% of the proteins, which demonstrated that most of the proteins could be placed into two groups, while the proteins with the more extreme pI_{est} values were placed in the remaining three groups. $2Step_5$ Cluster 4 contained only 2% of the proteins, which were the most acidic proteins (pI_{est}='Very Acidic') with a mean (±SD) pI_{est} of 3.7±0.3. $2Step_5$ Cluster 5 contained most of the proteins with a $4.0 \leq pI_{est} \leq 5.0$. The remaining 'Acidic' proteins, $5.0 \leq pI_{est} \leq 6.0$, were grouped in $2Step_5$ Cluster 2. The 'Very Basic' proteins were placed in $2Step_5$ Cluster 3 along with few of the 'Basic' proteins. The 2Step clustering algorithm appeared to be able to use more information present within the $\overline{Q}$ curve to group 'similar' proteins.

Additionally, there were also significant differences in the MW_{au} and $\ln(MW_{au})$ distributions between clusters, that may partially explain some of the overlap between pI_{est} values (Table 6.8; Figure 6.8a). The clusters (3 and 4) with the extreme pI_{est} values had a smaller mean MW_{au} (<32 kDa). Sporadic differences were also observed with the $diff_{lim}$ and pH_{cryst} distributions (Figure 6.8d). Although only slight differences were observed in pH_{cryst} distributions, every cluster had significantly different Q_{cryst}, $\overline{Q}_{cryst}$, and σ_{cryst} distributions (Figure 6.8bc). This mirrored the findings using pI_{est} Binning.

Using five clusters, the 2Step algorithm was able to separate the proteins into distinct groups as demonstrated by their $\overline{Q}$ curves. Each cluster exhibited a distinct distribution for the *Hidden Observables* (Q_{cryst}, $\overline{Q}_{cryst}$, and σ_{cryst}). However, similar to Binning by pI_{est}, each cluster

did not have a unique pH_{cryst} distribution, again demonstrating the potential for using *Hidden Observable* (Q_{cryst}, $\overline{Q}_{cryst}$, and σ_{cryst}) distributions and the *Hidden Controllables* (Q, $\overline{Q}$, or σ) for modeling purposes.

Table 6.7 Cross-tabulation of the clusters generated by pI_{est} Binning and two-step clustering with five clusters ($2Step_5$).

Cluster	Binning by pI_{est}					Total	% Total	Mean pI_{est}
	Very Acidic	Acidic	Neutral	Basic	Very Basic			
$2Step_5$ 1	2	383	887	501	121	1894	46.0	7.0
$2Step_5$ 2	549	788	48	1	0	1386	33.7	5.2
$2Step_5$ 3	0	0	2	62	327	391	9.5	9.6
$2Step_5$ 4	78	0	0	0	0	78	1.9	3.7
$2Step_5$ 5	347	17	1	0	0	365	8.9	4.4
Total	976	1188	938	564	448	4114	100.0	6.3
% Total	23.7	28.9	22.8	13.7	10.9	100.0		
Mean pI_{est}	4.5	5.5	6.6	8.2	9.7	6.3		

Table 6.8 The variable descriptors (mean and SD) for each of the $2Step_5$ clustering groups separated by their estimated specific charge curves from pH 4.0-10.0.

$2Step_5$ Cluster	n		*Features*		*Observables*					Random Variable
			pI_{est}	MW_{au}	σ_{cryst}	pH_{cryst}	$\overline{Q}_{cryst}$	Q_{cryst}	$diff_{lim}$	
1	1894	Mean	7.0 ^{A}	66.3 ^{A}	14.2 ^{A}	6.7 ^{A}	0.1 ^{A}	1.4 ^{A}	2.05 ^{A}	0.03 ^{A}
		SD	1.1	74.8	59.2	1.2	0.2	18.2	0.43	0.98
2	1386	Mean	5.2 ^{B}	70.4 ^{A}	-56.4 ^{B}	6.6 ^{AB}	-0.2 ^{B}	-14.1 ^{B}	2.04 ^{A}	0.02 ^{A}
		SD	0.5	87.0	69.6	1.3	0.2	26.9	0.46	1.00
3	391	Mean	9.6 ^{C}	31.6 ^{B}	124.3 ^{C}	6.7 ^{AB}	0.5 ^{C}	15.2 ^{C}	1.96 ^{AC}	0.00 ^{A}
		SD	0.7	33.1	92.2	1.5	0.4	20.4	0.44	1.03
4	78	Mean	3.7 ^{D}	20.0 ^{C}	-282.0 ^{D}	6.3 ^{B}	-1.2 ^{D}	-23.8 ^{D}	1.74 ^{B}	-0.03 ^{A}
		SD	0.3	20.7	113.7	1.4	0.5	26.8	0.38	1.04
5	365	Mean	4.4 ^{E}	40.6 ^{D}	-125.7 ^{E}	6.4 ^{B}	-0.5 ^{E}	-20.1 ^{E}	1.96 ^{C}	-0.02 ^{A}
		SD	0.3	46.0	92.9	1.4	0.3	28.6	0.44	1.02
KW Test		x^2	3264	401	1999	21.4	2064	1479	56.0	1.5
		$p<$	0.0001	0.0001	0.0001	0.0003	0.0001	0.0001	0.0001	0.833

Note: Groups labeled with different letters (A, B, C, D, or E) have significantly different distributions ($p<0.01$) as determined by a KS Test.

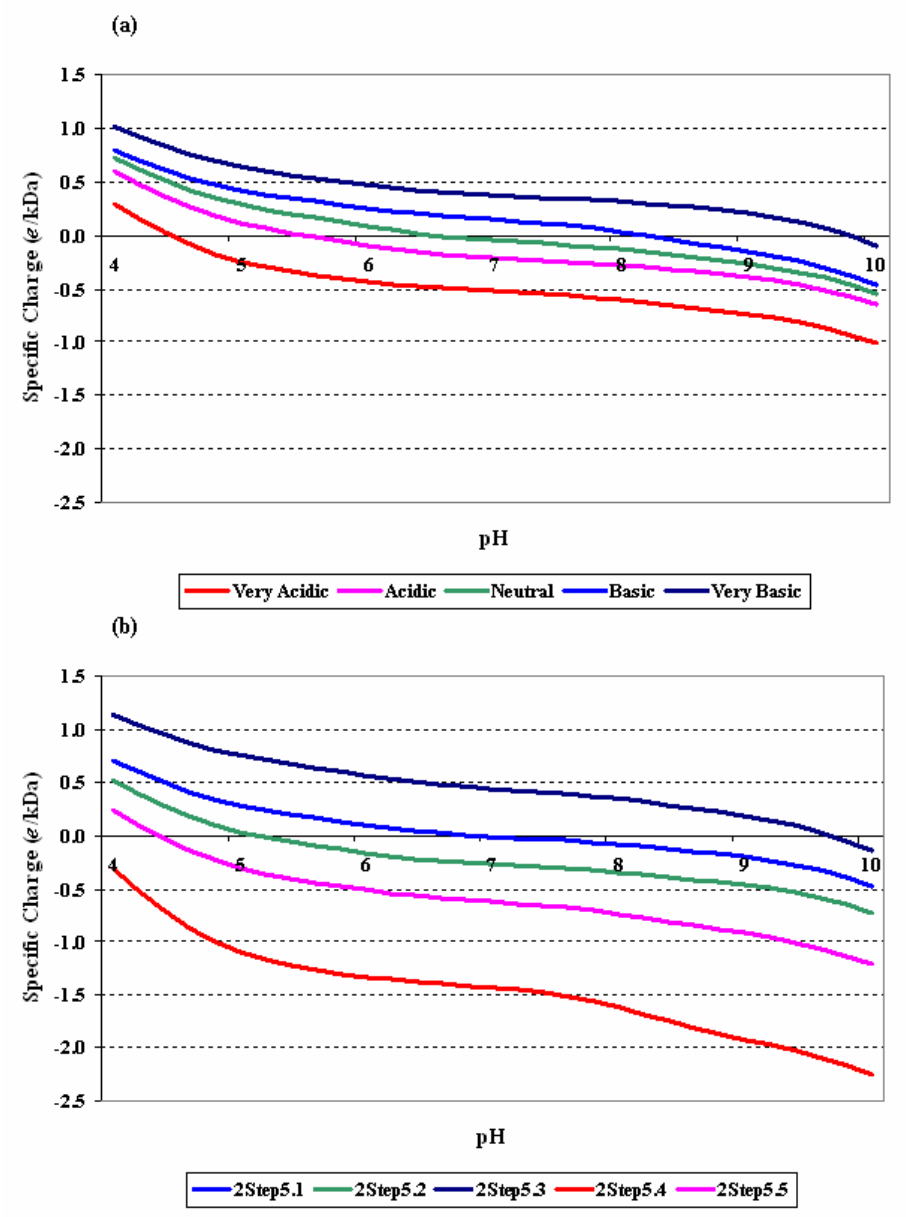


Figure 6.7 The mean $\overline{Q}$ curves for each (a) pI_{est} Bin and (b) $2Step_5$ Cluster.

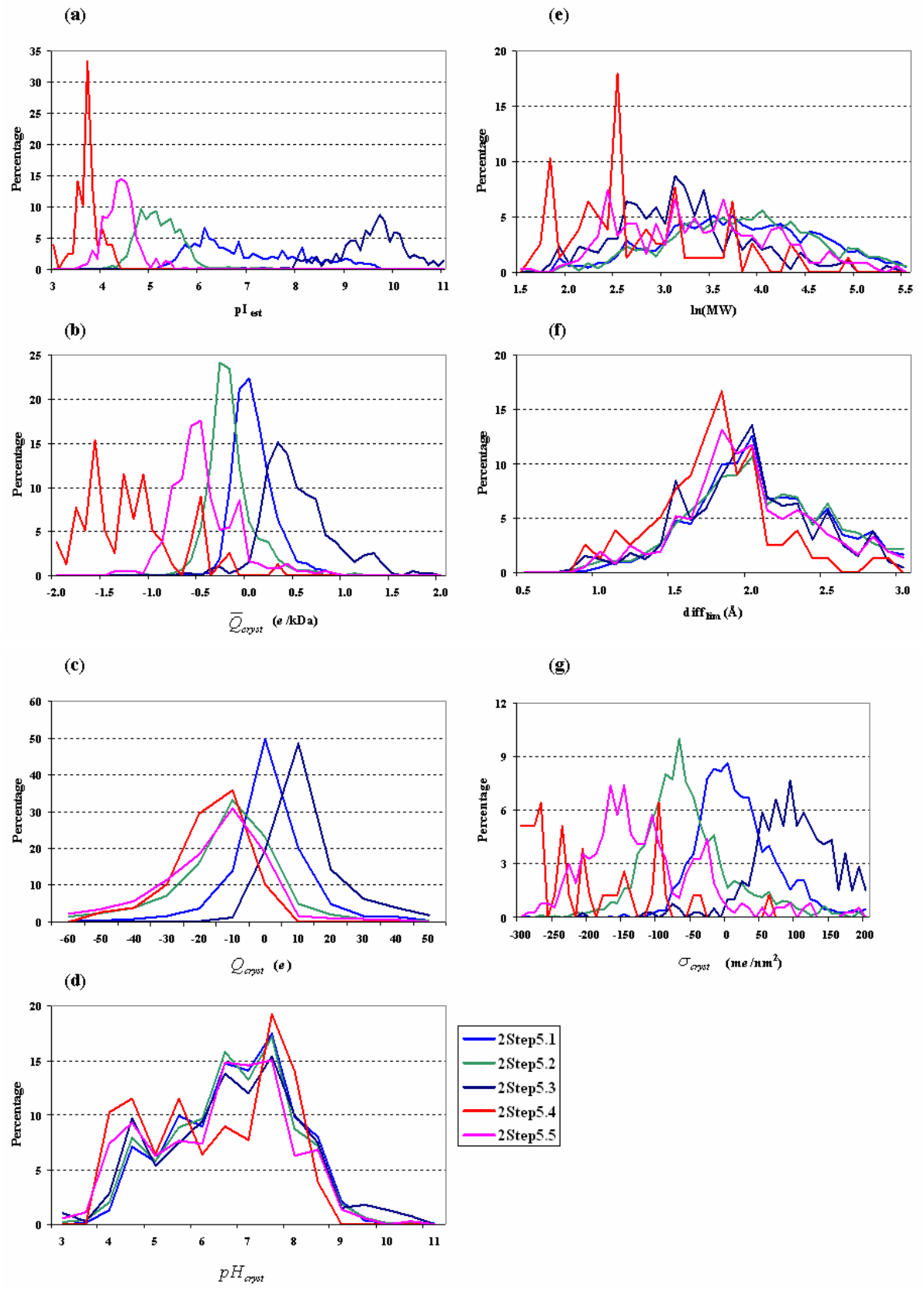

(a)
Percentage
pI est
(b)
Percentage
$\overline{Q}_{cryst}$ (e /kDa)
(c)
Percentage
Q_{cryst} (e)
(d)
Percentage
pH_{cryst}
(e)
Percentage
ln(MW)
(f)
Percentage
diff Hm (Å)
(g)
Percentage
σ_{cryst} (me /nm²)
2Step5.1
2Step5.2
2Step5.3
2Step5.4
2Step5.5

Figure 6.8 The $2Step_5$ distributions of (a) pI_{est}, (b) $\overline{Q}_{cryst}$, (c) Q_{cryst}, (d) pH_{cryst}, (e) $\ln(MW_{au})$, (f) $diff_{lim}$, and (g) σ_{cryst}.

6.2.1.2 Charge Range Test

After demonstrating the ability of the 2Step algorithm to cluster the proteins into distinct groups by their $\overline{Q}$ curve similarity, the Charge Range Test was performed on the clusters and compared to Baseline and pI_{est} Binning. Because there were no differences in the Charge Range Test between Baseline, $\ln(MW_{au})$ binning, and Random Binning, the latter two were not reported here (see Section 6.1.3). There was an obvious improvement in the amount of proteins captured by the Charge Range Test when comparing the cumulative percentage versus Baseline at all levels (Table 6.9; Figure 6.9a). A 14% increase in the Charge Range Test results over Baseline was observed at all levels. When compared to the pI_{est} Bin Charge Range Test results, 2Step clustering resulted in no improvement at the low intervals, Mean±0.1 and Mean±0.2, and a slightly better result at the Mean±0.3 level (+4%). Because of simplicity and ease of calculation, along with similar $\overline{Q}$ range results, pI_{est} binning should be used over 2Step clustering with 5 clusters. However, it was not known what the 'optimal' number of clusters should be, which was examined in the next section (Section 6.2.1.3).

Table 6.9 The Charge Range Test for each $2Step_5$ cluster and that of the Baseline group (Chapter 5).

Training Set	$2Step_5$ 1 N=1894	$2Step_5$ 2 N=1386	$2Step_5$ 3 N=391	$2Step_5$ 4 N=78	$2Step_5$ 5 N=365	Cum. % $2Step_5$	Baseline
Mean ±0.1	51.1	59.9	33.0	29.5	43.3	51.2	40.4
Mean ±0.2	78.6	78.6	56.8	37.2	59.5	74.0	60.1
Mean ±0.3	92.6	88.2	73.9	46.2	75.1	86.9	73.1
Test Set	$2Step_5$ 1 N=587	$2Step_5$ 2 N=424	$2Step_5$ 3 N=128	$2Step_5$ 4 N=9	$2Step_5$ 5 N=98	Cum. % $2Step_5$	Baseline
Mean ±0.1	48.0	56.8	32.8	44.4	48.0	49.4	37.5
Mean ±0.2	72.6	76.4	60.2	55.6	72.4	72.5	59.6
Mean ±0.3	90.1	88.2	75.8	77.8	81.6	87.2	71.8

Next, the clusters within the $2Step_5$ method were compared (Figure 6.9b). $2Step_5$ Clusters 1 and 2, the most populated clusters, both had a significant improvement at all levels over the Baseline group. $2Step_5$ Cluster 4 had the fewest training (n=78) and test (n=9) set proteins. This group appeared to capture the small 'Very Acidic' proteins, and had a lower Charge Range Test score than did all other groups within method and between methods (Baseline and pI_{est} Bins). Similarly, $2Step_5$ Cluster 3 represented the 'Very Basic' proteins and also had an accuracy below that of the Baseline group. These two groups represented more extreme cases and there was difficulty predicting the $\overline{Q}$ range. Thus, the cumulative percentage of protein structures captured by Charge Range Test was much more accurate at all levels. However, the $2Step_5$ clustering method performed no better than did binning by pI_{est}. A slight improvement of 4% was observed at the Mean ±0.3 interval. The next step was to attempt to use the 2Step clustering algorithm's ability to determine the number of clusters present within the data.

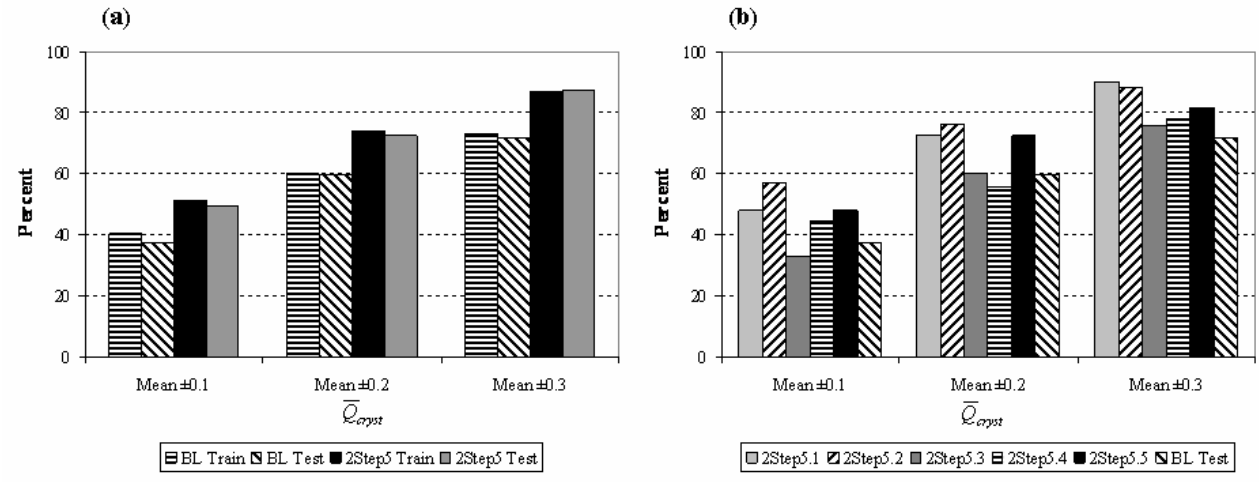


Figure 6.9 (a) The cumulative Charge Range Test results for $2Step_5$ clustering and Baseline (BL). (b) The breakdown of each $2Step_5$ cluster compared to Baseline on the test set proteins.

6.2.1.3 Determining the Optimum Number of Clusters

After demonstrating that protein structures could be separated by their $\overline{Q}$ curves, the next step was to determine how many clusters were present within the data. For this step, the 2Step

algorithm was allowed to choose the optimal number of clusters based upon Schwarz's Bayesian Information Criteria (BIC; Schwarz, 1978) using maximum likelihood distance.

Table 6.10 The optimal number of clusters chosen by the 2Step algorithm was six, using BIC and maximum likelihood distance.

Number of Clusters	BIC	BIC Change [a]	Ratio of BIC Changes [b]	Ratio of Distance Measures [c]
1	-174512			
2	-203369	-28856.6	1.000	1.478
3	-222729	-19359.9	0.671	2.816
4	-229271	-6542.5	0.227	1.069
5	-235360	-6088.7	0.211	1.310
6	-239884	-4524.3	**0.157**	**2.947**
7	-241078	-1194.3	0.041	1.206
8	-241981	-902.6	0.031	1.021
9	-242854	-873.2	0.030	1.055
10	-243655	-800.6	0.028	1.055
11	-244387	-731.7	0.025	1.650
12	-244627	-240.2	0.008	1.234
13	-244723	-96.6	0.003	1.092
14	-244768	-45.0	0.002	1.297
15	-244685	83.3	-0.003	1.003

[a] The changes are from the previous number of clusters in the table.

[b] The ratios of changes are relative to the change from the two cluster solution.

[c] Compares the distance measures between the current and previous number of clusters.

The 2Step algorithm determined that the optimal number of clusters for the training data was six using BIC analysis ($2Step_6$; Table 6.10). The $\overline{Q}$ curves of each protein cluster are shown in Figure 6.10. When compared to the $2Step_5$ clusters, the $2Step_6$ Clusters split $2Step_5$ Clusters 1 (mean pI_{est} of 7.0) into $2Step_6$ Clusters 4 (mean pI_{est} of 6.2) and 5 (mean pI_{est} of 8.4). There was also some other rearrangement of proteins, but there was a group of proteins with the same mean pI_{est} as in the $2Step_5$ Clusters (Table 6.11).

The BIC analysis resulted in six clusters with 2Step clustering. Although similar to the $2Step_5$ clusters, the addition of the extra cluster still maintained the statistically significant differences observed in the pI_{est} and the *Hidden Observable* distributions of Q_{cryst}, $\overline{Q}_{cryst}$, and σ_{cryst} (Table 6.12; Figure 6.11a-c,g). The pH_{cryst} distributions displayed few differences; with the most acidic cluster ($2Step_6$ Cluster 1) having lower pH_{cryst} distributions than clusters 4 and 5, which had a smaller percentage of proteins that crystallized at pH 4-4.5 (Figure 6.11d). Varying differences were also observed with the MW_{au} and $diff_{lim}$ distributions (Figure 6.11e-f). Only $2Step_6$ Clusters 3 and 4 had similar MW_{au} distributions. All other Clusters had MW_{au} distributions that were significantly different from one another. The $2Step_6$ clusters with the extreme pI_{est} distributions (Clusters 1 and 6) had significantly lower mean $diff_{lim}$ values (1.8 and 1.9 Å, respectively) than the other clusters which averaged 2.0-2.1 Å. No differences were observed in the random number distributions.

Table 6.11 Cross-tabulation of the 2Step clustering results using 5 ($2Step_5$) or 6 clusters ($2Step_6$).

$2Step_6$ Cluster	$2Step_5$ Cluster					Total	% Total	Mean pI_{est}
	1	2	3	4	5			
1	0	0	0	78	5	83	2.0	3.7
2	0	129	0	0	360	489	11.9	4.5
3	14	1215	0	0	0	1229	29.9	5.2
4	1143	42	0	0	0	1185	28.8	6.2
5	737	0	178	0	0	915	22.2	8.4
6	0	0	213	0	0	213	5.2	9.8
Total	1894	1386	391	78	365	4114	100.0	6.3
% Total	46.0	33.7	9.5	1.9	8.9	100.0		
Mean pI_{est}	7.0	5.2	9.6	3.7	4.4	6.3		

Table 6.12 The variable descriptors (mean and SD) for each of the $2Step_6$ clustering groups separated by their $\overline{Q}$ curves from pH 4.0-10.0.

$2Step_6$			*Features*		*Observables*					Random
Cluster	n		pI_{est}	MW_{au}	σ_{cryst}	$\overline{Q}_{cryst}$	Q_{cryst}	pH_{cryst}	$diff_{lim}$	Variable
1	83	Mean	3.7 ^{A}	20.1 ^{A}	-276.0 ^{A}	-1.2 ^{A}	-23.5 ^{A}	6.3 ^{A}	1.8 ^{A}	-0.05 ^{A}
		SD	0.3	20.6	115.0	0.5	26.3	1.4	0.4	1.03
2	489	Mean	4.5 ^{B}	46.8 ^{B}	-118.2 ^{B}	-0.4 ^{B}	-21.4 ^{B}	6.5 ^{AB}	2.0 ^{BD}	0.00 ^{A}
		SD	0.4	57.1	89.8	0.3	30.8	1.4	0.4	1.02
3	1229	Mean	5.2 ^{C}	70.8 ^{C}	-52.5 ^{C}	-0.2 ^{C}	-13.1 ^{C}	6.6 ^{AB}	2.0 ^{BC}	0.03 ^{A}
		SD	0.5	88.5	67.5	0.2	25.8	1.3	0.5	1.00
4	1185	Mean	6.2 ^{D}	73.9 ^{C}	-4.3 ^{D}	0.0 ^{D}	-2.1 ^{D}	6.7 ^{B}	2.1 ^{C}	0.01 ^{A}
		SD	0.6	84.3	55.6	0.2	20.5	1.2	0.4	0.95
5	915	Mean	8.4 ^{E}	51.1 ^{D}	49.8 ^{E}	0.2 ^{E}	7.8 ^{E}	6.7 ^{B}	2.0 ^{BD}	0.07 ^{A}
		SD	0.9	50.6	58.2	0.2	13.8	1.3	0.4	1.01
6	213	Mean	9.8 ^{F}	28.4 ^{E}	160.2 ^{F}	0.7 ^{F}	17.9 ^{F}	6.7 ^{AB}	1.9 ^{D}	-0.07 ^{A}
		SD	0.8	31.4	95.0	0.4	20.6	1.5	0.4	1.03
KW	x^2		3542	389.5	2187	2244	1716	15.4	73.7	5.3
Test	$p<$		0.0001	0.0001	0.0001	0.0001	0.0001	0.009	0.0001	0.379

Note: Groups labeled with different letters (A, B, C, D, E, or F) have significantly different distributions ($p<0.01$) as determined by a KS Test.

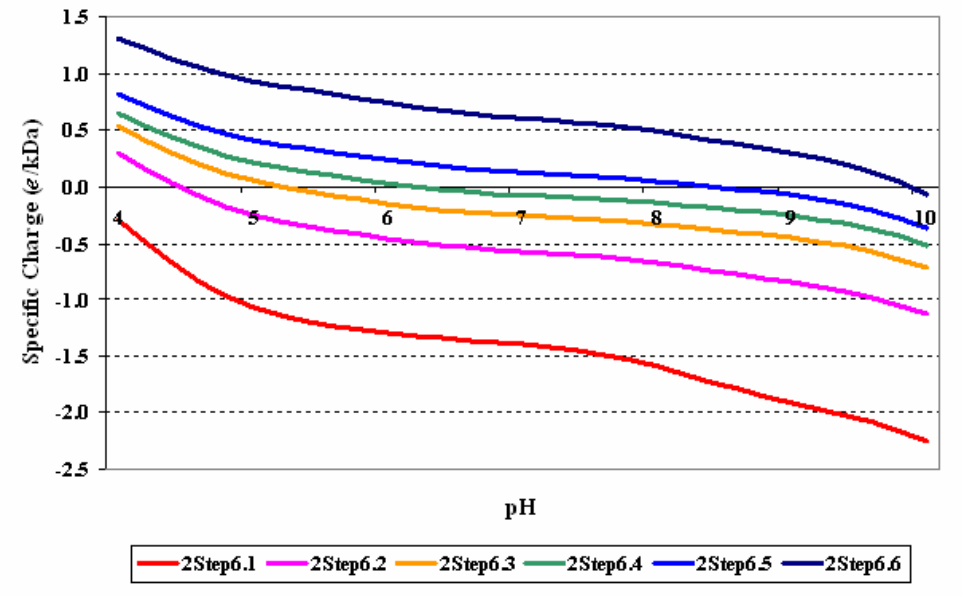


Figure 6.10 The mean $\overline{Q}$ curves for each $2Step_6$ cluster.

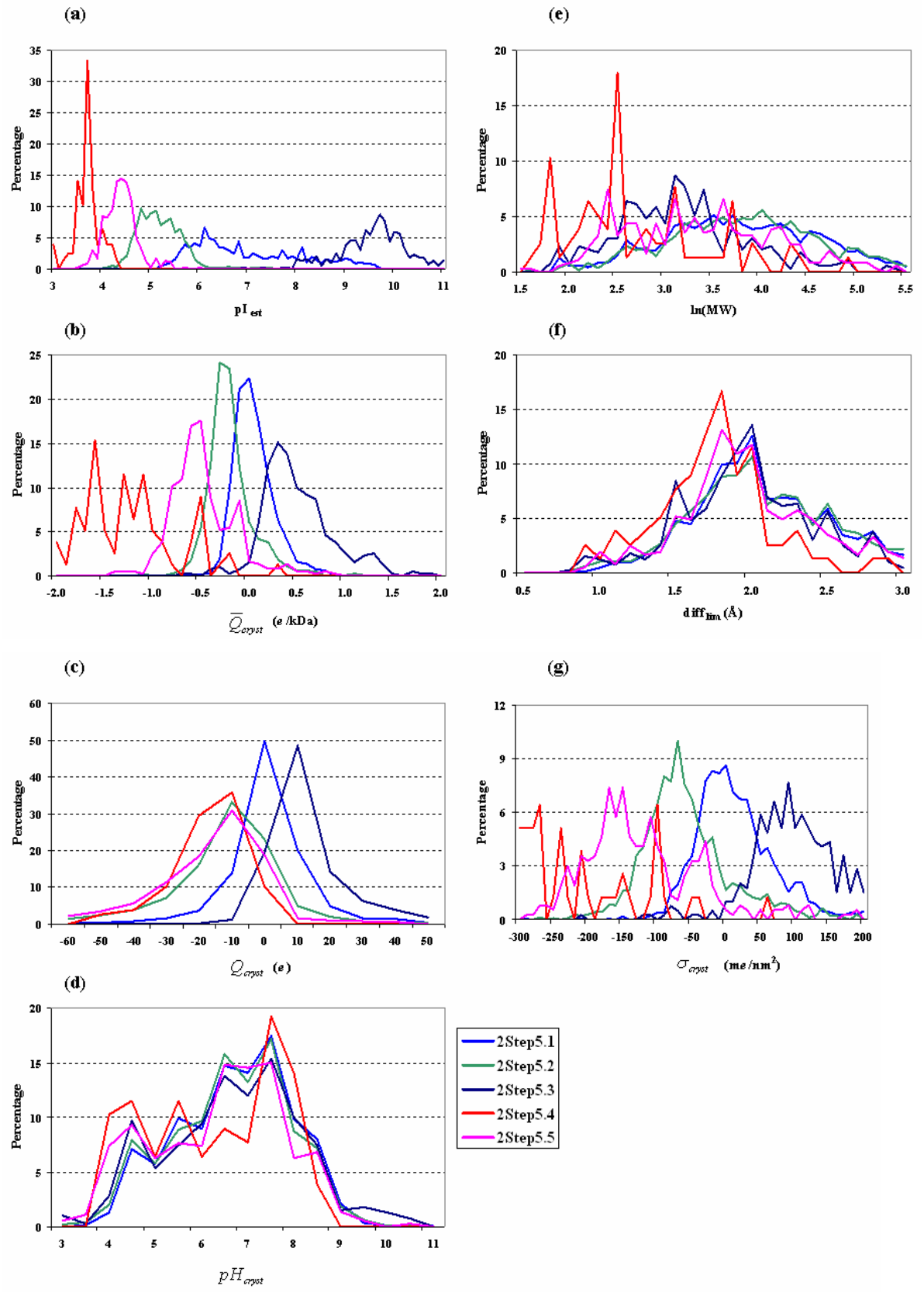

(a)
Percentage
pI est
(e)
ln(MW)
(b)
(f)
diff Hm (Å)
(c)
(g)
(d)
2Step5.1
2Step5.2
2Step5.3
2Step5.4
2Step5.5

Figure 6.11 The $2Step_6$ distributions of (a) pI_{est}, (b) $\overline{Q}_{cryst}$, (c) Q_{cryst}, (d) pH_{cryst}, (e) $\ln(MW_{au})$, (f) $diff_{lim,}$ and (g) σ_{cryst}.

6.2.1.4 Charge Range Test

The $2Step_6$ results of the Charge Range Test are shown in Table 6.12. Similar to the $2Step_5$ results (Table 6.9), the $2Step_6$ results demonstrated significant improvement over the Baseline group at all levels (Figure 6.12a). While the $2Step_5$ results showed little to no cumulative improvement over the pI_{est} Bin results, allowing the 2Step algorithm to select the number of clusters (6 clusters; $2Step_6$) resulted in a 4-5% improvement over both the pI_{est} Bin (Table 6.5) and $2Step_5$ results (Table 6.9) at all ranges. Therefore, this would seem to represent a more optimal separation of proteins.

After examining the cumulative $2Step_6$ results, the focus turned to the results within the method (Figure 6.12b). $2Step_6$ Clusters 1 and 6 represented the extremely acidic and basic proteins, respectively. Similar to previous results with pI_{est} binning, these extreme cases had lower prediction accuracy than the Baseline group. These two clusters performed poorly on both the training and test sets. $2Step_6$ cluster 2 performed similar to the Baseline group at all levels, although slightly better at the Mean±0.3 level. At the Mean±0.3 level, $2Step_6$ Clusters 3-5 all approached 90% accuracy in selecting the correct $\overline{Q}$ range for crystallization, which was similar for the 'Acidic,' 'Neutral,' and 'Basic' pI_{est} groups. A $\overline{Q}$ range of Mean±0.1 or Mean±0.2 seems to maximize the performance over the Baseline method, +17%.

Table 6.13 The Charge Range Test results for each $2Step_6$ cluster and Baseline (Chapter 5).

Training Set	$2Step_6$ 1 n=83	$2Step_6$ 2 n=489	$2Step_6$ 3 n=1229	$2Step_6$ 4 n=1185	$2Step_6$ 5 n=915	$2Step_6$ 6 n=213	Cum. % $2Step_6$	Baseline
Mean ±0.1	28.9	39.3	63.1	64.2	58.6	30.5	57.2	40.4
Mean ±0.2	37.3	60.7	81.0	86.1	81.5	50.7	77.7	60.1
Mean ±0.3	45.8	77.9	89.2	93.3	90.2	70.4	87.4	73.1
Test Set	$2Step_6$ 1 n=12	$2Step_6$ 2 n=132	$2Step_6$ 3 n=386	$2Step_6$ 4 n=365	$2Step_6$ 5 n=286	$2Step_6$ 6 n=65	Cum. % $2Step_6$	Baseline
Mean ±0.1	33.3	44.7	58.8	57.5	53.8	36.9	54.4	37.5
Mean ±0.2	41.7	63.6	78.5	82.2	79.0	58.5	76.7	59.6
Mean ±0.3	58.3	81.1	88.3	90.7	88.8	78.5	87.6	71.8

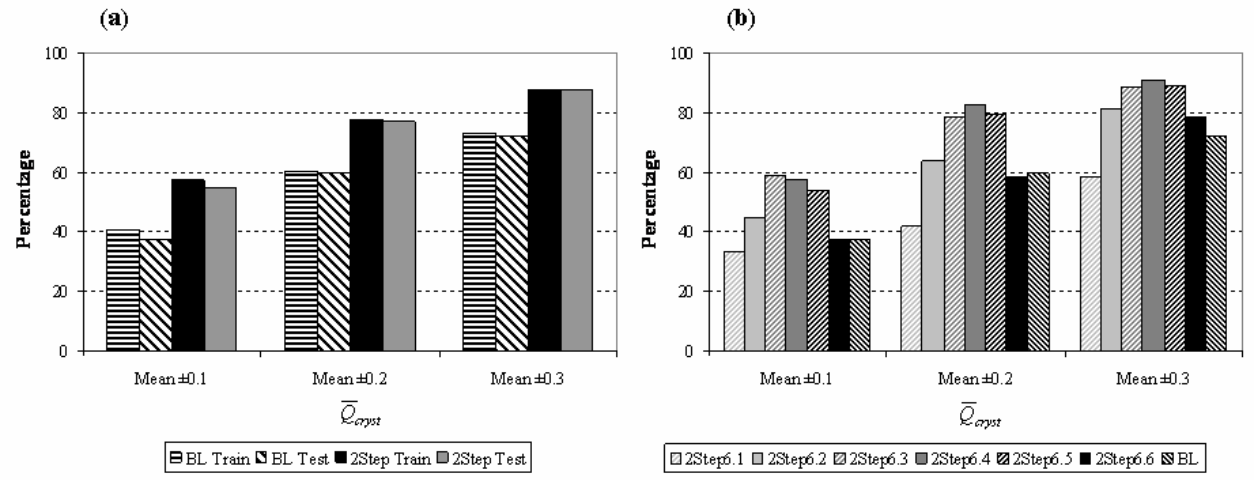


Figure 6.12 Comparison of (a) cumulative or (b) individual Charge Range Test results between $2Step_6$ clustering and Baseline.

6.2.2 Self-Organizing Maps (SOMs)

The self-organizing map (SOM) algorithm (Kohonen, 2001) usually requires the user to set the number of clusters apriori. The number of clusters may be set arbitrarily or by using some domain knowledge. Usually, the SOM clusters are a two-dimensional projection of a

multidimensional dataset with each cluster represented by an *x* and *y* value. The number of clusters within such a SOM equals x * y. Alternatively, there are a few dynamic SOM algorithms that allow the data to determine the number of clusters. One example was the GSOM algorithm (Hsu et al., 2003), which was described in Chapter 3. A dynamic SOM measures the error within a cluster and when this error reaches a threshold, another cluster is added adjacent to the best-fit node (see Section 4.7.2.2).

Initially, a 5x1 SOM (SOM_{5x1}) was chosen to compare the results with binning by the pI_{est} (Section 6.1). This size SOM would result in a maximum of five clusters with a spatial relationship between each cluster. Cluster 1 would be more similar to Cluster 2 than to Cluster 3. The input vector was the $\overline{Q}$ values of the $\overline{Q}$ curve from 4.0-10.0 every 0.2 pH units, the same values used in the previous section, 2Step clustering. The self-organizing map (SOM) algorithm was implemented using Clementine 8.0 (SPSS, Chicago, IL).

The default settings were used for the first Learning Phase (growing), the learning rate decay (linear), neighborhood (2), initial eta (0.3), and cycles (20). The second Learning Phase (smoothing) had a neighborhood of 1, an initial eta of 0.1, and 150 cycles.

The next step was to determine the optimum SOM dimensions (i.e. number of clusters). Two methods were attempted, a dynamic SOM and the Supervised SOM algorithm developed in Section 6.2.3. The dynamic SOM package used was the same one used in the Preliminary Results (Section 3.2.2), GSOMPak 1.0 Beta, kindly provided by Art Hsu (Hsu et al., 2003). Similar growth and smoothing parameters to those used in the initial SOM analysis in Clementine were used. The initial spread factor was set to 0.1 (the lowest setting), which should result in the fewest number of clusters. After the initial attempt with the GSOM package, a smaller amount of cycles was used for the smoothing phase (50) in an attempt to reduce the number of clusters in the final model.

6.2.2.1 SOM with 5 Clusters

Initially, five clusters were chosen to allow easy comparison to binning by the pI_{est} (Section 6.1.) and the other unsupervised clustering method, 2Step Clustering (Section 6.2.1). These clusters were created by using a 5x1 SOM (SOM_{5x1}). The proteins within the SOM_{5x1} clusters were first compared to binning by pI_{est} and then the other unsupervised clustering algorithm.

In the initial analysis (Section 6.1.1.2), the proteins were divided into 5 bins based upon the proteins pI_{est} values, 'Very Acidic' ($pI_{est} \leq 5.0$), 'Acidic' ($5.0 < pI_{est} \leq 6.0$), 'Neutral' ($6.0 < pI_{est} < 8.0$), 'Basic' ($8.0 \leq pI_{est} < 9.0$), and 'Very Basic' ($pI_{est} \geq 9.0$). These pI_{est} cut point values were determined with some limited domain knowledge and it was unknown if there were better ways to group the data. The unsupervised SOM algorithm broke up proteins differently with each group containing 14-30% of the proteins (Table 6.14). The resulting frequencies may not be surprising given the preponderance of 'Acidic' and 'Neutral' proteins in the training set. The SOM split the two acidic pI_{est} groups ($pI_{est} \leq 6.0$) into 3 clusters (Clusters 2-4) and combined most of the 'Basic' and 'Very Basic' proteins ($pI_{est} \geq 8.0$) into Cluster 0. The majority of 'Acidic' proteins fell into Cluster 2, along with a sizable amount in Cluster 3. The 'Very Acidic' proteins were divided between Clusters 3 and 4. All SOM_{5x1} groups contained some proteins labeled as 'Neutral'; however, the majority of these proteins fell in Clusters 1 and 2. The percentage of proteins within each group was more balanced than observed with the pI_{est} Binning where 10-14% of the proteins were found in each of the 'Basic' and 'Very Basic' pI_{est} Groups. The 'Very Acidic', 'Acidic', and 'Neutral' pI_{est} groups each had approximately 25% of all the training set proteins. The mean $\overline{Q}$ curves for each SOM_{5x1} cluster are shown in Figure 6.13. These curves, especially the SOM_{5x1} Clusters 0 and 4, were much closer together than were those obtained using 2Step clustering (Figure 6.7 and 6.10). This could be explained by these clusters having a larger proportion of proteins (14-20%) than did the most acidic and basic 2Step clusters (2-10%).

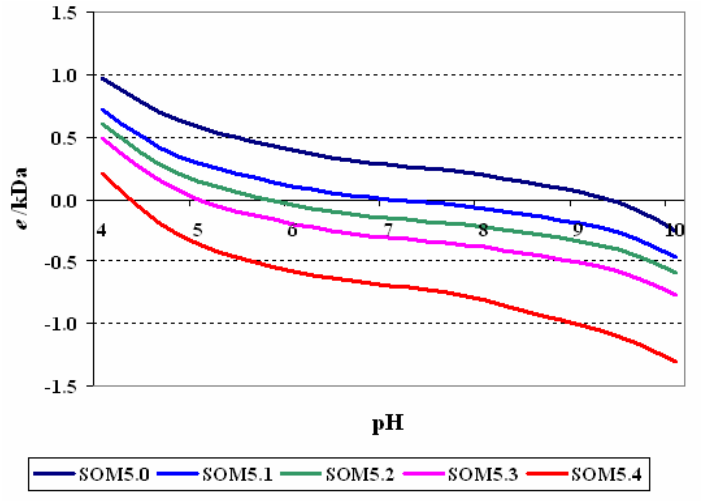


Figure 6.13 The SOM_{5x1} cluster's mean $\overline{Q}$ curves.

Table 6.14 Cross-tabulation of the SOM_{5x1} clusters vs. the (a) pI_{est} bins or (b) $2Step_5$ clusters.

(a) pI_{est} Bin

Cluster	Very Acidic	Acidic	Neutral	Basic	Very Basic	Total	% Total	Mean pI_{est}
SOM_{5x1} 0	0	0	52	328	441	821	20.0	9.0
SOM_{5x1} 1	0	31	561	226	7	825	20.1	7.0
SOM_{5x1} 2	65	844	296	9	0	1214	29.5	5.7
SOM_{5x1} 3	387	274	24	1	0	686	16.7	5.0
SOM_{5x1} 4	524	39	5	0	0	568	13.8	4.4
Total	976	1188	938	564	448	4114	100.0	6.3
% Total	23.7	28.9	22.8	13.7	10.9	100.0		
Mean pI_{est}	4.5	5.5	6.6	8.2	9.7	6.3		

(b) $2Step_5$ Cluster

Cluster	1	2	3	4	5	Total	% Total	Mean pI_{est}
SOM_{5x1} 0	431	0	390	0	0	821	20.0	9.0
SOM_{5x1} 1	824	0	1	0	0	825	20.1	7.0
SOM_{5x1} 2	639	575	0	0	0	1214	29.5	5.7
SOM_{5x1} 3	0	686	0	0	0	686	16.7	5.0
SOM_{5x1} 4	0	125	0	78	365	568	13.8	4.4
Total	1894	1386	391	78	365	4114	100.0	6.3
% Total	46.0	33.7	9.5	1.9	8.9	100.0		
Mean pI_{est}	7.0	5.2	9.6	3.7	4.4	6.3		

As mentioned previously, SOMs should preserve the spatial relationship between clusters, i.e. clusters that share borders or are closer together in space (1D or 2D) are more closely related to each other. For this analysis, Cluster 0 should be most closely related to Cluster 1, while most different from Cluster 4. This could be observed by either the $\overline{Q}$ curves (Figure 6.13) or the mean pI_{est} values of each cluster. Meanwhile, there are no defined relationships between 2Step clusters. These relationships may be inferred from each cluster's mean pI_{est} value or by calculating the distance between clusters on the mean $\overline{Q}_{cryst}$ values.

This was indeed the case, where Cluster 0 was composed of 'Very Basic' proteins (mean pI_{est} of 9.0) and Cluster 4 is composed of 'Very Acidic' proteins (mean pI_{est} of 4.4). Table 6.15 shows the Mean±SD of the variables examined, including the pI_{est}, MW_{au}, pH_{cryst}, Q_{cryst}, $\overline{Q}_{cryst}$, σ_{cryst}, $diff_{lim}$ (Å), and the random variable. All five groups had significantly different distributions of the pI_{est} (Figure 6.14a**),** Q_{cryst} (Figure 6.14c), $\overline{Q}_{cryst}$ (Figure 6.14b), and σ_{cryst} (Figure 6.14g). Although there were significant differences in the distributions of the pI_{est}, there were overlaps between all groups (Figure 6.14a). Clusters 0 and 1 had fairly broad pI_{est} distributions, while the others are more tightly distributed. In contrast, binning by the pI_{est} resulted in no overlaps of the pI_{est}. Clearly, the SOM algorithm was picking up other differences between the $\overline{Q}$ curves. Using a portion of the $\overline{Q}$ curves (pH 4.0-10.0) separated the data differently, regardless of the unsupervised clustering method (SOMs or Two-Step Clustering).

The $\overline{Q}_{cryst}$ distributions (Figure 6.14b) were very similar to those obtained when binning proteins by their pI_{est}, where all the distributions appeared relatively normal. This was another difference between the SOMs and the other unsupervised clustering method (2Step Clustering). The most acidic and basic clusters in the 2Step clustering method displayed $\overline{Q}_{cryst}$ distributions that were very erratic (Figures 6.8b and 6.11b) due to the smaller amount of proteins in these groups. There was also an interesting peak in the $\overline{Q}_{cryst}$ distribution of Cluster 4, which had a spike near the pI_{est}. Similar to the $\overline{Q}_{cryst}$, the other measures of charge (Q_{cryst} and σ_{cryst}) also appeared relatively normal with varying degrees of noise.

While some statistical differences in the pH_{cryst} and $diff_{lim}$ distributions did exist between the SOM_{5x1} clusters, most seem quite trivial for use in the laboratory (pH_{cryst}) or determining quality ($diff_{lim}$). The clusters all had pH_{cryst} means ranging from pH 6.4-6.7. The distribution of the pH_{cryst} is shown in Figure 6.14c. Cluster 4 had more proteins that crystallized under acidic conditions, especially pH 4.0, than did the other groups, while less at pH 6.0. Again, this lack of difference in the pH_{cryst} demonstrated the utility of using $\overline{Q}$ values for prediction rather than pH values. The distribution of $diff_{lim}$ was also found to be statistically different, but lacking any great differences as all groups had an average $diff_{lim}$ of 1.9-2.0 Å. Finally, no statistical differences were observed in the random number distributions.

Table 6.15 The variable descriptors (mean and SD) for each of the SOM_{5x1} groups separated by their estimated specific charge curves from pH 4.0-10.0.

SOM_{5x1} Cluster	n		*Features*		*Observables*					Random Variable
			pI_{est}	MW_{au}	σ_{cryst}	$\overline{Q}_{cryst}$	Q_{cryst}	pH_{cryst}	$diff_{lim}$	
0	821	Mean	9.0 ^{A}	42.5 ^{AB}	85.8 ^{A}	0.3 ^{A}	11.4 ^{A}	6.7 ^{AB}	2.0 ^{AC}	0.05 ^{A}
		SD	0.9	44.7	82.6	0.3	16.6	1.3	0.4	1.01
1	825	Mean	7.0 ^{B}	66.6 ^{AC}	15.9 ^{B}	0.1 ^{B}	2.6 ^{B}	6.7 ^{A}	2.0 ^{AB}	0.03 ^{A}
		SD	0.7	75.2	55.3	0.2	20.7	1.2	0.4	0.96
2	1214	Mean	5.7 ^{C}	75.6 ^{BC}	-25.5 ^{C}	-0.1 ^{C}	-7.4 ^{C}	6.7 ^{A}	2.1 ^{B}	0.01 ^{A}
		SD	0.5	83.6	59.7	0.2	20.4	1.2	0.4	1.00
3	686	Mean	5.0 ^{D}	67.6 ^{AC}	-63.3 ^{D}	-0.2 ^{D}	-15.3 ^{D}	6.5 ^{AB}	2.0 ^{AB}	0.02 ^{A}
		SD	0.5	92.5	70.0	0.2	27.2	1.3	0.4	0.99
4	568	Mean	4.4 ^{E}	42.1 ^{C}	-140.3 ^{E}	-0.6 ^{E}	-21.3 ^{E}	6.4 ^{B}	1.9 ^{C}	0.00 ^{A}
		SD	0.5	53.9	109.6	0.5	30.2	1.4	0.4	1.03
KW Test	x^2		3551	327	2128	2185	1694	19.3	43.3	1.2
	$p<$		0.0001	0.0001	0.0001	0.0001	0.0001	0.0007	0.0001	0.879

Note: Groups labeled with different letters (A, B, C, D, or E) have significantly different distributions ($p<0.01$) as determined by a KS Test.

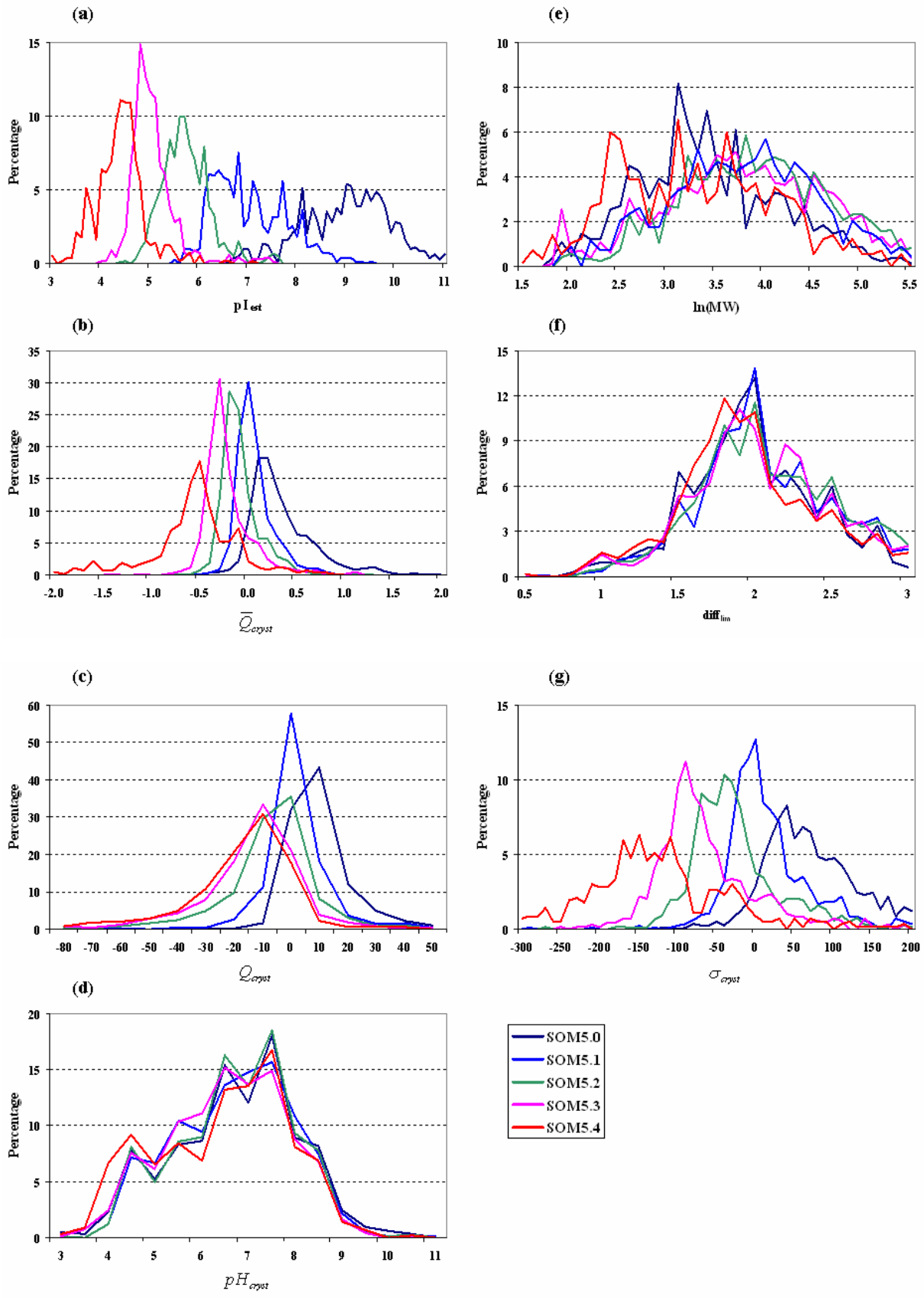

(a)
Percentage
pI est
(b)
Percentage
$\overline{Q}_{cryst}$
(c)
Percentage
Q_{cryst}
(d)
Percentage
pH_{cryst}
(e)
Percentage
ln(MW)
(f)
Percentage
$diff_{lin}$
(g)
Percentage
σ_{cryst}
SOM5.0
SOM5.1
SOM5.2
SOM5.3
SOM5.4

Figure 6.14 The SOM_{5x1} cluster distributions of the (a) pI_{est}, (b) $\overline{Q}_{cryst}$, (c) Q_{cryst}, (d) pH_{cryst}, (e) $\ln(MW_{au})$, (f) $diff_{lim}$, and (g) σ_{cryst}.

6.2.2.2 Charge Range Test

After demonstrating the ability to separate proteins into unique groups by their $\overline{Q}$ curves, the ability to correctly capture the $\overline{Q}_{cryst}$ of an independent test set was examined. The SOM_{5x1} clusters displayed an obvious improvement over Baseline (Table 6.16; Figure 6.15a). An approximate 13% increase was observed in the cumulative percentage of test set proteins bordered at all levels. When the Charge Range Test values were examined within cluster, striking differences were observed among clusters. The most acidic (SOM_{5x1} 4) and basic (SOM_{5x1} 0) protein clusters did not perform as well as the middle three clusters, although they performed similarly or better than the Baseline group. There were little differences between the SOM_{5x1} clusters, the $2Step_5$ clusters, and the pI_{est} Bins. The $2Step_6$ clusters performed slightly better (3%) than the SOM_{5x1} clusters at the Mean±0.2 and Mean±0.3 levels, but 7% better at the Mean±0.1 level. Therefore, it is recommended to use the $2Step_6$ clustering over SOM_{5x1} clusters.

Table 6.16 The Charge Range Test results for each SOM_{5x1} cluster and the Baseline values (Chapter 5).

Training Set	SOM_{5x1} 0 n=821	SOM_{5x1} 1 n=825	SOM_{5x1} 2 n=1214	SOM_{5x1} 3 n=686	SOM_{5x1} 4 n=568	Cum. % SOM_{5x1}	Baseline
Mean ±0.1	41.2	58.1	66.9	55.4	32.0	53.3	40.4
Mean ±0.2	65.8	85.2	84.4	79.7	57.4	76.3	60.1
Mean ±0.3	80.4	93.8	92.1	90.4	65.7	86.2	73.1
Test Set	SOM_{5x1} 0 n=263	SOM_{5x1} 1 n=246	SOM_{5x1} 2 n=373	SOM_{5x1} 3 n=223	SOM_{5x1} 4 n=141	Cum. % SOM_{5x1}	Baseline
Mean ±0.1	36.1	50.8	55.2	53.4	34.0	47.6	37.5
Mean ±0.2	66.9	74.0	79.6	77.6	61.7	73.4	59.6
Mean ±0.3	81.0	93.1	88.2	87.9	72.3	85.8	71.8

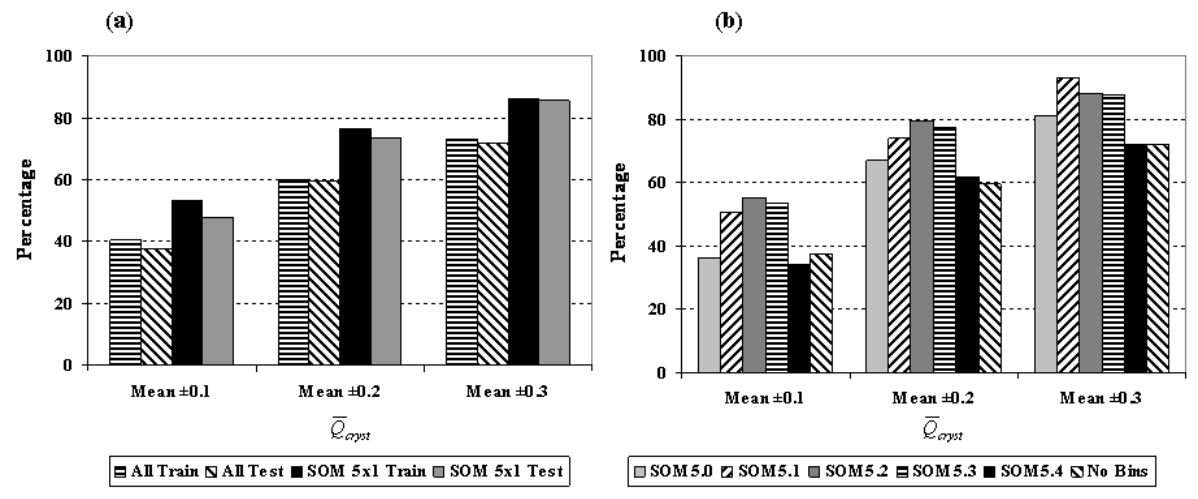


Figure 6.15 (a) The Charge Range Test results (a) between methods (cumulative percent) or (b) within grouping method, SOM_{5x1} and Baseline for proteins in the test set.

6.2.3 Supervised SOM

The second SOM method to determine the optimum number of clusters was developed for this particular purpose, the supervised SOM. The goal was to determine the maximum number of clusters, which all displayed significantly different distributions of an *Observable*, $\overline{Q}_{cryst}$ in this case. The $\overline{Q}_{cryst}$ distribution was chosen because of the differences observed between the pI_{est} bins in Section 6.1.1.2. As shown in all the previous examples, there were much larger differences observed in the $\overline{Q}_{cryst}$ distributions than in the pH_{cryst} distributions. The hypothesis was that the optimum number of clusters would perform better at modeling the $\overline{Q}_{cryst}$ distributions for prediction on an independent test set as judged by the Charge Range Test.

Initially, a 2x2 SOM (SOM_{2x2}) was created and incrementally increased in one dimension at a time until either a single non-statistically significant difference was found between any two cluster's $\overline{Q}_{cryst}$ distributions, as determined by a KS test, or the number of clusters reached 16 (Section 4.7.6). Sixteen clusters would be the maximum number of groups if the proteins were binned by their pI_{est} every 0.5 pH units from 4.0-12.0. Starting with an initial SOM_{2x2}, each

dimension of the SOM was increased one at a time starting with the x-dimension, i.e. the next SOM investigated would be SOM_{3x2}.

Although the initial use of the GSOM algorithm in Chapter 3 seemed promising, the implementation with the new training set proved otherwise. The GSOM applied to the PDB version #107 data set (n = 11,518) had a size of 61 clusters, which was large, but manageable. The attempts on the non-redundant data set ($nrPDB_{10.04.05}$) produced 993 clusters using the same parameters as in the SOM_{5x1}. The effort to analyze 1000 clusters would be quite time consuming. One difference between the analyses was that the current input *Feature* vector only used the $\overline{Q}$ curve from pH 4.0-10.0, while the initial analysis used the estimated titration curve from 1.0-15.0. The number of epochs (passes through the data) was then decreased in an attempt to decrease the overall size, but this had little to no effect on the overall size (1023 clusters). Although there is the possibility that there are actually several hundred distinct clusters in this small subsection of the protein universe, this method was abandoned for the time being.

The supervised SOM algorithm proved more manageable than the dynamic SOM. After some initial SOM analysis, it was observed that if the *x*-dimension increased and the *y*-dimension stayed at 2, no proteins were found in the 2nd *y*-dimension, creating a one-dimensional SOM. For example, a 4x2 SOM could have 8 possible clusters to populate. As it turned out, one dimension was not populated, resulting in only 4 populated clusters. Therefore, the *x*-dimension of the SOM was allowed to increase above 8, where a SOM_{8x2} would theoretically allow for sixteen groups. With the *y*-dimension remaining at 2, the *x*-dimension was allowed to increase to a maximum value of 16. Therefore, the strategy was altered so the SOM size could increase to 16x2, because the 2nd dimension would not be populated, resulting in only 16 clusters. Additionally, only minor differences were observed between the cluster assignments (2% of the proteins; 80/4114) in the one dimensional and two-dimensional SOM cluster assignments, such as between a 5x1 and 5x2 SOM.

Table 6.17 demonstrated that the 'optimum' SOM dimension was a 14x2 SOM (SOM_{14x2}). When a 15x2 SOM (SOM_{15x2}) was produced, there was a non-significant pairwise difference between clusters as determined by a KS test. When the 2nd dimension began to populate in the SOM_{3x3}, non-significant differences in $\overline{Q}_{cryst}$ distributions were immediately observed between clusters. One possible explanation was that a one-dimensional (1D) vector was used as input.

The differences in the 1D vector may not be able to be captured in the second dimension. Thus, the 1D vector was better represented in 1D space, which also allowed for an easier interpretation.

The SOM_{14x2} splits each SOM_{5x1} cluster into 3-4 additional clusters each with overlaps between the adjacent SOM_{5x1} clusters (Table 6.18). The 'Very Acidic' proteins are grouped more towards SOM_{14x2} Cluster 0, while the 'Very Basic' clusters are more concentrated toward SOM_{14x2} Cluster 13. The terminal clusters (0-1 and 12-13) had the fewest number of proteins and tended to be of smaller size with the most extreme pI_{est} values and the widest $\overline{Q}_{cryst}$ distributions (Figure 6.17). These terminal clusters also contained proteins with better resolution, which may be related to their small size. Each SOM_{14x2} cluster displayed significantly different distributions of pI_{est} and $\overline{Q}_{cryst}$, although some clusters may have the same mean value (Table 6.19). Many differences were also observed in the Q_{cryst} distributions, with overlapping distributions occurring towards each terminal end of the SOM_{14x2}. SOM_{14x2} Clusters 0 and 1 had significantly lower pH_{cryst} distributions than did several other clusters. The MW_{au} and $diff_{lim}$ distributions also displayed a complex pattern of differences.

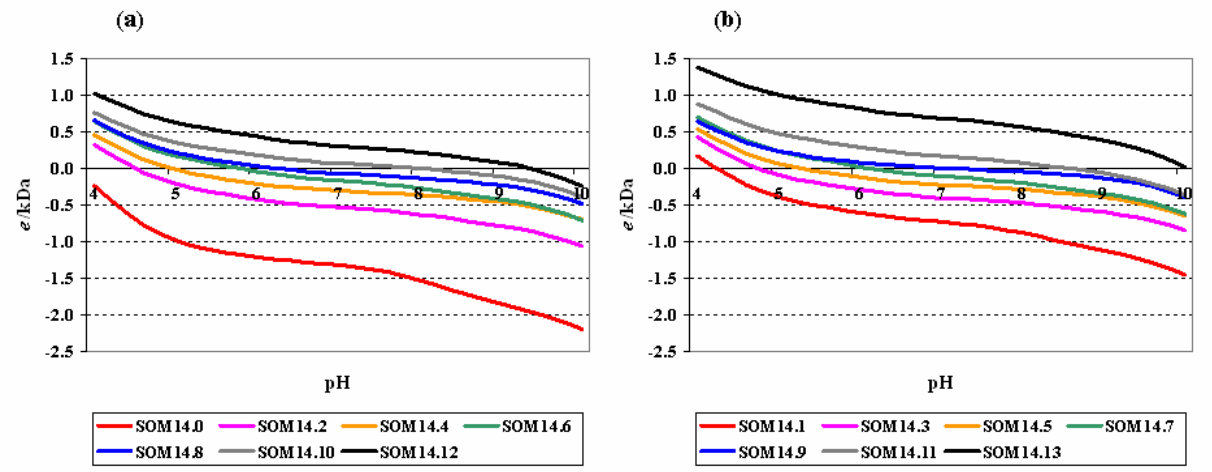


Figure 6.16 The mean $\overline{Q}$ curves for each SOM_{14x2} cluster.

Table 6.17 Determining the optimum SOM dimension using the supervised SOM algorithm in Section 4.7.6. For comparison, the number of non-significant (NS) pairwise differences in the Q_{cryst} and pH_{cryst} distributions between SOM clusters was also reported.

SOM Dimensions (x*y)	# of Possible Clusters	# of Empty Clusters	# Possible Pairwise Comparisons	# of Clusters - NS Pairwise		
				$\overline{Q}_{cryst}$	Q_{cryst}	pH_{cryst}
2x2	4	3		-	-	-
3x2	6	3	3	0	0	3
4x2	8	4	6	0	0	5
5X2	10	5	10	0	0	8
6x2	12	6	15	0	1	12
7x2	14	7	21	0	1	16
8x2	16	8	28	0	1	22
9X2	18	9	36	0	3	30
10x2	20	10	45	0	3	38
11x2	22	11	55	0	6	45
12x2	24	12	66	0	6	56
13x2	26	13	78	0	9	68
14x2	**28**	**14**	**91**	**0**	**6**	**78**
15x2	30	15	**104**	**1**	10	91
3x3	9	0	36	1	3	35
4x3	12	0	66	1	6	63

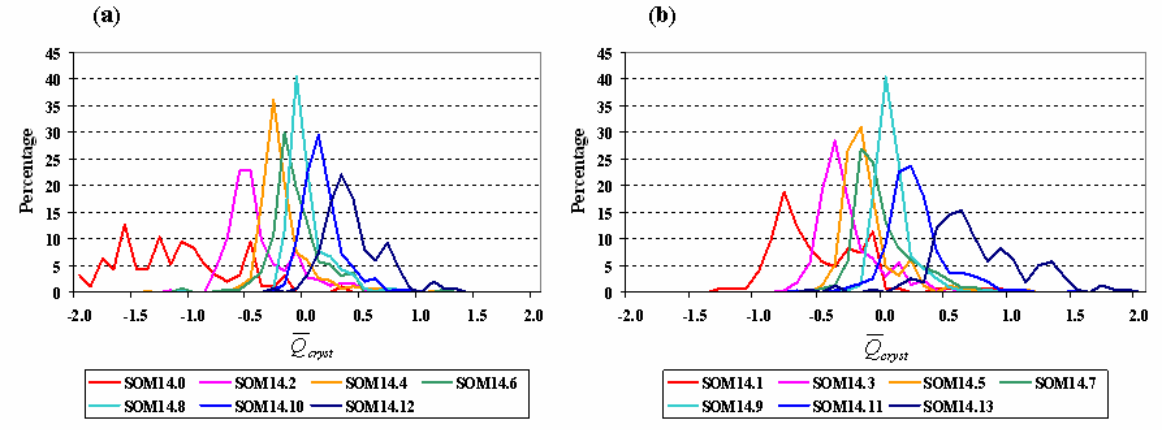


Figure 6.17 The distribution of the $\overline{Q}_{cryst}$ for each SOM_{14x2} cluster in the training set.

Table 6.18 Cross-tabulation of the SOM_{5x1} and SOM_{14x2} Clusters.

SOM_{14x2} Cluster	SOM_{5x1} Cluster					Total	% Total
	0	1	2	3	4		
0	0	0	0	0	95	95	2.3
1	0	0	0	0	122	122	3.0
2	0	0	0	0	245	245	6.0
3	0	0	0	165	106	271	6.6
4	0	0	5	381	0	386	9.4
5	0	0	295	112	0	407	9.9
6	0	0	354	28	0	382	9.3
7	0	51	355	0	0	406	9.9
8	0	216	201	0	0	417	10.1
9	0	350	4	0	0	354	8.6
10	178	202	0	0	0	380	9.2
11	336	6	0	0	0	342	8.3
12	150	0	0	0	0	150	3.6
13	157	0	0	0	0	157	3.8
Total	821	825	1214	686	568	4114	100.0
% Total	20.0	20.1	29.5	16.7	13.8	100.0	

Table 6.19 The variable descriptions for each of the SOM_{14x2} clusters.

SOM_{14x2} Cluster		*Features*		*Controllables*					Random Number
		pI_{est}	MW_{au}	σ_{cryst}	$\overline{Q}_{cryst}$	Q_{cryst}	pH_{cryst}	$diff_{lim}$	
0	Mean	3.8 ^{A}	22.3 ^{A}	-270.6 ^{A}	-1.2 ^{A}	-25.3 ^{A}	6.4 ^{AB}	1.8 ^{A}	-0.06 ^{A}
	SD	0.3	23.4	114.7	0.5	28.7	1.4	0.4	1.05
1	Mean	4.3 ^{B}	28.7 ^{B}	-132.8 ^{B}	-0.5 ^{B}	-17.9 ^{BC}	6.1 ^{A}	1.9 ^{AB}	0.01 ^{A}
	SD	0.4	25.0	105.7	0.4	25.5	1.5	0.4	1.00
2	Mean	4.6 ^{C}	46.9 ^{C}	-109.6 ^{C}	-0.4 ^{C}	-19.6 ^{BC}	6.5 ^{AC}	2.0 ^{BCF}	-0.01 ^{A}
	SD	0.4	54.7	83.4	0.3	28.9	1.4	0.4	1.03
3	Mean	4.8 ^{D}	66.8 ^{D}	-88.1 ^{D}	-0.3 ^{D}	-21.3 ^{AB}	6.6 ^{BC}	2.0 ^{CD}	0.02 ^{A}
	SD	0.4	100.0	70.6	0.2	33.3	1.3	0.5	1.02
4	Mean	4.9 ^{E}	70.4 ^{DE}	-63.5 ^{E}	-0.2 ^{E}	-15.1 ^{C}	6.6 ^{BC}	2.0 ^{BDE}	0.08 ^{A}
	SD	0.3	89.6	65.5	0.2	27.1	1.2	0.4	0.98
5	Mean	5.3 ^{F}	74.3 ^{DE}	-44.7 ^{F}	-0.2 ^{F}	-12.1 ^{D}	6.7 ^{BC}	2.0 ^{BDEF}	0.06 ^{A}
	SD	0.3	84.4	60.0	0.2	24.6	1.2	0.5	1.03
6	Mean	5.8 ^{G}	68.3 ^{DE}	-30.5 ^{G}	-0.1 ^{G}	-8.2 ^{E}	6.7 ^{C}	2.1 ^{CDE}	-0.07 ^{A}
	SD	0.5	73.4	70.4	0.3	20.4	1.3	0.5	0.97
7	Mean	6.1 ^{H}	75.0 ^{DE}	-10.3 ^{H}	0.0 ^{H}	-4.9 ^{F}	6.7 ^{BC}	2.1 ^{CDE}	0.02 ^{A}
	SD	0.5	90.5	64.7	0.2	23.6	1.2	0.4	1.00
8	Mean	6.2 ^{I}	77.0 ^{E}	-2.8 ^{I}	0.0 ^{I}	-1.4 ^{G}	6.7 ^{BC}	2.1 ^{D}	0.01 ^{A}
	SD	0.5	78.7	52.6	0.2	18.9	1.2	0.4	0.93
9	Mean	7.1 ^{J}	66.6 ^{DE}	13.7 ^{J}	0.0 ^{J}	3.0 ^{H}	6.7 ^{BC}	2.0 ^{BCF}	0.01 ^{A}
	SD	0.7	81.5	42.5	0.1	16.6	1.2	0.4	0.97
10	Mean	8.0 ^{K}	59.2 ^{D}	35.6 ^{K}	0.1 ^{K}	6.6 ^{I}	6.7 ^{C}	2.0 ^{BCF}	0.06 ^{A}
	SD	0.8	55.5	51.0	0.2	14.4	1.2	0.4	1.01
11	Mean	8.6 ^{L}	45.3 ^{C}	60.8 ^{L}	0.2 ^{L}	9.3 ^{J}	6.7 ^{C}	2.0 ^{BCF}	0.10 ^{A}
	SD	0.8	46.7	60.8	0.2	15.2	1.2	0.4	1.02
12	Mean	9.4 ^{M}	35.0 ^{B}	100.3 ^{M}	0.4 ^{M}	13.5 ^{K}	6.5 ^{AC}	2.0 ^{BCF}	0.02 ^{A}
	SD	0.6	41.8	65.3	0.3	17.2	1.4	0.5	1.02
13	Mean	10.0 ^{N}	24.9 ^{B}	175.9 ^{N}	0.7 ^{N}	18.2 ^{L}	6.8 ^{BC}	1.9 ^{AF}	-0.10 ^{A}
	SD	0.7	18.5	97.4	0.4	19.9	1.5	0.4	1.00
KW Test	x^2	3742	479	2258	2319	1773	33.2	81.7	13.0
(df=13)	$p <$	0.0001	0.0001	0.0001	0.0001	0.0001	0.002	0.0001	0.447

Note: Groups labeled with different letters (A-N) have significantly different distributions ($p<0.01$) as determined by a KS Test.

6.2.3.1 Charge Range Test

The next step was to examine each cluster's Charge Range Test results and compare the cumulative results with Baseline and the 2Step clustering methods. Large improvements in the Charge Range Test values were seen in all levels when compared to Baseline (Table 6.20; Figure 6.18a). The cumulative results show a +20% at the Mean±0.1 and Mean±0.2 levels for the test set proteins. These were followed by a +15% increase at the Mean±0.3.

When the individual differences within clusters were examined, there was a striking increase in accuracy at the low $\overline{Q}$ levels, such as the Mean±0.1 (Figure 6.18b). For example, at the Mean±0.1 $\overline{Q}$ level, four clusters had accuracy above 65%. This was an improvement of over 30% in accuracy over Baseline. While some individual clusters at the Mean±0.1 level in other methods produced 60% values, none exceeded 60%. Again, the clusters at either end of the 14x2 SOM had lower values, which were composed of 'Very Acidic' or 'Very Basic' proteins. However, these terminal clusters still produced accuracy levels comparable to Baseline (~35%).

Table 6.20 The Charge Range Test results for each SOM_{14x2} cluster and the Baseline (Chapter 5) values.
(a) Training Set

SOM_{14x2} Clusters	n	Mean ±0.1	Mean ±0.2	Mean ±0.3
0	95	25.3	37.9	47.4
1	122	18.9	39.3	65.6
2	245	38.0	64.9	82.9
3	271	53.5	78.6	87.1
4	386	64.0	86.8	92.0
5	407	75.4	85.5	89.9
6	382	62.0	78.5	87.7
7	406	45.6	78.6	89.2
8	417	71.0	89.2	94.2
9	354	82.5	90.7	95.5
10	380	69.7	86.8	93.2
11	342	64.0	80.7	86.8
12	150	47.3	70.0	86.7
13	157	31.2	54.1	72.6

Cum. % SOM_{14x2}	4,114	59.6	78.9	87.7
Baseline	4,114	40.4	61.8	73.1

(b) Test Set

SOM_{14x2} Clusters	n	Mean ±0.1	Mean ±0.2	Mean ±0.3
0	14	28.6	35.7	50.0
1	21	23.8	47.6	71.4
2	67	46.3	67.2	85.1
3	99	45.5	74.7	85.9
4	116	64.7	81.9	88.8
5	126	68.3	86.5	94.4
6	116	44.8	75.9	86.2
7	134	44.8	76.1	85.8
8	133	63.2	85.7	92.5
9	92	73.9	84.8	89.1
10	109	61.5	76.1	90.8
11	114	65.8	80.7	88.6
12	70	44.3	58.6	85.7
13	35	42.9	65.7	77.1
Cum. % SOM_{14x2}	1,246	56.0	77.0	87.7
Baseline	1,246	37.5	59.6	71.8

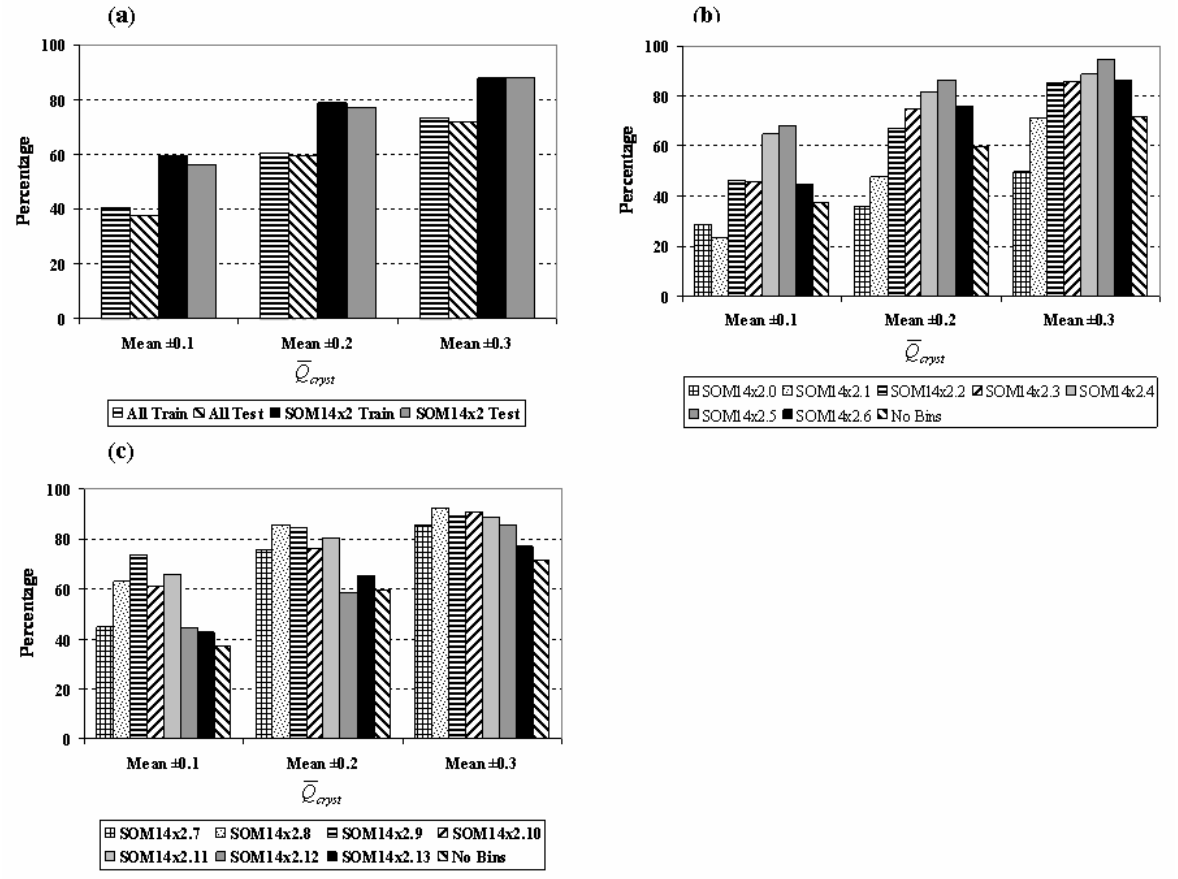


Figure 6.18 The Charge Range Test results (a) between methods (cumulative percent) or (b, c) within grouping method, SOM_{14x2} and Baseline for proteins in the test set.

6.2.4 Summary of Unsupervised Clustering Results

Similar to binning, the results in this section are based on the assumption that the Q-related *Observables* can be used as proxy variables for the pH_{cryst}. Using unsupervised clustering on a protein's $\overline{Q}$ curve, demonstrated improved benefits over using all proteins as one group (Baseline; Chapter 5). These methods increased the cumulative predictive percentage of $\overline{Q}_{cryst}$ by 10-20% over Baseline. An even more significant increase could often be found when examining the individual differences within each clustering technique. This demonstrated that there are groups of proteins in the known protein space that behave similarly in crystallization attempts. However, clusters with more extreme $\overline{Q}$ curves, as judged by very high or very low

pI_{est} values, did not perform as well in the Charge Range Test. Using the findings presented here should increase the number of proteins that yield crystals. This could be done by decreasing the sampling at the pH values where proteins are not likely to crystallize and increasing the sampling in the pH search space where proteins are likely to crystallize.

These results also reaffirmed the observation in Chapter 5 that knowledge of a protein's $\overline{Q}$ values over a given pH range could be used to potentially increase the probability of obtaining a crystal suitable for diffraction studies. Few differences were found in the distribution of pH_{cryst} between clusters within any method of grouping, binning or unsupervised clustering. Many more differences were found with the Q_{cryst}, $\overline{Q}_{cryst}$, and σ_{cryst} distributions. Additionally, these results demonstrated that there is knowledge present within a protein's sequence that can be used for the selection of solution pH ranges for initial crystallization screen design.

6.3 MODELING $\overline{Q}_{cryst}$ DISTRIBUTIONS WITH GAUSSIANS

In this section, attempts were made to model the $\overline{Q}_{cryst}$ distributions with Gaussians. If the $\overline{Q}_{cryst}$ distributions could be fit by Gaussians, this would allow one to easily compare the $\overline{Q}_{cryst}$ distributions between groups and methods, which would permit the identification of the 'best' method of grouping proteins by similarity. Knowledge of the $\overline{Q}_{cryst}$ distributions would also allow for simple prediction of probabilities over ranges of $\overline{Q}$ values, $P(\overline{Q} = \overline{Q}_{cryst} \mid data)$.

Equation 6.1

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

, where x = observed $\overline{Q}_{cryst}$ value, μ= mean $\overline{Q}_{cryst}$ value, and σ= SD (observed or best-fit SD)

The $\overline{Q}_{cryst}$ distributions in Chapter 5 and the previous sections (6.1-6.2) appeared as though they could be fit relatively well with Gaussians (Equation 6.1). However, the groups with the most extreme $\overline{Q}$ curves (lowest and highest mean pI_{est} values) often displayed erratic

$\overline{Q}_{cryst}$ distributions. In this section, attempts were made to model the $\overline{Q}_{cryst}$ distributions of two grouping methods with Gaussians.

The best fit Gaussian was chosen by minimizing the residual sum of square error (RSS; Equation 6.2) between the observed values (Y_i) and the Gaussian predicted $\overline{Q}_{cryst}$ values ($\hat{Y}_i$). The value of the standard deviation (SD) was varied by 0.01 $\overline{Q}_{cryst}$ units in order to determine the SD with the minimum RSS.

Equation 6.2. $RSS = \sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)$

After determining the best-fit Gaussian, the number of proteins within 1 SD of the best-fit curve for both the training and test sets was calculated. Then, in order to compare all groups, a common SD for all groups was chosen based on using the best fit SD of all proteins (rounding off to tenths). Gaussians (Equation 6.1) were initially fit to all training set proteins (Baseline; Section 6.3.1). After examining the calculated and observed distributions for all proteins, the $\overline{Q}_{cryst}$ distributions for each $2Step_6$ cluster was calculated (Section 6.3.2). These values then were compared to Baseline (Section 6.3.1).

6.3.1 Baseline (All Proteins)

Initially, a Gaussian was fit to the Baseline $\overline{Q}_{cryst}$ distribution using the observed mean and SD of -0.06 and 0.39, respectively (Figure 6.19). The Gaussian calculated based upon the observed SD fit fairly well, RSS of 0.0056. However, when the SD was allowed to vary, the best-fit SD was 0.30, with a RSS of 0.0017. The Gaussian with the best-fit SD fit the data very well (Table 6.19), with 72.5% of the training set proteins falling within 1 SD of the observed mean $\overline{Q}_{cryst}$ value. A similar amount of the test set proteins fell within 1 SD of the mean, 73%. These values, ~73%, were slightly more than would be expected from the portion of cases found within 1 SD of a normal distribution (67%). There were also slightly more proteins whose $\overline{Q}_{cryst}$ value was 1+ SD above the mean (~17%) than were 1+ SD below the mean $\overline{Q}_{cryst}$ value (~11%).

After demonstrating that a Gaussian could be fit to the $\overline{Q}_{cryst}$ distribution of all training set proteins, the ability to fit a Gaussian to the clustering results was examined. The two-step

clustering method with six clusters ($2Step_6$; Section 6.2.1.3) was used as the example for comparison.

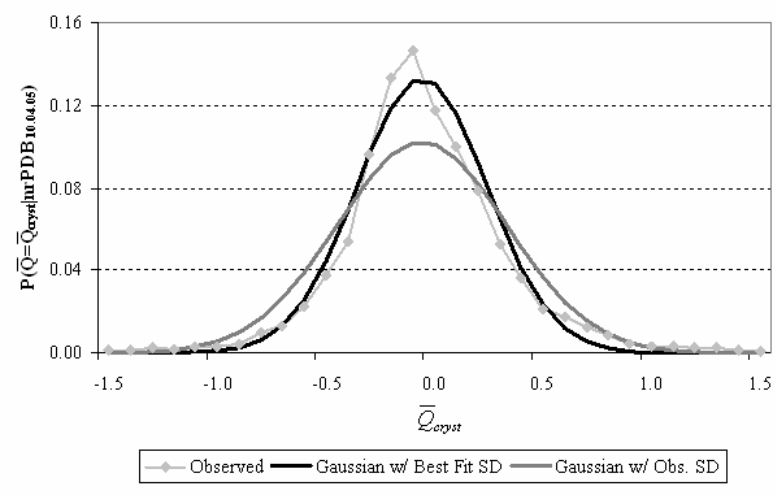


Figure 6.19 The Baseline $\overline{Q}_{cryst}$ distribution along with the Gaussians calculated from the observed and best-fit SD.

Table 6.21 The Baseline $\overline{Q}_{cryst}$ range for the best-fit Gaussian of the training set proteins and the percentage of proteins within 1, 1-2, or 2+ SD of the mean $\overline{Q}_{cryst}$ for both the training and test sets.

Grouping	Low $\overline{Q}_{cryst}$ Value	High $\overline{Q}_{cryst}$ Value	% Within Range (Train)	% Within Range (Test)
2+ SD Below	-	-0.8	3.4	1.8
1-2 SD Below	-0.7	-0.5	7.3	7.1
Within 1 SD	-0.4	0.2	72.5	73.0
1-2 SD Above	0.3	0.5	11.0	12.5
2+ SD Above	0.6	-	5.8	5..6

6.3.2 2Step Clustering 6 Clusters

In the previous section, a Gaussian was fit to the $\overline{Q}_{cryst}$ distribution of all proteins in the training set. In this section, the Baseline Gaussian was compared to Gaussians that were fit to each

cluster of the $2Step_6$ results, the optimal number of clusters as determined by 2Step clustering; Section 6.2.1.3.

The observed $\overline{Q}_{cryst}$ distributions of $2Step_6$ Clusters 1 and 6 were not as smooth as the other clusters. These clusters were also the ones with the fewest number of proteins within them, representing the proteins with the extreme pI_{est} values. The observed frequency distributions along with the calculated Gaussians from the observed and best-fit SD are pictured in Figure 6.20a-f. Figure 6.20g shows all of the $2Step_6$ best-fit Gaussians in one figure for comparison.

Table 6.22a shows the difference between the observed SD and the best-fit SD. There are varying degrees of differences. For instance, $2Step_6$ Cluster 1 had an observed and best-fit SD that was the same. The other best-fit SDs differed from the observed SDs by 0.03-0.07 $\overline{Q}_{cryst}$ units. Next the amount of proteins that fell within 1 SD of the best-fit Gaussian were calculated and compared to the Baseline Gaussian in Section 6.3.1.

The amount of proteins within 1 SD of the best-fit Gaussian ranged from 68-86% of the proteins within the cluster Table 6.22a. Overall, 81% of the proteins clustered by the 2Step algorithm fell within 1 SD of the best-fit Gaussians. This was an overall improvement of 9% over the best-fit Gaussian on the Baseline group in the previous section, which had 72.5% of the proteins within 1 SD of the best-fit Gaussian. However, this may be a little misleading, because the best-fit SD of all proteins was 0.30, while those in the $2Step_6$ distributions ranged from 0.16-0.52. This was the same problem with the CI_{50} test that led to the development of the Charge Range Test. In order for a more fair comparison, the amount of proteins within a common SD of ±0.2 or ±0.3 was calculated for each $2Step_6$ cluster and the Baseline group.

When a common SD was used for all clusters, an even more pronounced difference was observed between the $2Step_6$ clusters and the Baseline group (Table 6.22b). An overall improvement of 15-18% was observed for the $2Step_6$ clustering method. Several of the $2Step_6$ clusters with the tight distributions had ~80% and ~90% of their proteins within ±0.2 and ±0.3 *e*/kDa units of the group mean, respectively. Again the $2Step_6$ clusters with the more extreme $\overline{Q}$ curves, fewer proteins, and those with a larger best-fit SD (Clusters 1 and 6) did not perform as well as the other clusters. However, this was expected because these clusters involved proteins with more extreme $\overline{Q}$ curves and fewer representative cases.

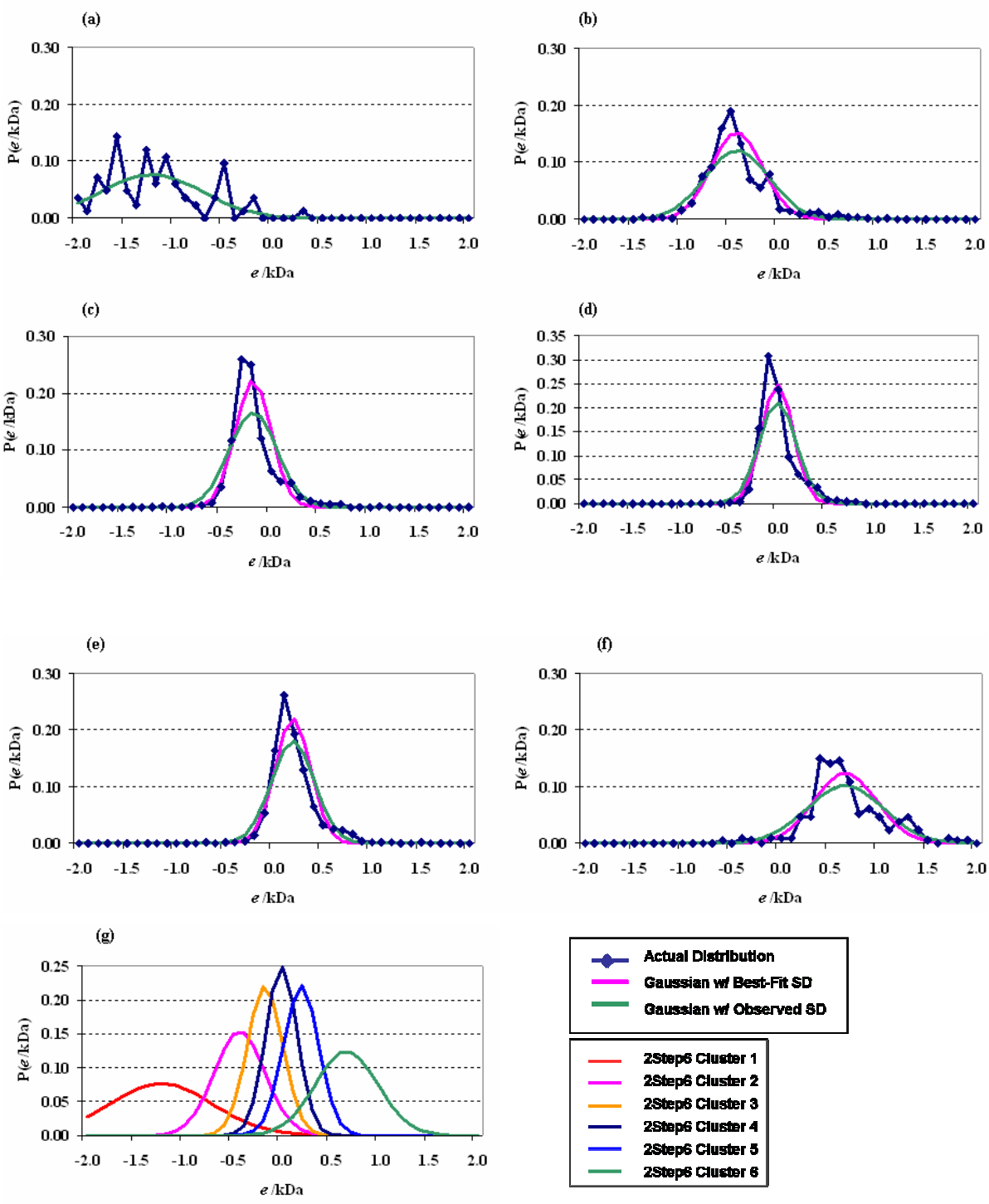


Figure 6.20 Modelling each $2Step_6$ cluster with a Gaussian (a-f). The best-fit $\overline{Q}_{cryst}$ Gaussian for each cluster was then shown in (g).

Table 6.22 The mean , observed SD, and best-fit SD of $\overline{Q}_{cryst}$ for each 2Step$_6$ cluster. The percentage of proteins within 1 SD of the best-fit $\overline{Q}_{cryst}$ Gaussian was calculated for (a) each Cluster or (b) a common SD based on the best-fit $\overline{Q}_{cryst}$ Gaussian on all proteins (Baseline).

(a)

Cluster	N	Mean $\overline{Q}_{cryst}$	Observed SD	Best-Fit SD	% Within 1 SD of Best-Fit
2Step$_6$ 1	83	-1.22	0.52	0.52	67.5
2Step$_6$ 2	489	-0.44	0.33	0.26	77.9
2Step$_6$ 3	1229	-0.18	0.24	0.18	81.0
2Step$_6$ 4	1185	-0.01	0.19	0.16	86.1
2Step$_6$ 5	915	0.19	0.22	0.18	81.5
2Step$_6$ 6	213	0.66	0.39	0.32	70.4
Cumulative Total	4114	-	-	-	81.4
Baseline	4114	-0.06	0.39	0.30	72.5

(b) Percent of proteins within a common SD of ±0.2 or ±0.3

2Step$_6$ Cluster	Common SD	% Within Common SD (0.2)	Common SD	% Within Common SD (0.3)
2Step$_6$ 1	±0.2	37.3	±0.3	45.8
2Step$_6$ 2	±0.2	60.7	±0.3	77.9
2Step$_6$ 3	±0.2	81.0	±0.3	89.2
2Step$_6$ 4	±0.2	86.1	±0.3	93.3
2Step$_6$ 5	±0.2	81.5	±0.3	90.2
2Step$_6$ 6	±0.2	50.7	±0.3	70.4
2Step$_6$ Total	±0.2	77.7	±0.3	87.4
Baseline	±0.2	59.3	±0.3	72.5

6.3.3 Modeling Summary

The two examples presented above demonstrated the ability to accurately model the $\overline{Q}_{cryst}$ distributions with Gaussians. The Gaussians were originally calculated based upon the mean and standard deviation (SD) of the training set's observed $\overline{Q}_{cryst}$ frequency distribution. Next, the SD was allowed to vary to determine the best-fit Gaussian by minimizing the residual sum of squares. In all but one case examined (6/7) the best-fit Gaussian had a lower SD (0.02-0.07 *e*/kDa units) than that calculated from the frequency distribution. Similar to earlier sections, a significant improvement over Baseline was observed when proteins were grouped by their $\overline{Q}$ curves. The resulting Gaussians could then be used to eliminate the erratic peaks (spikes) and valleys in the frequency distributions of the proteins with the more extreme estimated titration curves. This should result in a more accurate prediction of the $\overline{Q}_{cryst}$.

6.4 CHAPTER SUMMARY

In Chapter 5**,** two possible interpretations of the Q-related *Observables* were discussed. One interpretation was that, because these Q-related *Observables* were derived from both the protein *Features* and the *Primary Observable* (pH_{cryst}), there should be a high correlation between the variables. The other interpretation, although inconclusive, was that the Q-related *Observables* ($\overline{Q}_{cryst}$ and σ_{cryst}) could be used as proxy variables for the *Primary Observable* (pH_{cryst}). The results of this chapter are based on the second interpretation. Therefore, if the first interpretation is correct, the results of this section are meaningless.

In Chapter 5**,** the distributions of both the $\overline{Q}_{cryst}$ and σ_{cryst} were shown to be relatively normal, while displaying a high correlation to both *Features* and the *Primary Observable*. Several methods were also discussed on how to use this information for estimating a pH range that has a higher probability in generating crystals. Then, it was hypothesized that there are groups of proteins that can be identified, which display similar crystallization behavior. By

identifying these subgroups, the accuracy of capturing the $\overline{Q}_{cryst}$ within a narrow range of $\overline{Q}$ values can be improved over using all proteins as one group, Baseline. In this chapter, several methods were used to group proteins by similarity, both supervised and unsupervised. The first methods examined were supervised, based upon binning the proteins by their MW_{au}, pI_{est}, or randomly grouping the proteins (Section 6.1.1). Binning by the natural log transformed asymmetric unit molecular weight, $\ln(MW_{au})$, had no additional effect on the $\overline{Q}_{cryst}$ prediction compared to Baseline. Similarly, the randomly picked groups of proteins resulted in a similar level of accuracy to Baseline (Sections 6.1.2-6.1.3). However, when proteins were separated by their pI_{est}, a 12-14% improvement was observed in the Charge Range Test predictions of the $\overline{Q}_{cryst}$ on an independent test set (Table 6.5). This improvement was not equally split among all the pI_{est} groups. The models developed on proteins with moderate pI_{est} values, 5.0-9.0, performed better than those proteins with the extreme pI_{est} values, $pI_{est} < 5.0$ or $pI_{est} > 9.0$. One reason for this is the shape of the distributions. Those subgroups with a moderate pI_{est} are highly populated with a relatively normal distribution. However, there are fewer cases of proteins in the extreme pI_{est} groups and the $\overline{Q}_{cryst}$ distributions are very sporadic. A more detailed look into the crystallization conditions of these proteins may be warranted to determine if other molecules are present in the solution to neutralize the charge.

Because the pI_{est} represented only one point along the $\overline{Q}$ curve, unsupervised clustering techniques (Section 6.2) were examined for grouping proteins with similar $\overline{Q}$ curves from pH 4.0-10.0 in 0.2 pH unit intervals. Each clustering algorithm examined, Two-Step Clustering (2Step) and Self-Organizing Maps (SOMs), could have the number of clusters specified apriori or have the algorithms determine the 'correct' number of clusters. Initially, five clusters were examined and compared to pI_{est} binning, which resulted in five groups. Similar to pI_{est} binning, these two algorithms were able to improve prediction accuracy over the Baseline group. However, five clusters resulted in no improvement over the pI_{est} binning, as judged by the Charge Range Test. When the algorithms were allowed to select the number of clusters for the training set, six were chosen for 2Step and fourteen for SOMs. These clusters resulted in a slight improvement of 4-5 percent over pI_{est} binning (Table 6.23). Thus, the unsupervised clustering methods, which used a portion of the $\overline{Q}$ curve from pH 4.0-10.0, should be used over pI_{est}

binning for selecting $\overline{Q}$ ranges that have a higher probability in generating crystals. These ranges can then be translated into more specific pH ranges.

Table 6.23 The Charge Range Test summaries for (a) the training or (b) test set proteins.

(a)

$\overline{Q}_{cryst}$	Baseline	ln(MW_{au}) Bins	pI_{est} Bins	Random Clusters	$2Step_5$	$2Step_6$	SOM_{5x1}	SOM_{14x2}
Mean ±0.1	40.4	40.3	54.0	40.4±2.4	51.2	57.2	53.3	59.6
Mean ±0.2	60.1	60.1	74.6	60.1±1.8	74.0	77.7	76.3	78.9
Mean ±0.3	73.1	73.1	84.8	73.1±1.5	86.9	87.4	86.2	87.7

(b)

$\overline{Q}_{cryst}$	Baseline	ln(MW_{au}) Bins	pI_{est} Bins	Random Clusters	$2Step_5$	$2Step_6$	SOM_{5x1}	SOM_{14x2}
Mean ±0.1	37.5	36.8	49.0	37.5	49.4	54.4	47.6	56.0
Mean ±0.2	59.6	59.7	73.0	59.6	72.5	76.7	73.4	77.0
Mean ±0.3	71.8	71.3	83.1	71.8	87.2	87.6	85.8	87.7

Note: The numbers in the cells represent the percentage of correct $\overline{Q}_{cryst}$ within a range of the group mean $\overline{Q}_{cryst}$.

In the final section, attempts were made to model the $\overline{Q}_{cryst}$ distributions with Gaussians (Section 6.3). Gaussians were shown to fit the $\overline{Q}_{cryst}$ distributions very well for both the Baseline group and each $2Step_6$ cluster. Again, the clusters with extreme mean pI_{est} values could not be represented by a Gaussian as well. Using Gaussians will remove the peaks and valleys in the distributions and allow the prediction of the $\overline{Q}_{cryst}$ over any given $\overline{Q}$ range.

In conclusion, it appears that 'similar' proteins can be identified, which crystallize under similar $\overline{Q}$ ranges. Here, 'similarity' among proteins was based upon the $\overline{Q}$ curves, either one point along the curve (pI_{est}) or a section of the curve (pH 4.0-10.0). The belief is that using the successful solution conditions from 'similar' proteins should increase the probability of generating crystals suitable for diffraction studies over the current one-size-fits-all approach. There may be other *Hidden Features* that can further breakdown the proteins into groups with tighter $\overline{Q}_{cryst}$ distributions, some of which are discussed in Chapter 9. In the next chapter, an

example application using three target proteins undergoing crystallization attempts in the Rosenberg Lab (Department of Biological Sciences, University of Pittsburgh) is presented.

7.0 EXAMPLE OF APPLICATION

In this chapter, annotated examples are presented for three proteins, which are currently undergoing crystallization attempts in the Rosenberg Lab (Department of Biological Sciences) at the University of Pittsburgh. These proteins are α-Synuclein, Neuroblastoma apoptosis-related RNA binding protein (NAPOR-1), and Ultraviolet-damaged DNA-binding protein (UV-DDB). Currently, these proteins have failed to yield well-formed crystals suitable for X-ray diffraction studies. The steps involved in the process of selecting the most probable pH regions to set up experiments and recommendations are presented.

With the Protein Sequence-Properties Evaluation (PSPE) Framework developed in Chapter 4, these three proteins provided an opportunity for a real-world application. In Section 7.1, the protein sequences were examined and various parameters calculated that are used by the PSPE framework. These proteins were then matched to 'similar' proteins that have been successfully crystallized and deposited in the PDB as discussed in Chapter 6 (Section 7.2). After determining the best-matching group of proteins, prior probabilities were generated for the range of the $\overline{Q}$ values, $P(\overline{Q} = \overline{Q}_{cryst} \mid Group, nrPDB_{10.04.05})$, that are most likely to result in the formation of a crystal (Section 7.3). These $P(\overline{Q} = \overline{Q}_{cryst} \mid Group, nrPDB_{10.04.05})$ probabilities were then translated back into the pH space, $P(pH = pH_{cryst} \mid \overline{Q} = \overline{Q}_{cryst}, Group, nrPDB_{10.04.05})$, where the knowledge could be used directly in the lab. With this knowledge in hand, a researcher should theoretically increase the probability of growing a well-ordered protein crystal. This can be done by increasing the amount of experiments in pH areas more likely to generate crystals, while decreasing the experiments in areas not likely to grow crystals.

7.1 EXAMPLE PROTEINS

All three of these proteins have implications for human health in development or disease. α-Synuclein is found primarily in the pre-synaptic nerve terminals in the brain. This 14.5 kDa protein has been associated with plaque formation in neurodegenerative diseases, such as Alzheimer's (Jakes et al., 1994; Giasson et al., 2000) and Parkinson's disease (Giasson et al., 2000; Shimura et al., 2001; Zarranz et al., 2004).

NAPOR-1, which can also be found in the brain, has been linked to the early development and differentiation of the central nervous system. NAPOR-1 is involved in post-transcriptional regulation, i.e., turning on/off other gene products. This protein was induced during apoptosis (programmed cell death), which is an important part of the nervous system development (Choi et al., 1999; Levers et al., 2002).

The UV-DDB complex is composed of two chains, one small (48 kDa) and one large (127 kDa) subunit. People with an inheritable defect of this gene, called xeroderma pigmentosum, have a predisposition for skin cancer. When a cell's DNA is damaged from ultraviolet radiation, the tumor suppressor gene, p53, normally induces the expression of UV-DDB. UV-DDB is then involved with the nucleotide excision repair of the damaged DNA (Hwang et al., 1999; Rapic-Otrin et al., 2002). If this process is not working, the result could be the development of cancer.

Table 7.1 Three proteins within the Rosenberg Lab that are undergoing crystallization trials.

Protein	Number of subunits	AA Length	MW (kDa)	pI_{est}
α-Synuclein	1	140	14.5	4.4
NAPOR-1	1	490	52.3	8.6
UV-DDB	2	1567	174.8	6.0

The sequences of these proteins are shown in Figure 7.1, while some sequence derived parameters are displayed in Table 7.1. These proteins are quite different in terms of molecular weight (MW), ranging in size from 14.5-174.8 kDa, and a pI_{est}, ranging from 4.4-8.6. UV-DDB is the largest of the three and consists of two subunits, one large subunit composed of 1140

amino acids and one small subunit composed of 427 amino acids. Both α-Synuclein and NAPOR-1 consist of one protein subunit.

(a) α-Synuclein, monomeric, SwissProt P37840

```
>sp|P37840|SYUA_HUMAN Alpha-synuclein
MDVFMKGLSKAKEGVVAAAEKTKQGVAEAAGKTKEGVLYVGSKTKEGVVHGVATVAEKTKEQVTNVGGAVVT
GVTAVAQKTVEGAGSIAAATGFVKKDQLGKNEEGAPQEGILEDMPVDPDNEAYEMPSEEGYQDYEPEA
```

(b) NAPOR-1, monomeric, SwissProt Q9UL67

```
>tr|Q9UL67|Q9UL67_HUMAN Neuroblastoma apoptosis-related RNA binding
protein - Homo sapiens (Human).
MNGALDHSDQPDPDAIKMFVGQIPRSWSEKELKELFEPYGAVYQINVLRDRSQNPPQSKGCCFVTFYTRKAA
LEAQNALHNIKTLPGMHHPIQMKPADSEKSNAVEDRKLFIGMVSKKCNENDIRVMFSPFGQIEECRILRGPD
GLSRGCAFVTFSTRAMAQNAIKAMHQSQTMEGCSSPIVVKFADTQKDKEQRRLQQQLAQQMQQLNTATWGNL
TGLGGLTPQYLALLQQATSSSNLGAFSGIQQMAGMNALQLQNLATLAAAAAAAQTSATSTNANPLSTTSSAL
GALTSPVAASTPNSTAGAAMNSLTSLGTLQGLAGATVGLNNINALAVAQMLSGMAALNGGLGATGLTNGTAG
TMDALTQAYSGIQQYAAAALPTLYSQSLLQQQSAAGSQKEGPEGANLFIYHLPQEFGDQDILQMFMPFGNVI
SAKVFIDKQTNLSKCFGFVSYDNPVSAQAAIQAMNGFQIGMKRLKVQLKRSKNDSKPY
```

(c) UV-DDB, p127 (**large subunit**), SwissProt Q16531 (1140 AA)

```
>sp|Q16531|DDB1_HUMAN DNA damage binding protein 1 (DDB p127 subunit)
(DDBa) (UV-damaged DNA-binding protein 1) (UV-DDB 1)
MSYNYVVTAQKPTAVNGCVTGHFTSAEDLNLLIAKNTRLEIYVVTAEGLRPVKEVGMYGKIAVMELFRPKGE
SKDLLFILTAKYNACILEYKQSGESIDIITRAHGNVQDRIGRPSETGIIGIIDPECRMIGLRLYDGLFKVIP
LDRDNKELKAFNIRLEELHVIDVKFLYGCQAPTICFVYQDPQGRHVKTYEVSLREKEFNKGPWKQENVEAEA
SMVIAVPEPFGGAIIIGQESITYHNGDKYLAIAPPIIKQSTIVCHNRVDPNGSRYLLGDMEGRLFMLLLEKE
EQMDGTVTLKDLRVELLGETSIAECLTYLDNGVVFVGSRLGDSQLVKLNVDSNEQGSYVVAMETFTNLGPIV
DMCVVDLERQGQGQLVTCSGAFKEGSLRIIRNGIGIHEHASIDLPGIKGLWPLRSDPNRETDDTLVLSFVGQ
TRVLMLNGEEVEETELMGFVDDQQTFFCGNVAHQQLIQITSASVRLVSQEPKALVSEWKEPQAKNISVASCN
SSQVVVAVGRALYYLQIHPQELRQISHTEMEHEVACLDITPLGDSNGLSPLCAIGLWTDISARILKLPSFEL
LHKEMLGGEIIPRSILMTTFESSHYLLCALGDGALFYFGLNIETGLLSDRKKVTLGTQPTVLRTFRSLSTTN
VFACSDRPTVIYSSNHKLVFSNVNLKEVNYMCPLNSDGYPDSLALANNSTLTIGTIDEIQKLHIRTVPLYES
PRKICYQEVSQCFGVLSSRIEVQDTSGGTTALRPSASTQALSSSVSSSKLFSSSTAPHETSFGEEVEVHNLL
IIDQHTFEVLHAHQFLQNEYALSLVSCKLGKDPNTYFIVGTAMVYPEEAEPKQGRIVVFQYSDGKLQTVAEK
EVKGAVYSMVEFNGKLLASINSTVRLYEWTTEKELRTECNHYNNIMALYLKTKGDFILVGDLMRSVLLLAYK
PMEGNFEEIARDFNPNWMSAVEILDDDNFLGAENAFNLFVCQKDSAATTDEERQHLQEVGLFHLGEFVNVFC
HGSLVMQNLGETSTPTQGSVLFGTVNGMIGLVTSLSESWYNLLLDMQNRLNKVIKSVGKIEHSFWRSFHTER
KTEPATGFIDGDLIESFLDISRPKMQEVVANLQYDDGSGMKREATADDLIKVVEELTRIH
```

UV-DDB, p48 (**small subunit**), SwissProt Q02466 (427 AA)

```
>sp|Q92466|DDB2_HUMAN DNA damage binding protein 2 (Damage-specific DNA
binding protein 2) (DDB p48 subunit) (DDBb) (UV-damaged DNA-binding
protein 2) (UV-DDB 2)
MAPKKRPETQKTSEIVLRPRNKRSRSPLELEPEAKKLCAKGSGPSRRCDSDCLWVGLAGPQILPPCRSIVRT
LHQHKLGRASWPSVQQGLQQSFLHTLDSYRILQKAAPFDRRATSLAWHPTHPSTVAVGSKGGDIMLWNFGIK
DKPTFIKGIGAGGSITGLKFNPLNTNQFYASSMEGTTRLQDFKGNILRVFASSDTINIWFCSLDVSASSRMV
VTGDNVGNVILLNMDGKELWNLRMHKKKVTHVALNPCCDWFLATASVDQTVKIWDLRQVRGKASFLYSLPHR
HPVNAACFSPDGARLLTTDQKSEIRVYSASQWDCPLGLIPHPHRHFQHLTPIKAAWHPRYNLIVVGRYPDPN
FKSCTPYELRTIDVFDGNSGKMMCQLYDPESSGISSLNEFNPMGDTLASAMGYHILIWSQEEARTRK
```

Figure 7.1 Sequences in FASTA format for (a) α-Synuclein, (b) NAPOR-1, and (c) UV-DDB

7.2 IDENTIFICATION OF SIMILAR PROTEINS

All three methods (CI_{50}, Charge Range Test, and Probability Distributions) are presented to predict solution pH ranges that are more likely to result in the crystallization of these test proteins. All of these methods are currently based on some aspect of the protein's $\overline{Q}$ curve. In this chapter, each prediction method is examined using: (1) Baseline (Chapter 5) and (2) binning by the pI_{est} (Chapter 6, Section 6.1).

For the Baseline method, it was assumed that there are no subgroups of proteins and all proteins are modeled as one group. Therefore, the Baseline $\overline{Q}_{cryst}$ frequency distribution (Figure 5.5c) was used to predict solution pH ranges. This is accomplished by matching each $\overline{Q}$ value along the protein's $\overline{Q}$ curve to the corresponding $\overline{Q}_{cryst}$ frequency distribution of successfully crystallized proteins, $P(\overline{Q} = \overline{Q}_{cryst} \mid nrPDB_{10.04.05})$.

The second method assumes that there are groups of proteins that display similar crystallization behavior, as demonstrated in Chapter 6. The simplest method of grouping proteins uses the protein's pI_{est} to assign the protein to one of five groups (pI_{est} Bins), 'Very Acidic,' 'Acidic,' 'Neutral,' 'Basic,' and 'Very Basic.' The pI_{est} is just one point along the estimated titration curve (or $\overline{Q}$ curve) where the estimated net charge is equal to zero. The $\overline{Q}$ probabilities, $P(\overline{Q} = \overline{Q}_{cryst} \mid pI_{est}, nrPDB_{10.04.05})$, are then generated from the frequency of the $\overline{Q}_{cryst}$ values for the successfully crystallized protein structures based on all proteins (Baseline) or groups of proteins with similar pI_{est} or $\overline{Q}$ curve values.

7.3 PREDICTING THE PH RANGES

The hypothesis was that proteins with similar $\overline{Q}$ curves should crystallize under similar pH ranges. Although different proteins may have the same $P(\overline{Q} = \overline{Q}_{cryst} \mid nrPDB_{10.04.05})$ priors based on belonging to the same group of proteins, the pH probabilities may be quite different. This is

due to the translation of $\overline{Q}$ back into pH space. The three methods discussed in Chapter 4 were used to estimate the pH_{cryst} range, 50% confidence intervals (CI_{50}), Charge Range Test, and pH_{cryst} probability distribution. Generally, all methods are complementary within a given method of grouping, but may differ between grouping methods, such as binning by pI_{est} and 2Step clustering.

Based upon the amino acid sequences in Figure 7.1, the $\overline{Q}$ curves were first calculated for all three proteins (Figure 7.2). Then the $\overline{Q}_{cryst}$ priors were then chosen based on either all proteins as one group (Baseline) or binning the proteins by their pI_{est} (supervised) and were then compared in their prediction of $P(\overline{Q} = \overline{Q}_{cryst} \mid nrPDB_{10.04.05})$. These $P(\overline{Q} = \overline{Q}_{cryst} \mid nrPDB_{10.04.05})$ values were then translated into the pH space, where the information could be directly used in the laboratory, $P(pH = pH_{cryst} \mid \overline{Q} = \overline{Q}_{cryst}, nrPDB_{10.04.05})$.

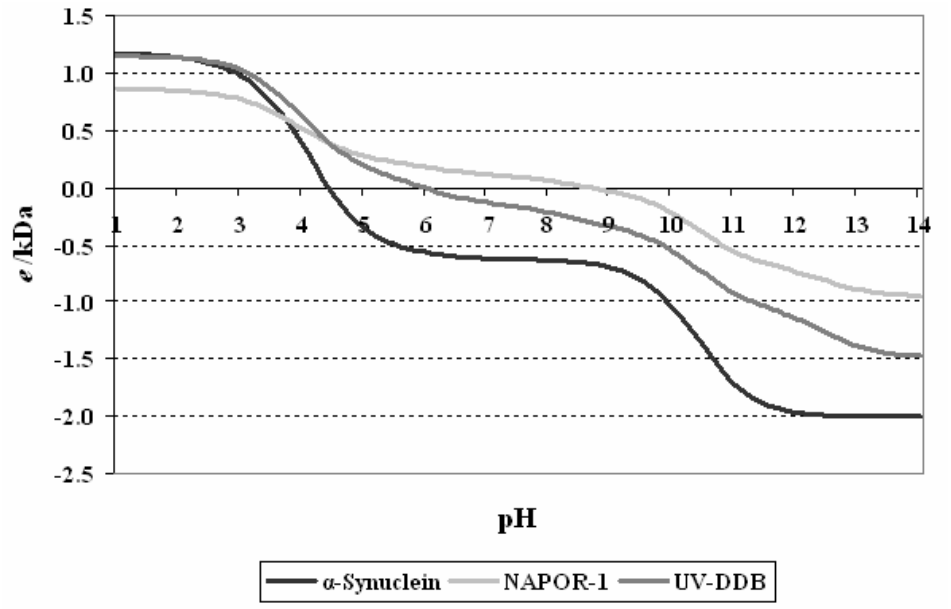


Figure 7.2 The $\overline{Q}$ curves for α-Synuclein, NAPOR-1, and UV-DDB.

7.3.1 α-Synuclein

The first clustering method discussed was binning proteins based upon their pI_{est}. α-Synuclein was labeled as 'Very Acidic' ($pI_{est} \leq 5$) with a pI_{est} of 4.4. The $P(\overline{Q} = \overline{Q}_{cryst} \mid nrPDB_{10.04.05})$ distribution was chosen based on the proteins in the training set, $P(\overline{Q} = \overline{Q}_{cryst} \mid nrPDB_{10.04.05}, pI_{est} = 'Very\_Acidic')$. The 976 'Very Acidic' proteins in this group generally crystallized with a negative charge with a mean $\overline{Q}_{cryst}$ value of -0.4 *e*/kDa. For the actual $\overline{Q}_{cryst}$ distribution, see Figure 6.3b in Chapter 6.

The first method examined, creates a confidence interval (CI) for the target protein based upon the middle 50% of the training set proteins within a group (CI_{50}). This CI_{50} range was then applied to the estimated titration curve of α-Synuclein to find the pH ranges that would result in a 50% chance of the pH_{cryst} being within that range if a crystal was produced.

The $\overline{Q}_{cryst}$ CI_{50} for Baseline was -0.3 to +0.1 *e*/kDa, while the CI_{50} for 'Very Acidic' proteins ranged from -0.6 to -0.2 *e*/kDa (Table 6.3). Figure 7.3 shows the CI_{50} for α-Synuclein as determined by using the pI_{est} = 'Very Acidic' proteins. When these $\overline{Q}$ values are translated back into the pH search space, a pH range of 4.3-4.9 (Baseline) or 4.7-8.4 (pI_{est} Bins) is suggested for the initial screens. The very narrow pH range (4.3-4.9) for the Baseline group is most likely due to the fact that α-Synuclein is classified as a 'Very Acidic' protein. The Baseline method was shown not to work as well with proteins with extreme pI_{est} values, $pI_{est} \leq 5.0$ or $pI_{est} \geq 9.0$, which are not well represented by the 'average' protein in the Baseline group. However, when the 'Very Acidic' proteins were used for priors, a much broader pH range is suggested for initial screens, 4.7-8.4. Based on the solution pH distribution of commercial screens (Appendix B; Table B2), 30% of the conditions can be removed from screening, because they have a solution pH that falls outside of this range (5.0-8.0). This would allow the researcher more flexibility in testing other conditions.

The second method uses the results from the Charge Range Test of each grouping method to estimate a probability of crystallization for the pH range that results in a $\overline{Q}$ value within the group's Mean $\overline{Q}_{cryst}$ ±0.1 to $\overline{Q}_{cryst}$ ±0.2. The mean $\overline{Q}_{cryst}$ value for the 'Very Acidic' proteins was

-0.4 *e*/kDa. The pH values that result in a $\overline{Q}$ of -0.4±0.1 *e*/kDa will have the same probability of crystallizing as that of the 'Very Acidic' proteins from Section 6.1.3. When these $\overline{Q}_{cryst}$ values are translated back into the pH search space, a pH range of 4.8-5.8 (Mean±0.1) or 4.7-8.4 (Mean±0.2) is suggested for α-Synuclein (Table 7.2). The probability that the pH_{cryst} will fall within these ranges is 47% and 70%, respectively.

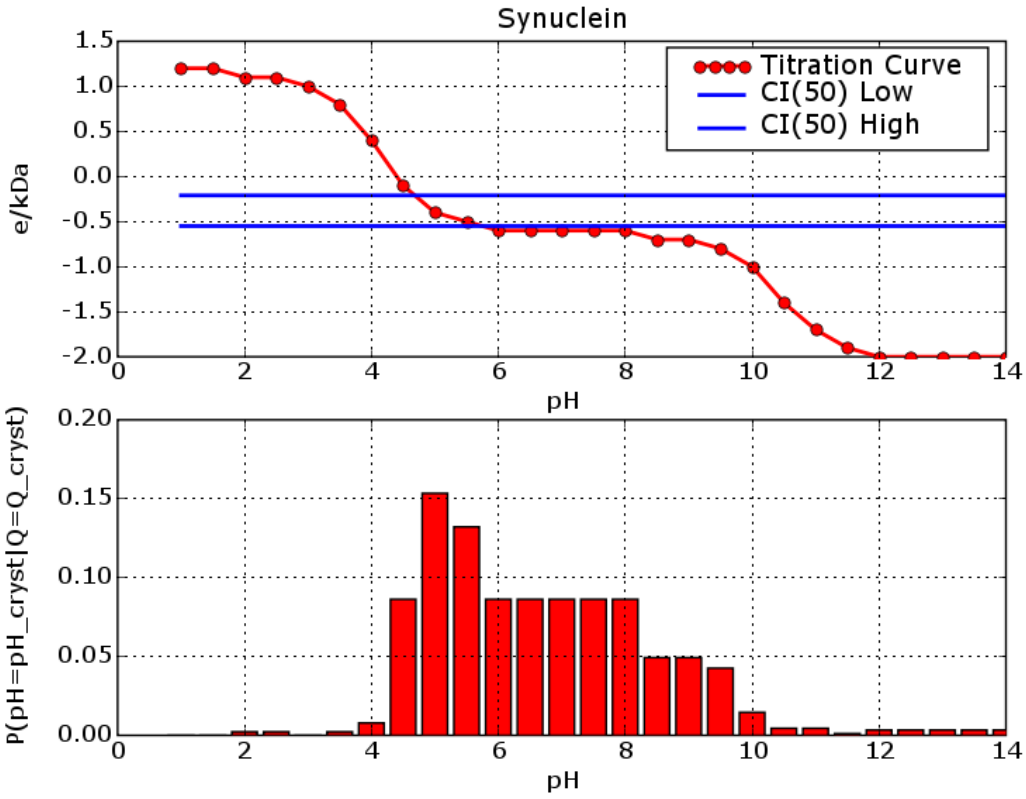


Figure 7.3 The $\overline{Q}$ curve with CI_{50} interval and pH_{cryst} probability distribution for α-Synuclein.

The next approach aimed at predicting a $\overline{Q}$ range for crystallization calculates a probability of each $\overline{Q}$ value from -2.0 to +2.0 *e*/kDa. These probabilities can then be carried over to the pH space by matching values along the estimated titration curve, $P(pH = pH_{cryst} \mid \overline{Q} = \overline{Q}_{cryst}, nrPDB_{10.04.05})$.

The pH probability distribution calculated for α-Synuclein is shown in Figure 7.3. When examining the probability curve, a spike in probability is observed around a pH of 5.0-5.5

for α-Synuclein. This spike is missed in the both the CI_{50} and Charge Range Test results, which do not show local maxima values. For example, all values between the CI_{50} lines are equally probable. It can be inferred from the Charge Range Test results that the mean $\overline{Q}_{cryst}$ value has the highest probability and as the $\overline{Q}$ moves away from the mean value, a decrease in probability can be expected. However, the Charge Range Test does not indicate how much of a loss in probability can be expected as the $\overline{Q}$ moves away from the mean. For the pH probability distribution a greater than 10% threshold was used as a high probability. The 10% probability threshold cutoff suggests an acidic pH range for both the Baseline group (4.5-5.0) and the pI_{est} bins (5.0-5.5).

Table 7.2 Suggested pH ranges for initial crystallization attempts of the three test proteins based on either Baseline or pI_{est} Binning.

	α-Synuclein			NAPOR-1			UV-DDB		
Method	pH_{Low}	pH_{High}	Prob.[2]	pH_{Low}	pH_{High}	Prob.[2]	pH_{Low}	pH_{High}	Prob.[2]
CI_{50} Baseline	4.3	4.9	50.0	6.4	10.3	50.0	5.2	9.0	50.0
CI_{50} pI_{est} Bins	4.7	8.4	50.0	4.6	8.0	50.0	6.3	9.0	50.0
Charge Range Test, pI_{est} Mean±0.1	4.8	5.8	42.6	4.6	8.0	56.7	5.7	8.2	58.2
Charge Range Test, pI_{est} Mean±0.2	4.7	8.4	63.5	4.3	9.1	82.1	5.2	9.0	79.4
10% Prob.[1] - Baseline	4.5	5.0	44.0	9.5	9.5	11.3	6.5	8.0	46.6
10% Prob.[1] - pI_{est}	5.0	5.5	44.0	6.5	8.0	50.9	6.5	9.0	83.3

[1] 10% Prob. = probability distribution with a 10% threshold for cut-off

[2] Probability based on the training set proteins within the group

All three methods chosen to predict a solution pH range, generally agree that an acidic pH range (5.0-6.0) would be more likely to crystallize α-Synuclein. It is also expected that the pI_{est} Bins will give a more accurate prediction of pH_{cryst} than the Baseline group. This is

primarily due to the very acidic pI_{est} of α-Synuclein, 4.6, which would not represent an 'average' protein in the Baseline group. Finally, the calculation of the pH_{cryst} probability distribution can detect local maxima, which the other two methods cannot.

7.3.2 NAPOR-1

Based upon its pI_{est} (8.6), NAPOR-1 was assigned to the 'Basic' group ($8 \leq pI_{est} < 9$). The mean (±SD) pI_{est} and $\overline{Q}_{cryst}$ of the 564 'Basic' proteins was 8.2±0.4 and +0.2±0.3 respectively.

Figure 7.4 shows the $\overline{Q}$ curve for NAPOR-1 and the CI_{50} range for the pI_{est}='Basic' proteins, +0.06 to +0.27. For this protein, different $\overline{Q}_{cryst}$ CI_{50} ranges are predicted by each method, Baseline (-0.25 to +0.14 *e*/kDa) and pI_{est} Bins (+0.06 to +0.27 *e*/kDa). Therefore, these $\overline{Q}$ ranges will predict quite different pH ranges for NAPOR-1, as shown in Table 7.2. When these $\overline{Q}$ values are translated into the pH search space, the $\overline{Q}_{cryst}$ CI_{50} range for Baseline suggests a Neutral-Basic pH range for crystallization, 6.4-10.3. The 'Basic' pI_{est} Bin suggests a $\overline{Q}$ range from pH 4.6-8.0. Similar to α-Synuclein, it is expected that the 'Basic' proteins would represent more similar proteins to NAPOR-1 than using all proteins.

Using the Mean±0.1 $\overline{Q}$ range for pI_{est}='Basic' proteins, a pH range of 4.6-8.0 is also suggested (Table 7.2). The estimated probability of NAPOR-1 crystallizing in this pH range is 53% based on the results in Section 6.1.3. This will reduce the number of screen conditions searched by 23.4% (Table B.2). When the $\overline{Q}$ range is increased to Mean±0.2, a larger pH range, 4.3-9.1, is suggested, along with an increased probability (78%) of crystallization. However, it should be noted that only 2.5% of the crystallization screens have a solution pH reported to be outside of this range. Therefore, it appears that the Mean±0.2 $\overline{Q}$ provides little benefit for NAPOR-1.

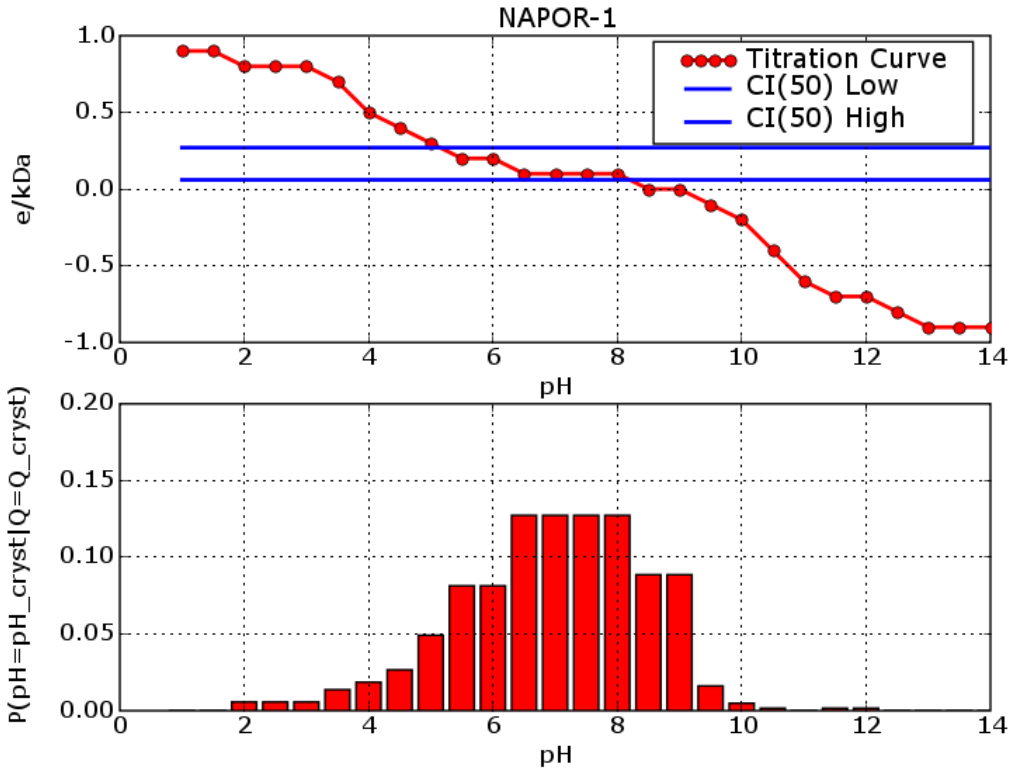


Figure 7.4 The $\overline{Q}$ curve with CI_{50} interval and pH_{cryst} probability distribution for NAPOR-1.

Using a 10% probability threshold, the Baseline group predicts only one pH value above the 10% threshold, pH 9.5 (Table 7.2). As discussed previously in the CI_{50} section, this appears unlikely given the 'Basic' nature of NAPOR-1. The pI_{est}='Basic' pH distribution predicts a relatively narrow solution pH range, 6.5-8.0 (Figure 7.4), which will reduce the crystallization conditions searched by 55.6%. A 51% probability for $P(pH = pH_{cryst} \mid \overline{Q} = \overline{Q}_{cryst}, nrPDB_{10.04.05})$ is predicted for this range. This will allow more extensive searches in more likely crystallization regions. Again some local maxima $P(pH = pH_{cryst} \mid \overline{Q} = \overline{Q}_{cryst}, nrPDB_{10.04.05})$ values (pH 6.5-8.0) are not visually obvious with the $\overline{Q}_{cryst}$ CI_{50} and Charge Range Test methods, but could be displayed with the probability distribution.

These results suggest different pH ranges depending upon the method used to group the structures. Binning by the pI_{est} predicts a wider pH range from either 4.6-8.0 (CI_{50} and Charge Range Test) to 6.5-8.0 (Probability distribution). A local maximum was again observed using

the probability distributions, $P(pH = pH_{cryst} \mid \overline{Q} = \overline{Q}_{cryst}, nrPDB_{10.04.05})$. Based upon these results, initial crystallization attempts for NAPOR-1 should focus on a narrow pH range of 6.5-8.0. The solution pH could be expanded to 4.6-8.0 if enough protein is available.

7.3.3 UV-DDB

UV-DDB was labeled 'Acidic' with a pI_{est} of 5.95, which straddles the pI_{est} binning border between the 'Acidic' ($5.0 < pI_{est} \leq 6.0$) and 'Neutral' ($6.0 < pI_{est} < 8.0$) protein groups. For this protein, a more specific pH range of 5.0-7.0 may be more appropriate as 'similar' proteins.

The CI_{50} range (Figure 7.5) for both methods was very similar, -0.25 to +0.14 *e*/kDa for Baseline and -0.25 to -0.06 *e*/kDa for the pI_{est} = 'Acidic' Bins. When these values are transferred into the pH search space, similar pH ranges will be predicted for UV-DDB with the Baseline method predicting a slightly large pH range (pH 5.2-9.0) than the pI_{est}='Acidic' Bin (pH 6.3-9.0; Table 7.2).

Using the Mean±0.1 $\overline{Q}_{cryst}$ range for pI_{est}='Acidic' proteins, a pH range of 5.7-8.2 is suggested (Table 7.2). The estimated probability of UV-DDB crystallizing in this pH range is 49% based on the Test Set results in Section 6.1.3. This will reduce the number of screen conditions searched by 39% (Appendix B, Table B.2). When the $\overline{Q}$ range is increased to mean $\overline{Q}_{cryst}$±0.2, a larger pH range, 5.2-9.0, is suggested, along with an increased probability of crystallization, 78%. However, only 10% of the crystallization screens have a solution pH reported to be outside of this range. This reduction will slightly reduce the initial conditions to search.

When the pH_{cryst} probability distribution was examined (Figure 7.5), two peaks in probability were a little more pronounced, pH 7.5 and 8.0. Each of these pH values had a $P(pH = pH_{cryst} \mid \overline{Q} = \overline{Q}_{cryst}, nrPDB_{10.04.05})$ over 15%. The probability values for 6.5-7.0 are slightly lower than 15%, but still well above the 10% threshold. The Baseline group suggests a slightly more narrow pH range, 6.5-8.0, than the pI_{est}='Acidic' Bin, pH 6.5-9.0.

All methods predict that UV-DDB is more likely to crystallize when the pH ranges from 6.5-8.0 with slight variations, down as low as pH 5.2 and as high as pH 9.0. Generally, these pH

values are the ones most sampled by the commercial screens. However, this information does suggest that crystallization attempts should be at a solution pH greater than 5.5. This would remove a subset of acidic conditions (9.2%) and allow the researcher to focus their efforts a little more.

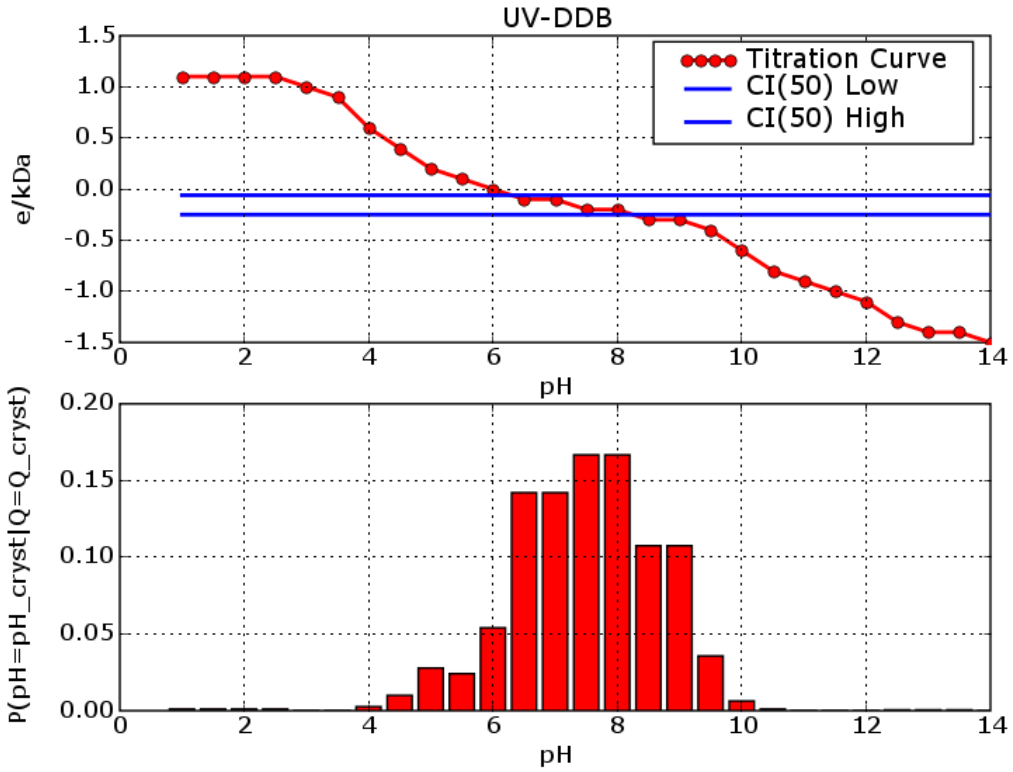


Figure 7.5 The $\overline{Q}$ curve with CI_{50} interval and pH_{cryst} probability distribution for UV-DDB.

7.4 CHAPTER SUMMARY

The three methods discussed in Chapter 5, namely 50% Confidence Intervals (CI_{50}), Charge Range Test, and $\overline{Q}_{cryst}$ probability distributions, were used to predict a solution pH range for crystallization of three proteins currently undergoing structural determination in the laboratory of Dr. John Rosenberg (Department of Biological Sciences, University of Pittsburgh), α-Synuclein, NAPOR-1, and UV-DDB. The goal is to predict the solution pH ranges that have a higher

probability of resulting in crystallization for the target protein given the protein's amino acid sequence.

This is accomplished by estimating the protein's $\overline{Q}$ curve from the amino acid sequence and then matching the $\overline{Q}$ values along the curve to the $\overline{Q}_{cryst}$ values of previously crystallized proteins, $P(\overline{Q} = \overline{Q}_{cryst} \mid nrPDB_{10.04.05})$. These probabilities can then be transferred into the pH search space by using the protein's $\overline{Q}$ curve, $P(pH = pH_{cryst} \mid \overline{Q} = \overline{Q}_{cryst}, nrPDB_{10.04.05})$. In addition to comparing the three different methods of predicting the pH_{cryst} ranges, two different methods of grouping proteins by similarity were compared. The Baseline grouping method did not consider subgroups of proteins, i.e. all proteins were considered equal. The second method used pI_{est} Binning (Section 6.1.1.2) to separate proteins into groups.

The Baseline approach assumes that there are no subgroups of proteins and that all proteins can be assigned probabilities based on a one-size-fits-all approach. Binning by the pI_{est} assumes that proteins with a similar pI_{est} can be used to improve the prediction over Baseline. As a result, when the target protein has a pI_{est} in the 'Acidic' to 'Basic' range, close to the overall mean pI_{est} (6.3±1.7), the predictions are generally not widely different from that obtained by the Baseline method. However, as the target protein's pI_{est} moves farther away from the overall pI_{est} mean, the probability differences can be quite different (Figure 7.6).

For example, the proteins α-Synuclein (pI_{est} = 4.4; Figure 7.6a) and NAPOR-1 (pI_{est} =8.6; Figure 7.6b) have pI_{est} values that are quite different than the training set's mean pI_{est} of 6.3. As a result the pH values with the highest probabilities using Baseline are much closer to the protein's pI_{est}., because the Baseline $\overline{Q}_{cryst}$ values with the highest probabilities are close to the pI_{est}, mean $\overline{Q}_{cryst}$ of -0.1 *e*/kDa. When the predictions are based on the matching the pI_{est} bin, there is large shift towards a more neutral pH where there are more negative (α-Synuclein) or positive (NAPOR-1) $\overline{Q}$ values. These $\overline{Q}$ values were shown to have a higher probability in crystallizing 'Acidic' and 'Basic' proteins, more closely representing the groups mean $\overline{Q}_{cryst}$ values (-0.4 *e*/kDa for 'Acidic' and +0.2 *e*/kDa for 'Basic' proteins; Figure 6.3; Table 6.1b).

Alternatively, when the target protein has a pI_{est} value that is closer to the mean pI_{est} value of the Baseline group (pI_{est} = 6.3), there isn't as much as a shift in probability values when comparing Baseline to the pI_{est} Binning. This was demonstrated with the target protein UV-DDB

(pI_{est} = 5.95; Figure 7.6c). There is a shift in pH probability values for UV-DDB, but there may be an additional problem binning UV-DDB by the pI_{est}. UV-DDB has a pI_{est} of 5.95, which is right at the pI_{est} separation point for 'Acidic' and 'Neutral' proteins (pI_{est} = 6.0). A better estimate of the pH_{cryst} probability may be obtained if all proteins within ±1.0 pI_{est} units are used to estimate the probability ($5.0 \leq pI_{est} < 7.0$). This remains an area of future research in case selection of 'similar' proteins, i.e. case-based reasoning.

As demonstrated with all three target proteins, each method of predicting the pH range for crystallization attempts are relatively complimentary, which they should be, because they are all based upon the same set of proteins. However, there are some differences between the methods. For example, the CI_{50} will suggest a pH_{cryst} range, but it gives equal probability among all pH values within the range. The Charge Range Test method can give a more accurate prediction by using the various levels of $\overline{Q}$, ±0.1, ±0.2 for the selection of appropriate ranges. The probabilities obtained using this method also may be less than or greater than 50%. This method does have its problems, because there may be differences in the pH_{cryst} probabilities below or above the mean $\overline{Q}_{cryst}$ value, such as between Mean-0.1 *e*/kDa and Mean+0.1 *e*/kDa. The Charge Range Test method may therefore miss some areas of local maximum probability. Therefore, it is felt that the probability distribution method will allow the researcher to make the most intelligent choice based upon any other limitations, such as a limited amount of protein.

The $P(pH = pH_{cryst} \mid \overline{Q} = \overline{Q}_{cryst}, nrPDB_{10.04.05})$ calculations may result in pH_{cryst} distributions from which any local maxima values can be determined. Of the three proteins examined, α-Synuclein was predicted to have a very narrow pH range of high probability, 5.0-5.5 (44%). A wider pH range of 4.5-8.0 also has a higher probability than randomly searching the pH space, but less than the 10% threshold probability. A more neutral pH is suggested for the basic NAPOR-1 (pI_{est} = 8.6), pH 6.5-8.0. This range is estimated to have a 51% chance of capturing the pH_{cryst}. A slightly larger pH range, 6.5-9.0, is suggested for the slightly acidic UV-DDB (pI_{est} = 5.95). An increase in the pH_{cryst} probability for this range is also observed, 83.3%. However, the lack of crystallization of these proteins might be due to something other than the solution pH, because previous attempts with commercial screens have failed. Other methods for translating back to the pH space might yield different probability distributions,

which could be explored in future work. Finally, the pH_{cryst} probabilities can be calculated over a narrow (pH 4.0-10.0) or broad (pH 1.0-14.0) range of pH values with narrow (0.1) or broad (0.5) pH intervals. For some proteins, the results will significantly reduce the pH values that should be searched, as with α-Synuclein, while other predictions will have little effect in reducing the pH search space.

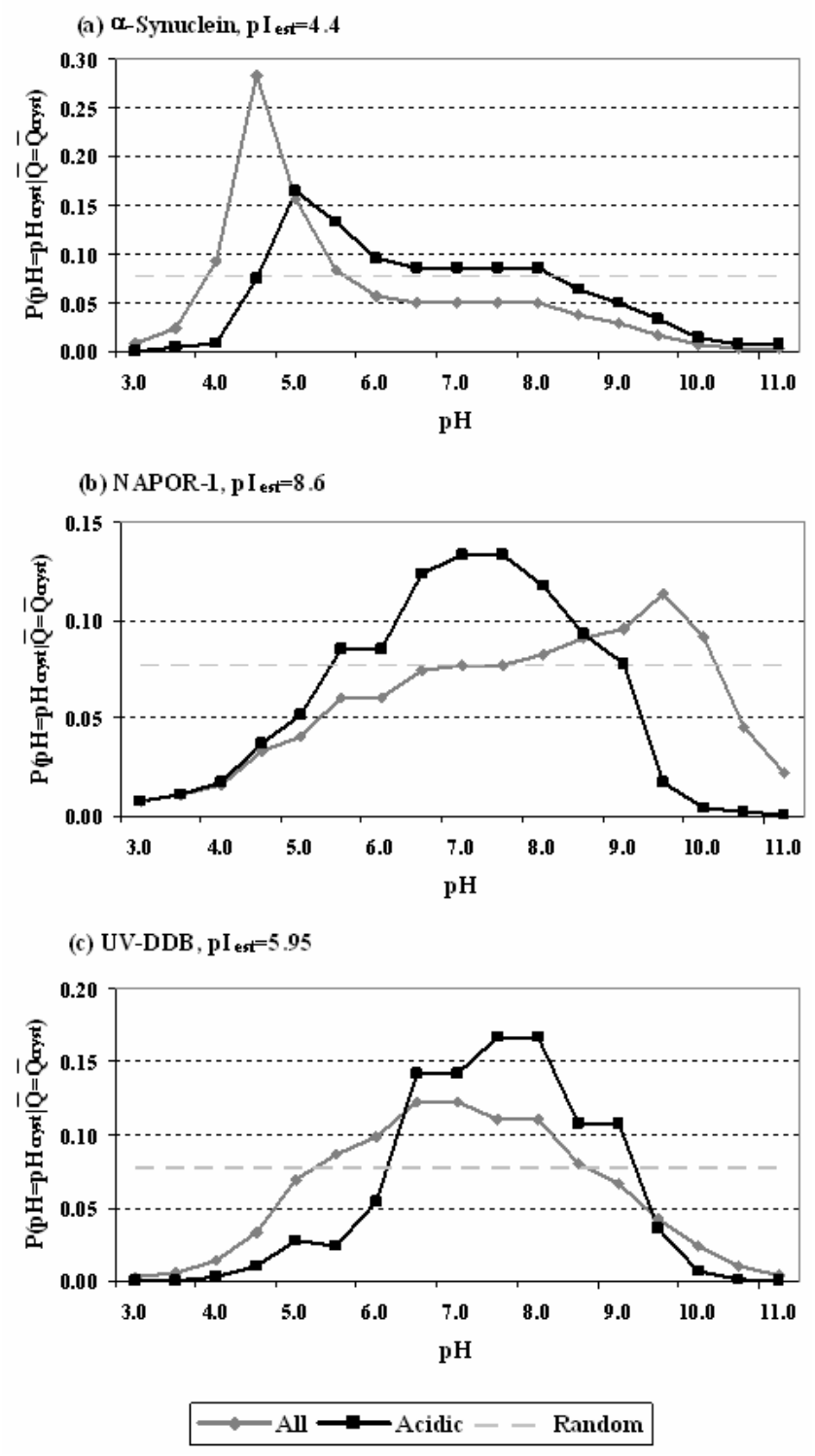


Figure 7.6 Comparing Baseline results with pI_{est} Binning, to estimate $P(pH = pH_{cryst} \mid \overline{Q} = \overline{Q}_{cryst}, nrPDB_{10.04.05})$ for the three example proteins.

8.0 DISCUSSION

Protein crystallization remains a major bottleneck for structural biologist, because of the idiosyncratic behavior of the proteins in solution. Currently, there are few, if any, rules for the selection of solution pH conditions for crystallization and the success rates are quite low. The initial results in Chapter 3 indicated that separating proteins into groups based upon their estimated isoelectric point (pI_{est}) or estimated titration curve may allow the researcher to intelligently select solution pH ranges that have a higher probability in generating crystals, while at the same time removing those pH ranges that have little chance of success. These results led to the development of a framework for identifying protein features and solution conditions that are related to successful crystallization. The Protein Sequence-Properties Evaluation (PSPE) framework was created (Chapter 4) to aid in the statistical evaluation of variables for predicting ranges of and prior probability distributions for protein crystallization conditions. Development of such a framework was largely motivated by the rapid growth and evolution of the Protein Data Bank. Because the pH_{cryst} is the most widely reported experimental variable for crystallization and the solution pH can largely control the protein's charge, various charge variables, including the specific charge ($\overline{Q}_{cryst}$) and average surface charge density (σ_{cryst}; Chapter 5) were estimated and their distributions examined in previously crystallized proteins. The principal observation of this study was that proteins appear to crystallize at low values of $\overline{Q}_{cryst}$ and σ_{cryst}. This framework, while tested only on the solution pH and related charge variables (Q_{cryst}, $\overline{Q}_{cryst}$, and σ_{cryst}), could be used in theory to identify any other crystallization variables (Section 8.1). Using a retrospective approach, the models developed in Chapters 5 and 6 were then tested and validated on an independent test set of more recent entries to the Protein Data Bank (PDB; Section 8.2). The possible success of this approach will hopefully lead to a more thorough

reporting of other crystallization solution parameters (Section 8.3), including a more accurate reporting of the pH_{cryst} (Section 8.4).

Many assumptions were made for the calculation of a protein's estimated net charge from the pH_{cryst}. With some modifications to these algorithms, a more accurate estimation of the $\overline{Q}_{cryst}$ and σ_{cryst} can be performed (Section 8.5). While the current results can be largely explained by physical chemistry, a more accurate calculation should increase the predictive power of the models and increase our understanding of the crystallization process (Section 8.6). With a more thorough understanding of the process, crystallization screens can be designed, which should improve upon the current low success rates (Section 8.7).

8.1 PSPE FRAMEWORK

The PSPE framework for identifying protein crystallization solution conditions based upon a protein's biophysical properties was developed in Chapter 4. The PSPE framework is an instantiation of the "scientific method" for the framing and testing of hypotheses in an informatics setting. This 'easy' framework was developed for testing hypotheses for explaining differences in protein behavior in crystallization screens. Initially, the framework was applied to proteins as one large group and indicated that the $\overline{Q}$ and σ may be possible proxy variables for the pH_{cryst} (Chapter 5). However, when proteins were separated into groups based upon similar biophysical properties (*Features*) that can be calculated prior to any crystallization attempts, an improvement was observed in the prediction of the estimated specific charge ranges for growing protein crystals (Chapter 6).

Previous research has failed to correlate the pI_{est} to the pH_{cryst}. The PSPE framework was used to identify *Features* and *Hidden Controllables* (Q, $\overline{Q}$, and σ), which could potentially be used to select solution pH ranges that have a higher probability of generating crystals. This can be done by using a protein's amino acid sequence to first calculate the protein's estimated titration curve. From the estimated titration curve, the $\overline{Q}$ and σ curves can be obtained by dividing the estimated net charge of the protein by either the MW_{au} or A_S,

respectively. Based upon the Q_{cryst} values of previously crystallized proteins, the $\overline{Q}$ and σ values of the target protein might be used to guide the researcher in selecting pH values that have a higher probability in forming crystals. Combined with other information, this knowledge could help lead to the development of protein-specific screens. Using this knowledge should increase the probability of generating crystals of suitable quality for diffraction studies by removing conditions not likely to result in crystals, while increasing the sampling in areas more likely to result in crystals.

It is also important to discuss the limitation of the PSPE Framework. The main limitation is that the PSPE Framework was tested on one example, identification of the $\overline{Q}$ and σ as possible proxy variables for the pH_{cryst}. This can be considered one example, because these variables are close approximations of each other, Spearman's rho of 0.995. However the major limitation is that this observation needs experimental validation. Another limitation is that the PSPE Framework is currently not automated. This can be accomplished; however, there is still the need for a highly involved user. The framework cannot automatically find hidden variables, hypotheses have to be developed and tested. Finally, another limitation is the availability of training and test set data. A researcher will likely need to generate new data sets from the PDB with each hypothesis. The dynamic nature of the database (continually evolving) also presents an interesting challenge as new protein structures are added with some others being replaced.

8.2 TEST SET VALIDATION

In order to validate any models or predictions, the solution pH ranges for crystallization were predicted on newer entries to the PDB. Alternatively, cross-validation can be used, which would also give an estimate of the prediction error. Due to a high level of bias present in the PDB, a non-redundant set of proteins should be used for both the training and test sets. Because the PDB is an evolving database, the models can be updated as more 3D structures are solved. This process can also be automated. However, the ideal validation would be the actual crystallization of target proteins. This experimental validation would demonstrate the value of recording more experimental parameters.

8.3 RECORD MORE INFORMATION

Similar to Peat et al. (2005), the present results suggest that researchers should include more standardized information when depositing structures in the PDB. This information would then be accessible in the mmCIF files for future analysis on the requirements for growing crystals. For example, if a commercial screen is used to grow the crystals, the screen and the well solution details should be listed. Biliverdin Reductase A (PDB ID: 1LC3) was a good example where Emerald Screen Wizard-I, condition 27 was used to grow the crystal. This information was found in the _exptl_crystal_grow.pdbx_details field of the mmCIF file. There are occurrences in the PDB where this happens, but very infrequently. With more information available, data mining may lead to an increase in the success rates (Luft et al., 2003; Walter et al., 2005). Additionally, researchers will be more likely to understand the conditions, including solution pH, which are required for the formation and growth of protein crystals suitable for diffraction studies.

8.4 REPORTED PH OF CRYSTALLIZATION

The saw-tooth pattern in the pH_{cryst} distribution strongly suggests that very few researchers are measuring the actual solution pH. Although commercial companies may report the pH of the buffer in their screens, it was demonstrated that the actual pH of the solution may be quite different (0.5-5.0 pH units) from the buffer listed or between the supplier's matched conditions (Bukrinsky et al., 2001; Wooh et al., 2003). Recently, Hampton Research published a list of the measured pH values of their well solutions. Although, many values were close to the pH of the buffer, some were quite different (Appendix B).

Additionally, these pH values are also often only appropriate for the well solution and not the protein solution. When making the protein solution, typically the well solution is mixed 50/50 with the protein solution, which may contain a different pH with different buffers than the well solution. This interaction between the reservoir and protein solutions may shift the solution pH at setup. Additionally, the solution pH can also drift during the crystallization process. This

would especially be systematic for reservoir solutions that are very acidic or very basic. Here the protein solutions are more likely to be buffered towards the neutral and hence shift the pH in that direction. These effects would broaden the $\overline{Q}$ and σ distributions, i.e. adding to the variance. Therefore, the pH of the well and protein solution should be listed separately. Based on current understanding, the pH_{cryst} present in the PDB is most often the pH of the well solution. A more accurate reporting of the protein solution's pH will allow a better estimate of the net charge of the protein.

8.5 LIMITATIONS WITH THE ESTIMATION OF NET CHARGE

Experimental titration curves are rarely measured due to the time and effort required to obtain the required information. Detailed pK_a measurements are generally not attempted until functional and structural information has accumulated, suggesting an interesting phenomenon. Computational methods of determining the pK_a values all require detailed structural knowledge. However, Q can be estimated using the HHE (Equation 2.1). The pK_a values used in most calculations are model values that have been determined from the isolated amino acid, which are generally quite accurate (Yang et al., 1993). It should be noted that there is some variation, depending upon the source. Differences in model pK_a values can be as large as 0.9 units for the same group. However, these average values are relatively accurate for surface exposed residues. Additionally, some sources model every amino acid's –NH3 and –COOH terminus as one value, such as Accelrys' GCG package, while others have individual values for each amino acid, such as that used in this dissertation (Nelson and Cox, 2000). Therefore, some differences may exist when computing the titration curves using different sources for model pK_a values.

Additionally, assumptions are made when calculating the estimated titration curves. The first assumption was that all like amino acids have the same pK_a value, which is not the case. Local environmental effects may cause significant shifts in pK_a values of individual residues. The amino acids with large observed pK_a shifts are often those that are important functionally or structurally. In a study of 24 crystallized proteins with known pK_a values determined by NMR spectroscopy, the pK_a values for the carboxyl groups of Aspartic acids averaged 3.9 in acidic (pI

< 5) and 3.1 in basic proteins (pI > 8). It was also observed that Aspartic acid residues with extreme pK_a values (>5.5) were often found in the active sites or ligand binding sites (Forsyth et al., 2002). For example, as the protein becomes more acidic, there is an increasing tendency for the negative charges to cluster, which is electrostatically unfavorable, thus raising their pK_a values. These effects are likely to be systematic.

The second assumption made when estimating the titration curve was that all charged residues are accessible. However, this is also known not to be the case. Currently, there is no way to determine *apriori* which residues are located on the surface, which is why the term 'estimated' titration curve is used. It might be possible to more closely estimate the number of charged amino acids on the surface, by adjusting for surface amino acid propensity prior to calculating estimated titration curve. For example, most Lysine, Glutamic acid, Aspartic acid, and Arginine residues are located on the surface of both monomeric (Miller et al., 1987a; Tsai et al., 1997) and oligomeric proteins (Janin et al., 1988), while over 50% of the Histidine, Cysteine, and Tyrosine residues are located in the interior. In addition to adjusting for surface propensity, if the protein is an oligomer, adjustments for the protein-protein interface amino acid propensity can also be made (Janin et al., 1988, Tsai et al., 1997) of oligomeric proteins prior to calculating estimated titration curve. These are possible areas for future study.

Charged amino acids are often found in serendipitous clusters, which bind ions, thus altering/shielding the charge. This reduces the net charge of the protein, which decreases the possible unfavorable electrostatic interactions between protein molecules allowing them to come into closer contact. Functional ion binding sites typically involve clusters of charged residues, e.g. cations, which are typically bound by several acidic groups. For example, the active site of inositol monophosphatase (PDB id: 2BJI; a homo-dimer) binds three Mg^{2+} ions. These three Mg^{2+} ions in each dimer are bound to the side chains of Glu70, Asp90, Asp93, and Asp220 (Gill et al., 2005). Negative ions can also be found bound to residues, however, they are usually involved in neutralizing positive charge. For example, lysozyme can bind four Cl- ions (Retailleau et al., 1997). Binding of ions, whether positively or negatively charged, may shift the estimated titration curve and pI (Green, 1931b; Retailleau et al., 1997). This complicates the calculation of the $\overline{Q}_{cryst}$ and σ_{cryst}, because there is currently no method to estimate the amount of ions that will bind apriori. Therefore, ion-binding effects are likely to be systematic as well.

However, it should be mentioned that the solution pH is not the only method of controlling the net charge. For example, nucleic acids are significantly more charged than are proteins, yet still manage to crystallize. Therefore, the protein's charge shouldn't be a barrier for crystallization. The nucleic acid crystallization literature describes many reagents for dealing with this strong negative charge, ranging from small divalent cations, such as Mg^{2+}, to small polyamines, such as spermine and putrascein, and including specialized reagents, such as cobalt hexamine. These reagents provide the necessary counter ions to neutralize the negative charge and help stabilize the structures for crystallization (Dock-Bregeon et al., 1999).

Due to difficulties estimating the A_S, the $\overline{Q}_{cryst}$ values appear to be a better variable to model than the σ_{cryst}. The $\overline{Q}$ values can be transformed into the σ values by taking the molecular weight to the appropriate power (Section 5.2). The A_S can also be adjusted for the predicted number of chains in the biological/asymmetric unit. Adjustments could also be made when considering the chains in the complex, whether they involve permanent, transient, or crystal contacts. Each type of contact buries a different amount of surface area on average with the permanent complexes burying the most and the crystal contacts burying the least (Jones and Thornton, 1996; Carugo and Argos, 1997; Dasgupta et al., 1997; Valdar and Thornton, 2001; Bahadur et al., 2003). Therefore, the primary amino acid sequence, particularly the composition of the charged amino acids, may play a critical role in suggesting pH ranges for obtaining crystals. The kinds of systematic effects discussed previously can be corrected for empirically by examining the pH_{cryst} values of proteins with similar amino acid composition, e.g. by pI binning. This would be especially worthwhile when designing the initial screens for a new protein.

8.6 PHYSICAL SIGNIFICANCE OF SPECIFIC CHARGE AND SURFACE CHARGE DENSITY

The distributions of both the $\overline{Q}_{cryst}$ and σ_{cryst} of previously crystallized proteins were examined. While the $\overline{Q}_{cryst}$ is a relatively clean number, which is well characterized thermodynamically

(specific charge), it does not provide the best explanative value for the problem. However, the σ_{cryst} has a more explanative value and allows for comparison to other charged biological models that crystallize, such as nucleic acids, but the calculations are based on a lot of interpretations and are rather soft. Errors in the estimation of the A_S may propagate when the σ_{cryst} is calculated due to uncertainties in the calculation of surface area. This would be more problematic for large proteins. These issues may lead to a decrease in the accuracy of the predicted pH_{cryst} range. Therefore, both the $\overline{Q}_{cryst}$ and σ_{cryst} were examined and how they relate to proteins with known 3D structures.

The distribution of the $\overline{Q}_{cryst}$ and σ_{cryst} around the pI_{est} (Figures 5.5) indicated that these proteins are close to neutral, electrostatically. It is likely that the crystallization process naturally selects for a low charge to limit negative electrostatic interactions between molecules, allowing them to come into close proximity to form crystal lattice contacts. An alternative hypothesis would be that the low charge is a consequence of natural selection for function and/or stability. The $\overline{Q}_{cryst}$ distributions were normally distributed and could be fit with a Gaussian (Section 6.3). A high correlation (0.995) between the two variables was also observed. Thus, a low $\overline{Q}$ is equivalent to a low σ. For comparison purposes, 0.1 *e*/kDa corresponds to one excess positively charged amino acid out of 91 amino acids, assuming an average molecular weight of 110 Daltons per amino acid. So the original idea of crystallizing around the isoelectric point was not far off. However, a low $\overline{Q}_{cryst}$ or σ_{cryst} does not necessarily translate into a pH_{cryst} being close to the pI_{est}. This is partially due to the shape of the protein's estimated titration curve.

However, there was a theoretical difficulty with the idea of crystallizing at the pI. The more charged residues a protein has on its surface, the higher the probability of unfavorable electrostatic interactions between molecules, which will not allow the molecules to come into close contact and form crystals (Ries-Kautt and Ducruix, 1999). Global electrostatic effects are problematic because they are necessarily long range. Electrostatic forces are quickly damped in high dielectric aqueous media, particularly so in the presence of a high salt concentration. It was therefore felt that electrostatics would be more important locally for crystallization i.e. between residues at (or near) lattice contacts especially between residues on different equivalent molecules (Tardieu et al., 2001).

The principal observation of this study is that proteins appear to crystallize at low values of $\overline{Q}$ and σ. One problem with this observation is that "low" is a relative term and the frame of reference requires careful examination. One frame of reference is provided by comparison to the known σ values for nucleic acids and phospholipid bilayers, whose surface charge densities are at least an order of magnitude greater than that of proteins at their pH_{cryst}. One problem with this frame of reference is that there is no pH at which proteins are as highly charged as nucleic acids. A more serious problem is that nucleic acids crystallize readily, demonstrating that a high σ is not a barrier to crystallization. Another frame of reference is in relation to the mean $\overline{Q}/\sigma$ values at "physiological pH" of 7.4 ($\overline{Q}_{pH=7.4}$); here, there is an approximately three-fold reduction in the mean of $\overline{Q}_{cryst}$ vs. the mean of $\overline{Q}_{pH=7.4}$. One problem with this frame of reference is that the pH reference value, 7.4, is questionable because many proteins function in physiological compartments where the pH is significantly different. A more serious problem is that the standard deviations of the two distributions are more than twice the shift between them. The shift is statistically significant because of the sample size, but cannot be used to make meaningful predictions about specific proteins. It should be noted that the preponderance of "domain knowledge" would be that the primary factors selecting low net charge are issues of protein folding and stability as well as issues of functionality.

It is interesting to note that in two studies (Dasgupta et al., 1997; Bahadur et al., 2003) that 4/5 of the most frequent residues found at crystal contacts were Aspartic acid, Glutamic acid, Lysine, and Arginine, which have side chains that can be ionized. A low $\overline{Q}_{cryst}$ or σ_{cryst} is consistent with the electrostatic effects being primarily local. Here it is argued that a low $\overline{Q}_{cryst}$ or σ_{cryst} implies a low probability of like charges in close proximity across a lattice contact. Even though the $\overline{Q}_{cryst}$ and σ_{cryst} can be largely controlled by the solution pH, there are other solution parameters involved.

8.7 APPLICATION TO SCREEN DESIGN

The above findings in possible combination with other case-based reasoning (Hennessy et al., 2000; Jurisica et al., 2001**)** and Bayesian methods (Hennessy et al., 2000; Rupp and Wang, 2004) may help lead to the development of more successful crystallization screens. Historically, crystallization conditions were initially searched in a lab specific manner. This changed with the development of the sparse matrix crystallization screens by Jancarik and Kim (1991). Designed using conditions that had succeeded in crystallizing other proteins, this screen initially sampled pH values approximately every 1.0 unit from 4.5 to 8.5 (4.6, 5.6, 6.5, 7.5, and 8.5). Using this method, a large number of conditions are examined by varying the partial combinations of the concentrations and type of salt, buffer, and precipitants. These screens caught on and were commercialized and are currently used by many researchers due to their ease of use and proven track record.

Since Jancarik and Kim's original sparse matrix screen, others have developed more specific sparse matrix screens for particular classes of biological molecules, such as RNA (Doudna et al., 1993; Scott et al., 1995), immunoglobulins (Harris et al., 1995), enzymes (Brzozowski and Walton, 2001), protein-protein complexes (Radaev and Sun, 2002; Radaev et al., 2006), and membrane proteins (Iwata, 2003). These studies and others demonstrated that particular classes of macromolecules have a preference for solution conditions that lead to successful crystallization (Samudzi et al., 1992; Farr et al., 1998; Hennessy et al. 2000; Jurisica et al., 2001; Gilliland et al. 2002; Kimber et al. 2003; Goh et al., 2004; Rupp and Wang, 2004). For example, Hennessey et al. (2000) found that ligand binding proteins and enzymes had significantly different pH_{cryst} distributions.

This prior knowledge can be encoded as Bayesian priors and used to generate probability distributions for various solution parameters, including the solution pH. These probabilities can be further combined over multiple crystallization variables as more data are collected to suggest regions in the crystallization search space more likely to produce well-ordered crystals. Even a collection of weak predictors can be informative and aid in crystallization design (Hennessy et al., 2000).

The present results support the idea of using prior knowledge to improve the probability of generating crystals from an initial screen. From an amino acid sequence, an estimated

titration curve can be easily calculated. This curve can then be translated into a $\overline{Q}$ or σ curve by taking the protein's MW or A_S into account. The $\overline{Q}_{cryst}$ or σ_{cryst} frequency distributions of the successfully crystallized proteins can then be used to suggest a pH range for the target protein(s). This can be accomplished by matching the pH values that result in a $\overline{Q} = \overline{Q}_{cryst}$ or $\sigma = \sigma_{cryst}$ value that has a high probability of generating a crystal.[2] However, careful consideration should be given to counter ions and other means of achieving neutrality, in addition to the solution pH, when the protein is highly acidic or basic. Several examples demonstrating possible scenarios for crystallizing a test protein are presented in Appendix C.

A protein from the test set, 1LRH (Auxin-Binding Protein 1), was used in the example. 1LRH was an example where the predicted pH_{cryst} range was extremely accurate. The CI_{50} range for $\overline{Q} = \overline{Q}_{cryst}$ was -0.25 to +0.14 *e*/kDa, which results in a 50/50 chance that the $\overline{Q}_{cryst}$ will fall within this range. When translated back into pH space, a pH range of ~4.8-6.0 was suggested for initial crystallization attempts, while the pH_{cryst} was 5.5. When such narrow distributions are predicted, a focused pH search (5.0-6.0) for the initial crystallization attempts is suggested.

Some proteins, such as UDP-N-Acetylglucosamine Enolpyruvyl Transferase (Figure 2.1; PDB id: 1A2N), have long flat portions along the estimated titration curve where the charge changes very slowly over a wide range of pH values. If a protein crystallizes in this flat region, it may be relatively insensitive to the solution pH and therefore crystallize over a wide range of pH values, all of which result in a similarly charged protein. If the protein crystallizes in an area of rapidly changing charge, near the pK_a values of the charged amino acids, then it may be highly sensitive to manipulation of solution pH. If a crystal produces a 'hit' in this region, a finer search of pH conditions would be suggested. These methods may be relatively effective in limiting the range of pH values searched in an initial crystallization screen. Alternatively, there may little reduction in pH space. When a significant reduction is observed in the pH range, the researcher will have greater flexibility in searching other parameters, such as the salt or precipitating agent, while using the same number of experiments.

[2] The findings and interpretation could differ based on the way these probabilities were propagated.

8.8 CONCLUSION

The goal of the research was to develop a Protein-Specific Properties Evaluation (PSPE) framework that could aid in the statistical evaluation of variables for predicting ranges of and prior probability distributions for protein crystallization conditions. Development of such a framework was motivated by the rapid growth and evolution of the Protein Data Bank. Features of the framework that has been developed include: (1) it is an instantiation of the "scientific method" for the framing and testing of hypotheses in an informatics setting, (2) the use of hidden variables, *i.e.* parameters which are analytic functions of quantities extracted from the database, and (3) a negative result is still useful; *i.e.* the recognition that a variable has minimal utility in predicting ranges or probabilities of crystallization allows energy and attention to be focused elsewhere, where it may be more productively employed.

The hidden variables examined in this study were related to the estimated net charge (Q) of the proteins under consideration. The Q is a function of the amino acid composition, the pH of the solution, and the assumed pKa values for the titratable amino acid residues. The size of the protein clearly has a significant impact on the magnitude of the Q. Therefore, two additional variables were introduced to mitigate this effect. The first variable was the estimated specific charge ($\overline{Q}$), which is the ratio of the Q to the protein mass, expressed here in units of *e*/kDa. The second variable was the estimated surface charge density (σ), which is the ratio of the Q to the estimated surface area of the protein; a convenient unit is m*e*/nm^2. Although the estimation of surface area is difficult, σ facilitates comparisons with other biologically relevant macromolecules. Additional Q-related quantities examined included the isoelectric point (pI), which is the pH at which the Q is zero, and measures of the shape of the titration curve, which is either $\overline{Q}$ or σ expressed as a function of the pH.

The principal observation of this study is that proteins appear to crystallize at low values of $\overline{Q}$ and σ. One problem with this observation is that "low" is a relative term and the frame of reference requires careful examination. One frame of reference is provided by comparison to the known σ values for nucleic acids and phospholipid bilayers, whose σ values are an order of magnitude greater than that of proteins. One problem with this frame of reference is that there are no pH values at which proteins are as highly charged as nucleic acids. A more serious

problem is that nucleic acids crystallize readily, demonstrating that high σ is not a barrier to crystallization.

Another frame of reference is in relation to the mean $\overline{Q}/\sigma$ values at "physiological pH" of 7.4 ($\overline{Q}_{pH=7.4}$). Here, there is an approximately three-fold reduction in the mean of $\overline{Q}_{cryst}$ (-0.6 *e*/kDa) vs. the mean of $\overline{Q}_{pH=7.4}$ (-0.17 *e*/kDa). One problem with this frame of reference is that the pH reference value, 7.4, is questionable because many proteins function in physiological compartments where the pH is significantly different. A more serious problem is that the standard deviations of the two distributions are more than twice the shift between them. The shift is statistically significant because of the sample size, but cannot be used to make meaningful predictions about specific proteins. It should be noted that the preponderance of "domain knowledge" would be that the primary factors selecting low net charge are issues of protein folding and stability as well as issues of functionality.

The results are sufficiently weak that no protein-specific prospective predictions appear possible although information of this type could be included with other weak predictors in a Bayesian predictor scheme. Additional work would be required to establish this; however that work is beyond the scope of the dissertation. Although many statistically significant correlations among Q-related quantities were noted, no evidence could be developed to suggest they were anything other than those expected from the additional information introduced with the hidden variables.

Thus, the principal conclusions of this PSPE analysis are: (1) the $\overline{Q}/\sigma$ and other Q-related variables are of limited value as prospective protein-specific predictors of ranges of values of crystallization conditions, (2) probabilities based on these variables may prove useful in a Bayesian sense, although that has not been demonstrated at this time, and (3) although this is a negative result, it is still useful in that it allows attention to be directed into more productive avenues.

9.0 FUTURE RESEARCH

The initial work presented here focused on the reported solution pH of crystallization (pH_{cryst}) for several reasons. The first reason was that the pH_{cryst} was the most reported crystallization parameter reported in the PDB. The second reason was that the solution pH has long been known to be one of the primary variables responsible for crystallization. The solution pH exhibits its primary influence by controlling the estimated net charge of the protein. For this dissertation, the *Controllables* were assumed to be independent variables. However, this is not the case as there are many types of interactions between the parameters, both known and unknown.

It was hypothesized in this dissertation that the Protein Sequence-Properties Evaluation Framework (PSPE) could be used to identify other important relationships between *Observables* and *Features*. There remain a lot of possible variables to examine (see Table 1.1 in Chapter 1); some are discussed in Sections 9.1-9.3. Additionally, other types of analyses can be examined, including multiparametric methods to examine the interactions among variables (Section 9.4), provided enough data is available (see Section 8.1).

In addition to other types of analyses, the same methods used here with some minor modifications might be able to be used to further increase the predictive range results (Section 9.5). Adjustments could be made to the calculation of estimated titration curves, such as adjusting for the propensity of the amino acid being found on the surface or at a protein-protein interface. Although the methods discussed within this dissertation might lead to an increase of the success rates of obtaining protein crystals, actual experimental verification is essential and should be explored further (Section 9.6).

9.1 OTHER CRYSTALLIZATION CONTROLLABLES AND OBSERVABLES

This analysis focused on the solution pH, but any of the other *Controllables* could be examined and analyzed if sufficient information is present as crystallization *Observables*. Although more information is accumulating on the crystallization conditions (*Observables*), much of it is in the form of free text. This presents some problems due to all the possible lexicons. For example, Peat et al. (2005) examined the precipitating agents used as listed in the PDB and found over 30 possible terms for "ammonium sulfate," many of which are misspellings or abbreviations. There would also be the possibility to develop methods to extract the information from the original research articles, but that would be an even greater challenge.

The temperature at which crystals grew is the next most reported variable and can now be queried from the PDB. The reported temperature is currently available (5/26/2006) for 57% of the crystal structures solved by X-ray diffraction (19,723/34,578). However, the bias in temperature selection is much greater than pH. Most researchers examine temperature effects at either room temperature (20 or 25 ^{o}C) or in a cold room/refrigerator (4 ^{o}C). Although the reporting of temperature is supposed to be in degrees Kelvin, there are many instances where the temperature is clearly in degrees Celsius, values of 4, 18, and 20.

9.2 OTHER PROTEIN FEATURES

In addition to analyzing other *Observables*, the effects of other *Features* can also be examined. Other *Features* that may prove valuable are measures of hydrophobicity, Aliphatic Index (stability), and Disorder. Because these *Features* are also based on sequence information, they may also be correlated to sequence-based *Observables*, such as the Q-related *Observables* examined in this dissertation. Additional variables such as the predicted secondary structure composition, i.e. Helix and Sheet, and subcellular location may also prove useful to include. As mentioned earlier in Section 2.2, these variables have correlated with success (0='No' or 1='Yes') of the structure determination process. Therefore, it would be expected that their values may indicate what crystallization conditions to use.

Another factor that may influence the ability of the protein to crystallize is the source of the protein. The source of each protein chain found in the PDB is available from PDBeast at http://www.ncbi.nih.gov/Structure/PDBEAST/pdbeast.shtml. It should be noted that there may be multiple sources for each crystallized complex. It is hypothesized that there may be differences in protein behavior based upon the Kingdom (Archaea, Bacteria, Eukaryota, or Virus).

Previous work has also suggested that certain subclasses of proteins crystallize under different conditions. Analysis could include identification of protein families, as defined in Pfam (Bateman et al., 2004), proteins with similar function, as defined by Gene Ontology (GO; The Gene Ontology Consortium, 2000), and other biological macromolecules. Although these analyses may prove fruitful, preliminary work on the identification and frequencies need to be developed.

9.3 OTHER BIOLOGICAL MACROMOLECULES

This analysis focused only on forming crystals from non-membrane proteins. There is no reason to believe that these methods could not be applied to NMR studies or membrane proteins. Sparse-matrix screens have been devised for membrane proteins, demonstrating that membrane proteins also have areas in the crystallization search space that have higher probabilities of generating crystals. Analysis on membrane proteins can even include predicting type of detergent if enough information is available.

The PSPE framework should also work with nucleic acids or DNA/RNA-protein complexes. Similar to membrane proteins, sparse-matrix screens have also been developed for these particular classes of biomacromolecules. It would be hypothesized that nucleic acids or these complexes may display similar crystallization behavior that sets them apart from other biological macromolecules even though the charged nature of nucleic acids may complicate analysis.

9.4 OTHER ANALYSIS METHODS

Indications are that the primary sequence composition may be used to predict the $\overline{Q}$ ranges that are more likely to result in the crystallization of a protein. Binning and two unsupervised clustering methods were used to group proteins by similarity. Although each method has shown promise there is no reason to believe that there may be better methods available for analysis. One method that should be more thoroughly explored is case-based reasoning, which has previously been shown to be useful for selecting crystallization conditions (Hennessey et al., 2000; Jurisica et al., 2001). Other methods that predict a single *Observable* value can be explored, such as Linear Regression, Neural Networks, and Decision Trees. Preliminary results for Linear Regression and Neural Networks are discussed in Appendix E. These analyses can be explored using the Charge Range Test and compared to the methods described in this dissertation.

9.5 CORRECTION FACTORS

Several modifications were discussed in Chapter 8 that could be explored to improve prediction of the $\overline{Q}_{cryst}$ ranges, which can then be translated into the pH_{cryst}. An example would be to make adjustments in the calculation of estimated titration curves for the amino acid's propensity for surface or protein-protein interface. Modification can also be made based on whether the multimeric proteins are homo-oligomeric or hetero-oligomeric proteins. Other adjustments in charge can also be made for the binding of metal ions (positive charge), some of which would be known apriori, such as the binding of iron to hemoglobin. Removal of the extreme cases (outliers) and using the next best example in the nrPDB may also slightly increase the prediction accuracy.

9.6 EXPERIMENTAL VERIFICATION

A critical factor for validation would be the actual crystallization of a target protein that has proved difficult to crystallize. Another possible form of validation would be to take some proteins from the PDB with low quality measures (poor resolution, r-factors, etc.) and attempt to improve quality by recrystallizing them using the predictions made using the models developed in this dissertation. The quickest way to do this would be pre-selecting proteins that have narrow regions of high probability. Finally, experiments can be performed using proteins that crystallize over a wide range of pH values. The actual titration curves can be determined and the slopes of the titration curve where the protein crystallizes can be examined.

APPENDIX A: Python Scripts

PYTHON SCRIPTS

```
# The purpose of this program is to identify PDB ids and ALL OF THEIR SUBUNITS/CHAINS
# and calculate some parameters based on protein sequence - amino acid counts, pKa curves, estimated pI
# slope at the theoretical pI

#copyright 2004 David S. Dougall and University of Pittsburgh, Pittsburgh, PA 15260

import re, string, os
from Bio import Fasta
class PDB:                  # create a class PDB that represents one PDB entry
        def __init__(self, id, subunit, type, length, sequence):
                self.id = id
                self.subunit = subunit
                self.type = type
                self.length = length
##              self.description = description
                self.sequence = sequence
                self.ala = 0            #ALA    ALANINE
                self.cys = 0            #CYS    CYSTEINE
                self.asp = 0            #ASP    ASPARTIC ACID
                self.glu = 0            #GLU    GLUTAMIC ACID
                self.phe = 0            #PHE    PHENYLALANINE
                self.gly = 0            #GLY    GLYCINE
                self.his = 0            #HIS    HISTIDINE
                self.ile = 0            #ILE    ISOLEUCINE
                self.lys = 0            #LYS    LYSINE
                self.leu = 0            #LEU    LEUCINE
                self.met = 0            #MET    METHIONINE
                self.asn = 0            #ASN    ASPARAGINE
                self.pro = 0            #PRO    PROLINE
                self.gln = 0            #GLN    GLUTAMINE
                self.arg = 0            #ARG    ARGININE
                self.ser = 0            #SER    SERINE
                self.thr = 0            #THR    THREONINE
                self.val = 0            #VAL    VALINE
                self.trp = 0            #TRP    TRYPTOPHAN
                self.tyr = 0            #TYR    TYROSINE
                self.x = 0                      #Xaa    Unknown AA
                self.total = 0
```

```
        self.pos_charge = 0         # lysine + histidine + arginine
        self.neg_charge = 0         # aspartic acid + glutamic acid + cysteine + tyrosine
        self.abs_charge = 0         # absolute value of charge i.e. neg + pos charge (sum of charges)
        self.mw = 0.0               # molecular weight (kDa)
        self.start = ''             # start = first amino acid
        self.end = ''               # end = last amino acid
        self.pi = 0                 # estimated pI
        self.slope_pi = []          # pk curve slope at the pI
        self.pk_list = []           # create an empty list to store net charge values ph 1-13 by 0.2
        self.protein = 0            # binary flag for type = protein
        self.protein_het = 0        # binary flag for type = protein-het
        self.nucleic = 0            # binary flag for type = nucleic
        self.nucleic_het= 0         # binary flag for type = nucleic-het
        self.asa = 0.0              # solvent accessible surface area
        self.num_chains = 1         # number of chains in the asymmetric unit

def sequence_parser(self):
        # this subroutine will parse the biological sequence (i.e. amino acids)
        # COUNT ALL OF THE 20 AMINO ACIDS IN THE SEQUENCE
        # also added flags for molecular type
        if curPDB.type == 'protein':
                curPDB.protein = 1
# type = protein present in PDB entry
        elif curPDB.type == 'protein-het':
                curPDB.protein_het = 1          # type = protein-het present in PDB entry
        elif curPDB.type == 'nucleic':
                curPDB.nucleic = 1              # type = nucleic present in PDB entry
        elif curPDB.type == 'nucleic-het':
                curPDB.nucleic_het = 1          # type = nucleic-het present in PDB entry
        if (curPDB.protein or curPDB.protein_het):          # parse only types of protein or protein-het
                aa_list = ['A', 'C', 'D', 'E', 'F', 'G', 'H', 'I', 'K', 'L', 'M', 'N', 'P', 'Q', 'R', 'S', 'T', 'V', 'W','Y', 'X']
                # list of amino acids
                self.ala = string.count(self.sequence, 'A')         #ALA   ALANINE
                self.cys = string.count(self.sequence, 'C')         #CYS   CYSTEINE
                self.asp = string.count(self.sequence, 'D')         #ASP   ASPARTIC ACID
                self.glu = string.count(self.sequence, 'E')         #GLU   GLUTAMIC ACID
                self.phe = string.count(self.sequence, 'F')         #PHE   PHENYLALANINE
                self.gly = string.count(self.sequence, 'G')         #GLY   GLYCINE
                self.his = string.count(self.sequence, 'H')         #HIS   HISTIDINE
                self.ile = string.count(self.sequence, 'I')         #ILE   ISOLEUCINE
                self.lys = string.count(self.sequence, 'K')         #LYS   LYSINE
                self.leu = string.count(self.sequence, 'L')         #LEU   LEUCINE
                self.met = string.count(self.sequence, 'M')         #MET   METHIONINE
                self.asn = string.count(self.sequence, 'N')         #ASN   ASPARAGINE
                self.pro = string.count(self.sequence, 'P')         #PRO   PROLINE
                self.gln = string.count(self.sequence, 'Q')         #GLN   GLUTAMINE
                self.arg = string.count(self.sequence, 'R')         #ARG   ARGININE
                self.ser = string.count(self.sequence, 'S')         #SER   SERINE
                self.thr = string.count(self.sequence, 'T')         #THR   THREONINE
                self.val = string.count(self.sequence, 'V')         #VAL   VALINE
                self.trp = string.count(self.sequence, 'W')         #TRP   TRYPTOPHAN
                self.tyr = string.count(self.sequence, 'Y')         #TYR   TYROSINE
                self.x  = string.count(self.sequence, 'X')   #Xaa   ANY AA

                self.total = self.ala + self.cys + self.asp + self.glu + self.phe + self.gly +\
                        self.his + self.ile + self.lys + self.leu + self.met + self.asn +\
```

```
                        self.pro + self.gln + self.arg + self.ser + self.thr + self.val + \
                        self.trp + self.tyr + self.x
                self.pos_charge = self.lys + self.his + self.arg        # lysine + histidine + arginine
                self.neg_charge = self.asp + self.glu + self.cys + self.tyr     # aspartic acid + glutamic
acid + cysteine + tyrosine
                self.mw = (self.ala*89.09 + self.cys*121.15 + self.asp*133.10 + self.glu*147.13 +\
                        self.phe*165.19 + self.gly*75.07 + self.his*155.16 + self.ile*131.17 +\
                        self.lys*146.19 + self.leu*131.17 + self.met*149.21 + self.asn*132.12 +\
                        self.pro*115.13 + self.gln*146.15 + self.arg*174.20 + self.ser*105.09 +\
                        self.thr*119.12 + self.val*117.15 + self.trp*204.23 + self.tyr*181.19 +
self.x*136.9)-(self.total-1)*18.02
                # x value for molecular weight is the average mw of all 20 amino acids
                # this is in daltons.  it will be divided by 1000 when writing the data to get kDa
                # subtract 18.02 for every protein bond - 1 H2O
                self.start = self.sequence[0]                   # save first amino acid
                self.end  = self.sequence[self.total-1]         # save last amino acid

        def pi_calc(self, start_ph, end_ph, ph_step, save, slope):
                # Calculate the estimated pI
                # I use recursion to narrow down the pI (successive approximation)
                # all pk values taken from Harpers Biochemistry (e-book) through HSLS
                # these values compare very closely to Voet & Voet Biochemistry, 1990
                #seq_list = ['A, 'C' 'D', 'E',  'F',  'G',  'H',  'I',  'K',  'L',
                #                        'M', 'N',  'P',  'Q',  'R',  'S',  'T',  'V',  'W',  'Y']
                pk_nh3 =  [ 9.69, 10.28, 9.60, 9.67, 9.00, 9.60, 9.17, 9.68, 8.95, 9.60, \
                                        9.21, 8.80, 10.96, 9.13, 9.04, 9.15, 9.62, 9.62, 9.39, 9.11]
                pk_coo3 = [     2.34, 1.96, 1.88, 2.19, 1.83, 2.34, 1.82, 2.36, 2.18, 2.36, \
                                        2.28, 2.02, 1.99, 2.17, 2.17, 2.21, 2.11, 2.32, 2.38, 2.20]
                pk_side = [8.18, 3.65, 4.25, 10.07, 6.0, 10.53, 12.48]
                # pk_side C, D, E, Y, H, K, R
                low_ph = 0
                ph = start_ph
                while ph <= end_ph:
                        c_parts01 = 10**(ph - pk_side[0])               # cysteine
                        c_zi = c_parts01 /(c_parts01 + 1)
                        c_nzi = c_zi * self.cys

                        d_parts01 = 10**(ph - pk_side[1])               # aspartic acid
                        d_zi  = d_parts01 /(d_parts01 + 1)
                        d_nzi = d_zi * self.asp

                        e_parts01 = 10**(ph - pk_side[2])               # glutamic acid
                        e_zi  = e_parts01 /(e_parts01 + 1)
                        e_nzi = e_zi * self.glu

                        y_parts01 = 10**(ph - pk_side[3])               # tyrosine
                        y_zi  = y_parts01 /(y_parts01 + 1)
                        y_nzi = y_zi * self.tyr

                        neg_charge_sum = c_nzi + d_nzi + e_nzi + y_nzi  # sum negative side chain aa

                        h_parts01 = 10**(ph - pk_side[4])               # histidine
                        h_zi = 1 /(h_parts01 + 1)
                        h_nzi = h_zi* self.his

                        k_parts01 = 10**(ph - pk_side[5])               # lysine
```

```
                k_zi = 1 /(k_parts01 + 1)
                k_nzi = k_zi* self.lys

                r_parts01 = 10**(ph - pk_side[6])                    # arginine
                r_zi = 1 /(r_parts01 + 1)
                r_nzi = r_zi* self.arg

                pos_charge_sum = h_nzi + k_nzi + r_nzi                # sum positive side chain aa
                aa_list = ['A', 'C', 'D', 'E', 'F', 'G', 'H', 'I', 'K', 'L', 'M', 'N', 'P', 'Q', 'R', 'S', 'T', 'V', 'W', 'Y']
                start_zi = 0
                end_zi = 0
                count = 0
                for aa in aa_list:
                        for start_aa in self.start:
                                if start_aa == aa:
                                        # start amino acid - NH3 end
                                        start_parts01 = 10**(ph - pk_nh3[count])
                                        start_zi = start_zi + 1 /(start_parts01 + 1)
                        count = count +1
                count = 0
                for aa in aa_list:
                        for end_aa in self.end:
                                if end_aa == aa:
                                        # end amino acid - Carboxyl end
                                        end_parts01 = 10**(ph - pk_coo3[count])
                                        end_zi = end_zi + end_parts01 /(end_parts01 + 1)
                        count = count + 1
                # calculate charges (net, sum of squares, and total charges)
                self.total_charge = (start_zi + pos_charge_sum - neg_charge_sum - end_zi)
                # append charge values to respective lists
                if save:
                        self.pk_list.append(str(self.total_charge) )    # add total (net) charge to list
                if self.total_charge > 0:
                        low_ph = ph
                # obtain lowest positive charge ph
                ph = ph + ph_step
        # increase pH to next step
        if not slope:
                if ph_step > 0.0001:
                        # don't calculate steps too small
                        self.pi_calc(low_ph, low_ph + ph_step, ph_step/2, 0, 0)
                # recursion - will decrease the range of ph values to look at and decrease the step
                else:
                        self.pi = low_ph

def pi_slope(self):
        # this function is for calculating the slope around the pI
        # pi_calc(self, start_ph, end_ph, ph_step, save, slope)
        low_value  = self.pi - 0.1
        high_value = self.pi + 0.1
        self.slope_pi.append(str(low_value))
        self.slope_pi.append(str(high_value))
        self.pi_calc(low_value, high_value, 0.1, 1, 1)

def asa_calc(self):
        # this function calculates the solvent accessible surface area of the structure
```

```
                # uses monomer equation (Miller Janin, Lesk, & Chothia (1987) JMB 196: 641656
                # subtracts 'standard-size' interface of 1600 Angstroms squared per every additional chain
                #(Lo Conte, Chothia & Janin (1999) JMB 285: 2177-2198)
                self.asa = 6.3*((self.mw)**0.73) - (self.num_chains-1)*1600

        def write_data(self):
        # this is my function to write the data to file
                list = [self.id, '\t', self.subunit, '\t', self.type, '\t', self.num_chains, '\t', self.length, '\t', \
                                self.sequence, '\t', self.ala, '\t', self.cys, '\t', self.asp, '\t', \
                                self.glu, '\t', self.phe, '\t', self.gly, '\t', self.his, '\t', \
                                self.ile, '\t', self.lys, '\t', self.leu, '\t', self.met, '\t', \
                                self.asn, '\t', self.pro, '\t', self.gln, '\t', self.arg, '\t', \
                                self.ser, '\t', self.thr, '\t', self.val, '\t', self.trp, '\t', \
                                self.tyr, '\t', self.x, '\t', \
                                self.pos_charge, '\t', self.neg_charge, '\t',\
                                self.mw/1000, '\t', self.asa, '\t', self.start, '\t', self.end, '\t', self.pi, '\t',\
                                self.protein, '\t', self.protein_het, '\t', self.nucleic, '\t', self.nucleic_het, '\t']
                for value in list:
                        output.writelines(str(value))
                for value in self.pk_list:                      # write net charge (pk curve) values
                        output.writelines(str(value))
                        output.writelines('\t')
                for value in self.slope_pi:                     # write values for calculation of the slope at the pI
                        output.writelines(str(value))
                        output.writelines('\t')
                output.writelines('\n')

def title_parser(fasta_record):
        #this function will break up the fasta title into parts found in the PDB seqres file
        title = re.compile(r'(?P<pdb_id>[a-zA-Z0-9]+)_'
                                        r'(?P<subunit>[a-zA-z\:0-9\!\_\-\+\#\=\:\>\<\|]+)?  ?mol:'
                                        r'(?P<type>[a-zA-Z\-]+) length:'
                                        r'(?P<length>[0-9]+) ? ? ? ? ?'
                                        r'(?P<descrip>[a-zA-Z0-9,-:;\[\]\(\)\'_*=\.\/ ]+)')
        result = title.search(fasta_record.title)               # search only in title line
        pdb_id = result.group('pdb_id')                         # PDB ID
        subunit = result.group('subunit')                       # protein subunit
        if subunit == None:                                     # if no subunit set subunit to 99
                subunit = "99"
        type = result.group('type')                             # type: protein, nucleic, et al.
        length = result.group('length')                         # LENGTH OF PROTEIN IN AMINO ACIDS
        description = result.group('descrip')                   # DESCRIPTION
        outputList = []                                         # initialize a list named outputList - to return a list
        outputList.append(pdb_id)
        outputList.append(subunit)                              # 1
        outputList.append(type)                                 # 2
        outputList.append(length)                               # 3
##              outputList.append(description)                  # 4  # I don't save currently
        outputList.append(fasta_record.sequence)                # 5
        return(outputList)                                      # return a list for subunit parsing

##############################################################################################
##############################  Main Program #################################################
##############################################################################################
```

```
os.chdir(r'C:\Documents and Settings\dougalld\My Documents\School Material\PDB')  # change directory to where PDB_SEQRES.txt file found
# this is a list of all PDB files in fasta format - i.e. sequence information
outputFile = os.path.join("c:\\", "Documents and Settings", "dougalld", "My Documents", "School Material", "PDB", "pdb_012006_aa_vector_protein.txt")  # set output file path
output = open(outputFile, 'w+')                         # open output file for fasta file parse
#caption = "PDB_ID\tsubunit\tType\tTotal_charge\tPos_charge\tNeg_charge\tSS_charge\n"
# for sum of square charge calc
caption = "PDB_ID\tsubunit\tType\tnum_chains\taa_length\tSequence\tA\tC\tD\tE\tF\tG\tH\tI\tK\tL\tM\tN\tP\tQ\tR\tS\tT\tV\tW\tY\tX\tpos_charge\tneg_charge\tmw\tasa\tstart_aa\tend_aa\ttheor_pi\tp_flag\tp-het_flag\tn_flag\tn-het_flag\t"
        # make caption a string for writing
output.writelines(caption)          # column headings printed to file
a = 1.0
line_list = [ ]                     # this will containt column headings
# net charge header
while a <= 14:
        line_list.append(('pk_' + str(a)))
        line_list.append('\t')
        a = a + 0.1
line_list.append('net_low\tnet_pi\tnet_high\tph_low\tph_high\t')
line_list.append('\n')
for line in line_list:              # this method or below both work
        output.write(line)          # column headings printed to file
end_ph = 14                         # initial maximum ph value
start_ph = 1                        # initial minimum ph value
ph_step = 0.1                       # original ph step
parser = Fasta.RecordParser()
file_name = open('pdb_seqres_013105.txt') # file with full list of fasta files#
iterator = Fasta.Iterator(file_name, parser)
#print file_name
pdb_id = [ ]                                # create empty list named pdb_id
theor_pi = []                               # initialize empty list named theor_pi
prevPDB = PDB(0, 0, 0, 0, 0)                # initialize empty PDB file class
save = 1
slope = 0
while 1:
        cur_record = iterator.next()        # obtain first FASTA record
        if cur_record is None:              # after last fasta pdb_id will break
                prevPDB.pi_calc(start_ph, end_ph, ph_step, save, slope)     # calculate theoretical pI
                prevPDB.pi_slope()          # obtain values to calculate slope
                prevPDB.asa_calc()          # calculate the solvent accessible surface area
                prevPDB.write_data()        # save data to file
                break

        pdb_id = title_parser(cur_record)   # sends current record for title parsing
        # returns a list with [0-4] values
        if pdb_id not in (0,'0'):
                curPDB = PDB(pdb_id[0], pdb_id[1], pdb_id[2], pdb_id[3], pdb_id[4])#, pdb_id[5])
                curPDB.sequence_parser()
                if (prevPDB.id == curPDB.id):
                                # try to identify if subunits present by seeing multiple pdb_id entries
                        curPDB.subunit = prevPDB.subunit + curPDB.subunit
                                # put subunits together into one value
                        if curPDB.type != prevPDB.type:             # see if different types
                                curPDB.type = prevPDB.type + curPDB.type
```

```
                                # put types together into one value
                        if curPDB.protein < prevPDB.protein:
                                curPDB.protein = prevPDB.protein
                                # type = protein present in PDB entry
                        if curPDB.protein_het < prevPDB.protein_het:
                                curPDB.protein_het = prevPDB.protein_het
                                # type = protein-het present in PDB entry
                        if curPDB.nucleic < prevPDB.nucleic:
                                curPDB.nucleic = prevPDB.nucleic
                                # type = nucleic present in PDB entry
                        if curPDB.nucleic_het < prevPDB.nucleic_het:
                                curPDB.nucleic_het = prevPDB.nucleic_het
                                # type = nucleic-het present in PDB entry
                        if curPDB.nucleic != 1 and curPDB.nucleic_het!= 1:
                                curPDB.length = int(curPDB.length) + int(prevPDB.length)
                                # adjust length of protein by new subunit's aa length
        ##              if curPDB.description != prevPDB.description:
        ##                      curPDB.description = prevPDB.description + curPDB.description
                        curPDB.sequence = prevPDB.sequence + curPDB.sequence
                                # put aa sequences together into one value
                        curPDB.ala = prevPDB.ala + curPDB.ala           #ALA    ALANINE
                        curPDB.cys = prevPDB.cys + curPDB.cys           #CYS    CYSTEINE
                        curPDB.asp = prevPDB.asp + curPDB.asp           #ASP    ASPARTIC ACID
                        curPDB.glu = prevPDB.glu + curPDB.glu           #GLU    GLUTAMIC ACID
                        curPDB.phe = prevPDB.phe + curPDB.phe           #PHE    PHENYLALANINE
                        curPDB.gly = prevPDB.gly + curPDB.gly           #GLY    GLYCINE
                        curPDB.his = prevPDB.his + curPDB.his           #HIS    HISTIDINE
                        curPDB.ile = prevPDB.ile + curPDB.ile           #ILE    ISOLEUCINE
                        curPDB.lys = prevPDB.lys + curPDB.lys           #LYS    LYSINE
                        curPDB.leu = prevPDB.leu + curPDB.leu           #LEU    LEUCINE
                        curPDB.met = prevPDB.met + curPDB.met           #MET    METHIONINE
                        curPDB.asn = prevPDB.asn + curPDB.asn           #ASN    ASPARAGINE
                        curPDB.pro = prevPDB.pro + curPDB.pro           #PRO    PROLINE
                        curPDB.gln = prevPDB.gln + curPDB.gln           #GLN    GLUTAMINE
                        curPDB.arg = prevPDB.arg + curPDB.arg           #ARG    ARGININE
                        curPDB.ser = prevPDB.ser + curPDB.ser           #SER    SERINE
                        curPDB.thr = prevPDB.thr + curPDB.thr           #THR    THREONINE
                        curPDB.val = prevPDB.val + curPDB.val           #VAL    VALINE
                        curPDB.trp = prevPDB.trp + curPDB.trp           #TRP    TRYPTOPHAN
                        curPDB.tyr = prevPDB.tyr + curPDB.tyr           #TYR    TYROSINE
                        curPDB.x   = prevPDB.x   + curPDB.x             #Xaa    ANY AA
#                       curPDB.total = prevPDB.total + curPDB.total     # adjust total aa length
                        curPDB.pos_charge = curPDB.pos_charge + prevPDB.pos_charge
                                # lysine + histidine + arginine
                        curPDB.neg_charge = curPDB.neg_charge + prevPDB.neg_charge
                                # aspartic acid + glutamic acid + cysteine + tyrosine
                        curPDB.mw = curPDB.mw + prevPDB.mw
                        curPDB.start = curPDB.start + prevPDB.start     # put start amino acids together
                        curPDB.end   = curPDB.end + prevPDB.end         # put end amino acids together
                        curPDB.num_chains = prevPDB.num_chains +1       # add 1 to number of chains
                        prevPDB = curPDB
        # assign prev_id = pdb_id for adjustment of parameters
                else:
                        if prevPDB.id not in (0, '0'):   # so initial value of prev_id not sent to seq_parser function
                                if prevPDB.protein_het < 1:
                                        if prevPDB.nucleic < 1:
```

```
                                                    if prevPDB.nucleic_het < 1:
                                                            prevPDB.pi_calc(start_ph, end_ph, ph_step, save,
slope)          # calculate pI
                                                            prevPDB.pi_slope()  #obtain values to calculate slope
                                                            prevPDB.asa_calc()
# calculate solvent accessible surface area
                                                            prevPDB.write_data()    # save data to file
                                prevPDB = curPDB
        # current pdb_id now becomes previous pdb_id - works
output.close()
file_name.close().
```

calc_charge.py

```
# The purpose of this program is to calculate the slope of the pk_curve at the experimental ph of all pdb ids and
# ALL OF THEIR SUBUNITS based on amino acid counts and starting and end amino acids
import re, string, os
class PDB:                              # create a class PDB that represents one PDB file
        def __init__(self):
                self.id = 'X'           # PDB ID
                self.mw = 0.0           # Molecular Weight
                self.exp_ph = 0.0 # reported ph of crystallization value
                self.ala = 0            #ALA    ALANINE
                self.cys = 0            #CYS    CYSTEINE
                self.asp = 0            #ASP    ASPARTIC ACID
                self.glu = 0            #GLU    GLUTAMIC ACID
                self.phe = 0            #PHE    PHENYLALANINE
                self.gly = 0            #GLY    GLYCINE
                self.his = 0            #HIS    HISTIDINE
                self.ile = 0            #ILE    ISOLEUCINE
                self.lys = 0            #LYS    LYSINE
                self.leu = 0            #LEU    LEUCINE
                self.met = 0            #MET    METHIONINE
                self.asn = 0            #ASN    ASPARAGINE
                self.pro = 0            #PRO    PROLINE
                self.gln = 0            #GLN    GLUTAMINE
                self.arg = 0            #ARG    ARGININE
                self.ser = 0            #SER    SERINE
                self.thr = 0            #THR    THREONINE
                self.val = 0            #VAL    VALINE
                self.trp = 0            #TRP    TRYPTOPHAN
                self.tyr = 0            #TYR    TYROSINE
                self.start = 'X'        # start amino acids
                self.end = 'X'          # end amino acids
##              self.slope = 0          # pk curve slope at the experimental pH
                self.step = 0.1         # ph step to take
                self.total_charge = 0.0 # list of total charge values for slope calculation
                self.z_list = [ ]       # estimated net charge vs ph
                self.zkda_list= [ ]     # estimated net charge/kDa vs pH
                self.z = 0.0        # the protein's charge at the crystallization experimental pH
                self.z_kda = 0.0    # estimated net charge/kDa at the reported pH of crystallization

        def line_parser(self, pdb_record):
                #this function will break up the lline into parts needed for the exp ph slope calculation
                # all values needed to send to pk calculations
```

```
        sequence = re.compile(r'(?P<pdb_id>[a-zA-Z0-9\.]+)\t'
                              r'(?P<mw>[0-9.]+)\t'
                              r'(?P<pi>[0-9.]+)\t'
                              r'(?P<exp_ph>[0-9.]+)\t'
                              r'(?P<ala>[0-9]+)\t'
                              r'(?P<cys>[0-9]+)\t'
                              r'(?P<asp>[0-9]+)\t'
                              r'(?P<glu>[0-9]+)\t'
                              r'(?P<phe>[0-9]+)\t'
                              r'(?P<gly>[0-9]+)\t'
                              r'(?P<his>[0-9]+)\t'
                              r'(?P<ile>[0-9]+)\t'
                              r'(?P<lys>[0-9]+)\t'
                              r'(?P<leu>[0-9]+)\t'
                              r'(?P<met>[0-9]+)\t'
                              r'(?P<asn>[0-9]+)\t'
                              r'(?P<pro>[0-9]+)\t'
                              r'(?P<gln>[0-9]+)\t'
                              r'(?P<arg>[0-9]+)\t'
                              r'(?P<ser>[0-9]+)\t'
                              r'(?P<thr>[0-9]+)\t'
                              r'(?P<val>[0-9]+)\t'
                              r'(?P<trp>[0-9]+)\t'
                              r'(?P<tyr>[0-9]+)\t'
                              r'(?P<start>[A-Z]+)\t'
                              r'(?P<end>[A-Z]+)')
        result = sequence.search(pdb_record)        # search in every line
        self.id = result.group('pdb_id')            # PDB ID
        self.mw = float(result.group('mw'))         # molecular weight
        self.pi = float(result.group('pi'))         # estimated pI
        self.exp_ph = float(result.group('exp_ph')) # experimental ph
        self.ala = int(result.group('ala'))         # ALA  ALANINE
        self.cys = int(result.group('cys'))         #CYS  CYSTEINE
        self.asp = int(result.group('asp'))         #ASP  ASPARTIC ACID
        self.glu = int(result.group('glu'))         #GLU  GLUTAMIC ACID
        self.phe = int(result.group('phe'))         #PHE  PHENYLALANINE
        self.gly = int(result.group('gly'))         #GLY  GLYCINE
        self.his = int(result.group('his'))         #HIS  HISTIDINE
        self.ile = int(result.group('ile'))         #ILE  ISOLEUCINE
        self.lys = int(result.group('lys'))         #LYS  LYSINE
        self.leu = int(result.group('leu'))         #LEU  LEUCINE
        self.met = int(result.group('met'))         #MET  METHIONINE
        self.asn = int(result.group('asn'))         #ASN  ASPARAGINE
        self.pro = int(result.group('pro'))         #PRO  PROLINE
        self.gln = int(result.group('gln'))         #GLN  GLUTAMINE
        self.arg = int(result.group('arg'))         #ARG  ARGININE
        self.ser = int(result.group('ser'))         #SER  SERINE
        self.thr = int(result.group('thr'))         #THR  THREONINE
        self.val = int(result.group('val'))         #VAL  VALINE
        self.trp = int(result.group('trp'))         #TRP  TRYPTOPHAN
        self.tyr = int(result.group('tyr'))         #TYR  TYROSINE
        self.start = result.group('start')          # start list of first amino acids
        self.end = result.group('end')              # end list of last amino acids
##      self.start_ph = float(self.exp_ph) - 0.1    # lower pH value
##      self.end_ph = float(self.exp_ph) + 0.1      # higher pH value
```

```
    def x_calc(self, start_ph, end_ph):
        # this is my attempt to calculate theoretical pI
        # I use recursion to narrow down the pI (successive approximation)
        # all pk values taken from Harpers Biochemistry (e-book) through HSLS
        # these values compare very closely to Voet & Voet Biochemistry, 1990
        #seq_list = ['A, 'C' 'D', 'E',  'F',  'G',  'H',  'I',  'K',  'L',
        #                    'M', 'N',  'P',  'Q',  'R',  'S',  'T',  'V',  'W',  'Y']
        pk_nh3 =  [ 9.69, 10.28, 9.60, 9.67, 9.00, 9.60, 9.17, 9.68, 8.95, 9.60, \
                            9.21, 8.80, 10.96, 9.13, 9.04, 9.15, 9.62, 9.62, 9.39, 9.11]
        pk_coo3 = [2.34, 1.96, 1.88, 2.19, 1.83, 2.34, 1.82, 2.36, 2.18, 2.36, \
                            2.28, 2.02, 1.99, 2.17, 2.17, 2.21, 2.11, 2.32, 2.38, 2.20]
        pk_side = [8.18, 3.65, 4.25, 10.07, 6.0, 10.53, 12.48]
        # pk_side C, D, E, Y, H, K, R
##      low_ph = 0
        ph = start_ph
        while ph <= end_ph:
            c_parts01 = 10**(ph - pk_side[0])                   # cysteine
            c_zi = c_parts01 /(c_parts01 + 1)
            c_nzi = c_zi * self.cys
            d_parts01 = 10**(ph - pk_side[1])                   # aspartic acid
            d_zi  = d_parts01 /(d_parts01 + 1)
            d_nzi = d_zi * self.asp
            e_parts01 = 10**(ph - pk_side[2])                   # glutamic acid
            e_zi  = e_parts01 /(e_parts01 + 1)
            e_nzi = e_zi * self.glu
            y_parts01 = 10**(ph - pk_side[3])                   # tyrosine
            y_zi  = y_parts01 /(y_parts01 + 1)
            y_nzi = y_zi * self.tyr
            neg_charge_sum = c_nzi + d_nzi + e_nzi + y_nzi      # sum negative side chain aa
            h_parts01 = 10**(ph - pk_side[4])                   # histidine
            h_zi = 1 /(h_parts01 + 1)
            h_nzi = h_zi* self.his
            k_parts01 = 10**(ph - pk_side[5])                   # lysine
            k_zi = 1 /(k_parts01 + 1)
            k_nzi = k_zi* self.lys
            r_parts01 = 10**(ph - pk_side[6])                   # arginine
            r_zi = 1 /(r_parts01 + 1)
            r_nzi = r_zi* self.arg
            pos_charge_sum = h_nzi + k_nzi + r_nzi               # sum positive side chain aa
            aa_list = ['A', 'C', 'D', 'E', 'F', 'G', 'H', 'I', 'K', 'L', 'M', 'N', 'P', 'Q', 'R', 'S', 'T', 'V', 'W', 'Y']
            start_zi = 0
            end_zi = 0
            count = 0
            for aa in aa_list:
                for start_aa in self.start:
                    if start_aa == aa:              # start amino acid - NH3 end
                        start_parts01 = 10**(ph - pk_nh3[count])
                        start_zi = start_zi + 1 /(start_parts01 + 1)
                count = count +1
            count = 0
            for aa in aa_list:
                for end_aa in self.end:
                    if end_aa == aa:                # end amino acid - Carboxyl end
                        end_parts01 = 10**(ph - pk_coo3[count])
                        end_zi = end_zi + end_parts01 /(end_parts01 + 1)
                count = count + 1
```

```
                        # calculate charges (net, sum of squares, and total charges)
                        self.total_charge = (start_zi + pos_charge_sum - neg_charge_sum - end_zi)
                        if (round(ph,1) == round(self.exp_ph, 1)):
                                self.z = self.total_charge
                                self.z_kda = (self.z / self.mw)
                        # append charge values to respective lists
                        self.z_list.append(str(self.total_charge))      # add total (net) charge to list
                        z_kda = (self.total_charge/self.mw)
                        self.zkda_list.append(str(z_kda))
                        ph = ph + self.step                             # increase pH to next step

        def write_data(self):
                # this is my function to write the data to file
                list = [self.id, '\t', self.mw, '\t', self.pi, '\t',self.exp_ph, '\t', \
                                self.ala, '\t', self.cys, '\t', self.asp, '\t', \
                                self.glu, '\t', self.phe, '\t', self.gly, '\t', self.his, '\t', \
                                self.ile, '\t', self.lys, '\t', self.leu, '\t', self.met, '\t', \
                                self.asn, '\t', self.pro, '\t', self.gln, '\t', self.arg, '\t', \
                                self.ser, '\t', self.thr, '\t', self.val, '\t', self.trp, '\t', \
                                self.tyr, '\t', self.start, '\t', self.end, '\t',\
                                self.z, '\t', self.z_kda, '\t']
                for value in list:
                        output.writelines(str(value))
##              for value in self.z_list:               # write estimsted titration curve (z vs. pH)
##                      output.writelines(value)
##                      output.writelines('\t')
                for value in self.zkda_list:            # write estimsted titration curve (z_kda vs. pH)
                        output.writelines(value)
                        output.writelines('\t')
                output.writelines('\n')
###############################################################################
########################  Main Program ########################################
###############################################################################
os.chdir(r'C:\Documents and Settings\dougalld\My Documents\School Material\PDB')
# change directory to where data file found
# this is a list of all PDB files with a valid experimental pH and xrayed with amino acid sequence information
outputFile = os.path.join("C:\\", "Documents and Settings", "dougalld", "My Documents", "School Material",
"PDB", "pdb_110705_charge_calc_results.txt")    # set output file path
##os.chdir(r'D:\Bioinformatics\PDB Files')                      # change directory to where data file found
### this is a list of all PDB files with a valid experimental pH and xrayed with amino acid sequence information
##outputFile = os.path.join("d:\\", "Bioinformatics", "PDB Files", "pdb_040505_charge_at_ph.txt")
# set output file path
output = open(outputFile, 'w+')                         # open output file for fasta file parse
caption = "PDB_ID\tmw\tpI\texp_ph\tA\tC\tD\tE\tF\tG\tH\tI\tK\tL\tM\tN\t"\
                "P\tQ\tR\tS\tT\tV\tW\tY\tstart_aa\tend_aa\tz\tz_kda\t"          # make caption a string for writing
output.writelines(caption)              # column headings printed to file
a = 1.0                                 # initialize counter
z_list = [ ]                            # this will contain column headings
# net charge header
while a <= 14:                          # create title line for titration curve
        z_list.append(('z_' + str(a)))
        z_list.append('\t')
        a = a + 0.1
z_list.append('\n')
for line in z_list:                     # this method or below both work
        output.write(line)              # column headings printed to file
```

```
a = 1.0                                  # re-initialize counter
zkda_list = [ ]                          # this will containt column headings
# net charge/kDa header
while a <= 14:                           # create title line for surface charge density curve
        zkda_list.append(('z_' + str(a)))
        zkda_list.append('\t')
        a = a + 0.1
#line_list.append('net_low\tnet_pi\tnet_high\tph_low\tph_high\t')
zkda_list.append('\n')
for line in zkda_list:                                  # this method or below both work
        output.write(line)                              # column headings printed to file
file_name = open('pdb_110705_input_charge_calc.txt')        # file with full list of valid pdb ids
start_ph =  1.0             # initial pH for titration curve calculation
end_ph   = 14.0             # final pH for titration curve calculation
for line in file_name:
        line = line.rstrip()
        curPDB = PDB()                                  # initialize a new instance of class PDB
        curPDB.line_parser(line)                        # sends current record for line parsing
        if curPDB.id != 'X':
                curPDB.x_calc(start_ph, end_ph)   # sends current object for pk calculation
                curPDB.write_data()               # sends current record for writing to file
output.close()
if not line:
        file_name.close()
```

nrpdb_testset.py

The purpose of this program is to get a list of nrPDB ids with pH values

```
# downloaded nrPDB and filtered out non-acceptable, non-x-ray,
# with exp. pH, no membrane proteins, resolution <= 3.0, and no nucleic/hetero containing structures
# sorted by blast group 80 and then rank blast 80
import re, string, os
class PDB:                  # create a class PDB that represents one PDB chain
   def __init__(self):      # initialize values required values:
      self.id = 'aaaa'          # PDB ID
      self.mmdb = 0             # MMDB ID
      self.group_80 = 0          # Group 80 Membership
      self.rank_80 = 0          # Group 80 Membership Rank
      self.exp_ph = 0.0          # Experimental pH

   def line_parser(self, line):
      # this function will parse the line into a list of variables
      # the line should look like pdb_id..group_e80..rank_e80..exp_ph
      title = re.compile(r'(?P<pdb_id>[a-zA-Z0-9]+)\t'\
                  r'(?P<mmdb>[0-9]+)\t'\
                   r'(?P<group_80>[0-9]+)\t'\
                   r'(?P<rank_80>[0-9]+)\t'\
                   r'(?P<exp_ph>[0-9.]+)\n')
      result = title.search(line)                   # search in line
      self.id           = result.group('pdb_id')
      self.mmdb         = result.group('mmdb')
      self.group_80      = int(result.group('group_80'))
      self.rank_80       = int(result.group('rank_80'))
      self.exp_ph        = float(result.group('exp_ph'))

   def write_data(self):
```

```
        # this is my function to write the data to file
        list = [self.id, '\t', self.mmdb, '\t', self.group_80, '\t', self.rank_80, '\t', self.exp_ph, '\n']
        for value in list:
           output.writelines(str(value))

##def train_parse(line):
##    # this function will parse the line into a list of variables
##    # the line should look like group_80\n
##    title = re.compile(r'(?P<group_80>[0-9]+)\n')
##    result = title.search(line)                                    # search in line
##    return(result.group('group_80'))
#############################################################################
############################  Main Program ##################################
#############################################################################
os.chdir(r'C:\Documents and Settings\dougalld\My Documents\School Material\PDB')
# change directory to where nrpdb_with_ph.txt file found
# set input file
inputFile = os.path.join("c:\\", "Documents and Settings", "dougalld", "My Documents", "School Material", "PDB",
"new_testset_nov_2005.txt")
# sorted by group and then rank
# this is a list of all nrPDB files that have a pH value (acceptable, xray method, no membranes, aa length >20)
outputFile = os.path.join("c:\\", "Documents and Settings", "dougalld", "My Documents", "School Material",
"PDB", "nrpdb_111405_testset.txt")    # set output file path
output = open(outputFile, 'w+')              # open output file for fasta file parse
test_list = [ ]                  # list of new entries for test set
line_list = [ ]                   # this will contain column headings
line_list.append('pdb_id\tmmdb_id\tgroup_80\trank_80\texp_ph\n')    # column headings
output.writelines(line_list)                        # write headings to ouput file
input_file = open(inputFile)                        # file with list of nrPDB entries ordered by group and rank
curPDB  = PDB()                                     # current line pdb entry
prevPDB = PDB()                                     # previous pdb entry with unique group_ni and highest ranking
##uniquePDB = [ ]                                   # list of unique PDB ids
##mmdb_group_80 = [ ]                               # list of MMDB Group 80 members
##mmdb_id = [ ]                                     # list of MMDB IDs
for line in input_file:            # read each line one at a time
   curPDB.line_parser(line)       # parse line into PDB entry
   if (curPDB.group_80 == prevPDB.group_80):
      # see if same group number - if same group number and lower rank continue
      curPDB.id = prevPDB.id
      curPDB.mmdb = prevPDB.mmdb
      curPDB.group_80 = prevPDB.group_80
      curPDB.rank_80 = prevPDB.rank_80
      curPDB.exp_ph = prevPDB.exp_ph
   else:              # new group number blast 10e-80 replace previous PDB entry with current PDB entry
      if prevPDB.id not in test_list:
         test_list.append(prevPDB.id)
         test_list.append('\n')
      prevPDB.write_data()
   prevPDB.id = curPDB.id
   prevPDB.mmdb = curPDB.mmdb
   prevPDB.group_80 = curPDB.group_80
   prevPDB.rank_80 = curPDB.rank_80
   prevPDB.exp_ph = curPDB.exp_ph
print test_list
input_file.close()             # close input  file
output.close()                # close output file
```

APPENDIX B: Commercial Crystallization Screens

COMMERCIAL SCREENS

Table B.1 lists the commercial screens examined from four commercial companies (Hampton Research, Emerald BioStructures, Jena Biosciences, and Molecular Dimensions). Table B.2 lists the pH values of the buffers listed in the screens.

Table B.1 Commercial screens examined for the reported pH of the buffer solutions.

Vendor	Screen	N
Hampton Research	Crystal Screen I	50
	Crystal Screen II	48
	Index	96
	Cryo	50
	Lite	50
MD*	Structure Screen 1 & 2	96
	Heavy & Light	96
	PACT	96
	Clear Strategy 1	96
	Clear Strategy 2	96
Jena Bioscience	JBScreen Classic Kits 1-10	240
Emerald BioStructures	Wizard Screens 1 & 2	96
Total		1110

* MD = Molecular Dimensions Inc.

Table B.2 The reported buffer pH values for the commercial screens listed in Table B.1.

pH	Frequency	Percent
3.5	3	0.3
4.0	4	0.4
4.2	6	0.5
4.5	14	1.3
4.6	65	5.8
5.0	10	0.9
5.5	65	5.8
5.6	34	3.1
6.0	18	1.6
6.2	5	0.4
6.5	192	17.3
7.0	43	3.9
7.5	224	20.1
8.0	35	3.1
8.5	203	18.3
9.0	15	1.3
9.5	10	0.9
10.5	4	0.4
Not Listed	162	14.6
Total	1112	100.0

When the pH_{cryst} values were obtained from the PDB were compared to the reported buffer pH values listed in the 18 different screens, an overlap was observed (Figure B.1), which can most likely explain this observed pattern. Positive values indicate areas where the commercial screens may be over sampling. Negative values indicate possible areas where commercial screens are under sampling. It should be noted that ~15% of commercial screens do not list a pH value.

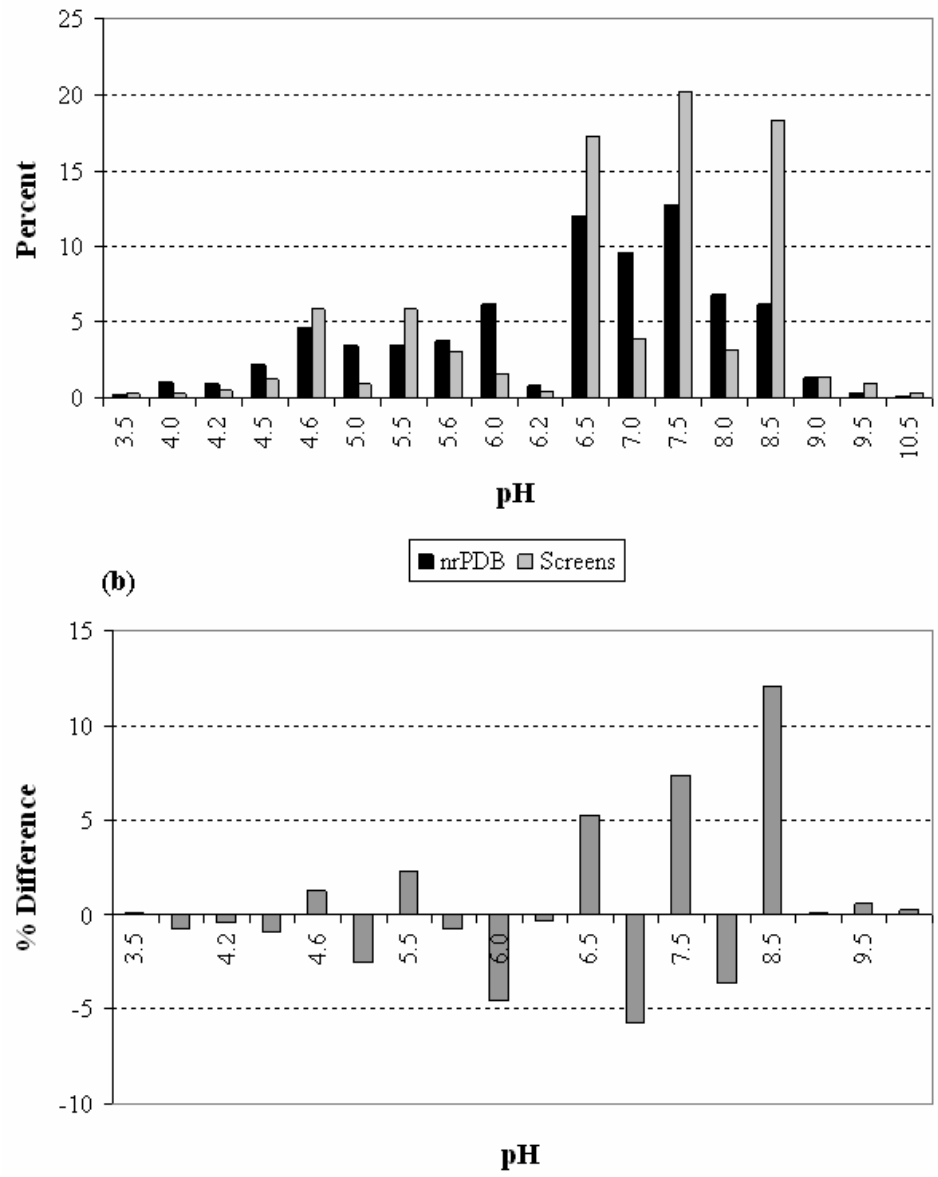


Figure B.1 (a) An overlap between the pH_{cryst} distribution and the buffer pH as reported from four commercial protein crystallization companies (Emerald BioStructures, Hampton Research, Jena Biosciences, and Molecular Dimensions). (b) The difference between the (% pH of screens) - (% pH of training set).

APPENDIX C: Annotated Example

ANNOTATED EXAMPLE

In Chapter 2, the methods used for determining the three-dimensional (3D) structure of proteins and nucleic acids were discussed. Current approaches, while rarely successful, need to be augmented to increase the throughput of crystallized structures. This chapter describes an annotated example using three proteins from the independent test set to demonstrate the usefulness of the *Features* (pI_{est}) to predict the *Observables* ($\overline{Q}_{cryst}$) examined in this dissertation, which can then be used to suggest more probable pH ranges that are more likely to result in crystallization.

C.1 INTRODUCTION

One hypothesis of this dissertation was that the Q-related *Observables*, particularly $\overline{Q}_{cryst}$, can be used as a proxy variables for the pH_{cryst}. This offers researchers the possibility to intelligently select pH ranges for initial crystallization screens. These methods in theory should increase the success rate of generating crystals suitable for diffraction studies by guiding the researcher to search areas in the crystallization search space that have a higher probability of success.

For any protein undergoing crystallization attempts, the amino acid sequence should be known apriori, although the protein's function may remain unknown. Thus, from the input of a

protein's amino acid sequence(s) (Figure C.1) an estimated titration curve (estimated net charge plotted as a function of solution pH) can be easily calculated (Figure C.2a). By accounting for the molecular weight or estimated solvent accessible surface area, the estimated titration curve can be transformed into a $\overline{Q}$ or σ curve (Figure C.2b). From these curves, prior *Observable* ($\overline{Q}_{cryst}$) probabilities derived from previously crystallized proteins in the PDB can be used to calculate the probability of crystallization for a test protein across the $\overline{Q}$ or σ curve, $P(\overline{Q} = \overline{Q}_{cryst} \mid PDB)$ or $P(\sigma = \sigma_{cryst} \mid PDB)$ (Figure C.3). These probabilities can then be translated back into pH search space to estimate the probability that $pH = pH_{cryst}$ given that $\overline{Q} = \overline{Q}_{cryst}$, $P(pH = pH_{cryst} \mid \overline{Q} = \overline{Q}_{cryst}, PDB)$, by matching the pH values that result in the given $\overline{Q}$ value along the $\overline{Q}$ curve. The same kind of probabilities can be calculated for $\sigma = \sigma_{cryst}$, $P(pH = pH_{cryst} \mid \sigma = \sigma_{cryst}, PDB)$. The researcher is then able to rationalize the selection of solution pH values that have a higher probability in generating crystals. Several methods are discussed in this dissertation on how to accomplish this, two of which are shown here. The example proteins are selected from 'Acidic' proteins in the test set, $pI_{est} = \text{'}Acidic\text{'}$ (Section 6.1.1.2).

```
>1LRH_A mol:protein length:163  auxin-binding protein 1
SCVRDNSLVRDISQMPQSSYGIEGLSHITVAGALNHGMKEVEVWLQTISPGQRTPIHRHSCEEVFTVLKGKGTLLM
GSSSLKYPGQPQEIPFFQNTTFSIPVNDPHQVWNSDEHEDLQVLVIISRPPAKIFLYDDWSMPHTAAVLKFPFVWD
EDCFEAAKEQL
>1LRH_B mol:protein length:163  auxin-binding protein 1
SCVRDNSLVRDISQMPQSSYGIEGLSHITVAGALNHGMKEVEVWLQTISPGQRTPIHRHSCEEVFTVLKGKGTLLM
GSSSLKYPGQPQEIPFFQNTTFSIPVNDPHQVWNSDEHEDLQVLVIISRPPAKIFLYDDWSMPHTAAVLKFPFVWD
EDCFEAAKEQL
>1LRH_C mol:protein length:163  auxin-binding protein 1
SCVRDNSLVRDISQMPQSSYGIEGLSHITVAGALNHGMKEVEVWLQTISPGQRTPIHRHSCEEVFTVLKGKGTLLM
GSSSLKYPGQPQEIPFFQNTTFSIPVNDPHQVWNSDEHEDLQVLVIISRPPAKIFLYDDWSMPHTAAVLKFPFVWD
EDCFEAAKEQL
>1LRH_D mol:protein length:163  auxin-binding protein 1
SCVRDNSLVRDISQMPQSSYGIEGLSHITVAGALNHGMKEVEVWLQTISPGQRTPIHRHSCEEVFTVLKGKGTLLM
GSSSLKYPGQPQEIPFFQNTTFSIPVNDPHQVWNSDEHEDLQVLVIISRPPAKIFLYDDWSMPHTAAVLKFPFVWD
EDCFEAAKEQL
>1IMV_A mol:protein length:398  PIGMENT EPITHELIUM-DERIVED FACTOR
MAHHHHHHMASLTVPAHVPSAAEDCEQLRSAFKGWGTNEKLIISILAHRTAAQRKLIRQTYAETFGEDLLKELDRE
LTHDFEKLVLVWTLDPSERDAHLAKEATKRWTKSNFVLVELACTRSPKELVLAREAYHARYKKSLEEDVAYHTTGD
HRKLLVPLVSSYRYGGEEVDLRLAKAESKILHEKISDKAYSDDEVIRILATRSKAQLNATLNHYKDEHGEDILKQL
EDGDEFVALLRATIKGLVYPEHYFVEVLRDAINRRGTEEDHLTRVIATRAEVDLKIIADEYQKRDSIPLGRAIAKD
TRGDYESMLLALLGQEED
>1AVB_A mol:protein length:226     Arcelin-1
SNDASFNVETFNKTNLILQGDATVSSEGHLLLTNVKGNEEDSMGRAFYSAPIQINDRTIDNLASFSTNFTFRINAK
NIENSAYGLAFALVPVGSRPKLKGRYLGLFNTTNYDRDAHTVAVVFDTVSNRIEIDVNSIRPIATESCNFGHNNGE
KAEVRITYDSPKNDLRVSLLYPSSEEKCHVSATVPLEKEVEDWVSVGFSATSGSKKETTETHNVLSWSFSSNFI
>1AVB_B mol:protein length:226     Arcelin-1
SNDASFNVETFNKTNLILQGDATVSSEGHLLLTNVKGNEEDSMGRAFYSAPIQINDRTIDNLASFSTNFTFRINAK
NIENSAYGLAFALVPVGSRPKLKGRYLGLFNTTNYDRDAHTVAVVFDTVSNRIEIDVNSIRPIATESCNFGHNNGE
KAEVRITYDSPKNDLRVSLLYPSSEEKCHVSATVPLEKEVEDWVSVGFSATSGSKKETTETHNVLSWSFSSNFI
```

Figure C.1 Amino acid sequences of the three test set proteins in FASTA format.

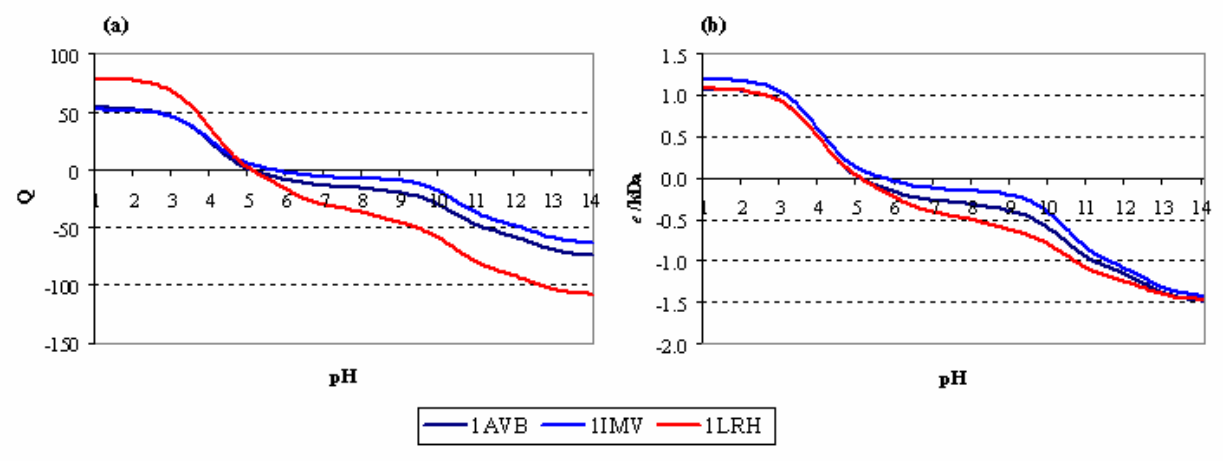


Figure C.2 The (a) estimated titration curves and (b) $\overline{Q}$ curves for the three proteins from the PDB test set, 1AVB, 1IMV, and 1LRH.

C.2 EXAMPLE PROTEINS FROM THE TEST SET

A typical example demonstrating three possible scenarios using three pI_{est} ='*Acidic*' proteins (PDB IDs: 1LRH, 1IMV, and 1AVB) from the independent test set is discussed here, where 'Acidic' proteins have a 5.0 < pI_{est} ≤ 6.0. These three proteins have 1-4 protein chains in the asymmetric unit, a molecular weight of 44-74 kDa, and a $diff_{lim}$ between 1.9-2.9 Å (Table C.1). One of the goals of this research was to predict the solution pH ranges to search for growing crystals given the test protein's sequence as input. Because proteins denature at the extreme pH ranges and few proteins were found to crystallize at these extreme pH ranges, probabilities were calculated for the pH range from 3.0-11.0 in 0.5 unit increments. It should be noted that the pH ranges < 4.0 (3.0 and 3.5) and > 10.0 (10.5 and 11.0) routinely have probabilities near zero. This was accomplished by calculating the $\overline{Q}$ or σ curve for the test protein and then using the probability density curve for previously crystallized proteins. Examples are shown for (1) using the prior probabilities for all proteins in the training set, (2) a subgroup of proteins based upon the test protein's pI_{est}, and using the pH_{cryst} priors based on the pH_{cryst} distribution of all

proteins in the training set. For reference, a 'Random' probability was also calculated, which assumes that each of the fourteen pH values between pH 4.0 and 10.0 (in 0.5 unit increments) examined has an equal chance of forming a crystal, i.e. 1/14 (7.7%). A 10% threshold is used to represent areas of high probability and to compare the pH ranges selected by the different methods.

Table C.1 The three example proteins used for the annotated example.

ID	# Chains in AU	MW (kDa)	pI_{est}	pH_{cryst}	pI_{est} Group	$diff_{lim}$
1LRH	4	73.6	5.1	5.5	Acidic	1.9
1IMV	1	44.3	5.7	6.2	Acidic	2.9
1AVB	2	49.9	5.1	4.8	Acidic	1.9

C.2.1 1LRH

The first scenario, represented by 1LRH (Auxin-binding protein 1), is an example where the predicted pH_{cryst} probability range was extremely accurate and could have significantly reduced the pH space to search. Both methods using the $\overline{Q}$ curve predicted a maximum pH probability between a pH of 5.5 and 6.0, which correctly captured the pH_{cryst} for 1LRH, 5.5 (Figure C.3a). Using a 10% probability threshold suggests a pH range of 5.0-6.5 (All proteins) or 5.5-6.5 ($pI_{est} = 'Acidic'$). However, using the pH_{cryst} priors based on all proteins misses the listed pH_{cryst} value, while suggesting a pH range of 6.5-7.5 for all target proteins. Assuming all pH values with a probability above the 10% threshold should be searched, the priors based on 'All' proteins, $P(pH = pH_{cryst} \mid \overline{Q} = \overline{Q}_{cryst}, nrPDB_{10.04.05})$, or the priors for $pI_{est} = 'Acidic'$ proteins, $P(pH = pH_{cryst} \mid \overline{Q} = \overline{Q}_{cryst}, pI_{est} = 'Acidic', nrPDB_{10.04.05})$, would reduce the initial screen conditions by 53%, only searching the solution conditions with a pH of 5.6 and 6.5 (Table C.2b). However, the pH probability distribution, based on $pI_{est} = 'Acidic'$ proteins gives a tighter pH range with a high probability to search, 5.5-6.0, reducing. When such narrow distributions are predicted, a focused pH search (5.5-6.0) to concentrate the initial crystallization attempts would

be suggested. This would allow the researcher greater flexibility in searching other parameters such as the salt or precipitating agent, while using the same number of experiments.

C.2.2 1IMV

The second scenario, represented by 1IMV (Pigment epithelium-derived factor), is an example where there is a relatively long stretch of pH values that have a similar probability (~10%). The probability distribution based on 'All' proteins suggests a pH range of 6.5-7.5 should be searched (Figure C.3b), which all predict a probability near the 10% threshold. The pH_{cryst} for 1IMV was 6.2, which falls near the peak stretch of this probability distribution. By only choosing the pH values above 10%, the number of experiments can be reduced by 44%. By using the priors for $pI_{est} = 'Acidic'$ proteins the reduction in experiments decreases only by 23%, removing all pH = 4.6-5.6 experiments. However, there is now a large drop-off in probability of crystals between pH 6.0 and 6.5, while an increase in probability is observed at more basic pH values, 8.5-9.0. The $diff_{lim}$ for 1IMV was 2.9 Å (Table C.1). It would be interesting to see if using a more neutral-basic pH solution could improve the resolution of the crystals. pH_{cryst}.

C.2.3 1AVB

The final example scenario was for 1AVB (Arcelin-1), which proved difficult, because this acidic protein (pI_{est} 5.1) crystallized below its pI_{est}, (pH_{cryst} = 4.8). Using prior probabilities from all proteins in the training set indicated an acidic range of pH values to search, 5.0-6.5. This pH range does not capture the pH_{cryst}, but just misses. However, this method eliminates 53% of the experiments, the experimental conditions with a pH value above 6.5 solutions. When the priors are used for $pI_{est} = 'Acidic'$ proteins, a broader pH range, 5.5-8.0, is suggested. However, the pH range from 5.0-6.5 still exhibits a much higher probability than the 7.0-8.0 range. Using the prior for acidic proteins reduces the overall experiment by 30%. Now, the pH_{cryst} is not captured by the prediction, but a very large drop-off in prediction is observed between pH 5.0 and 5.5. The training data used for generating the probabilities had few 'Acidic' structures

(~17%) that crystallized below their pI_{est}, which results in low probabilities for those pH ranges. However, it remains unknown whether the crystal structure of 1AVB ($diff_{lim}$ 1.9 Å) could be improved, i.e. lower resolution, at more neutral pH values.

C.3 SUMMARY

This procedure of using a protein's biophysical properties (*Features*), such as the protein's pI_{est}, to predict a $\overline{Q}_{cryst}$ range (*Hidden Observable*) and thus a $pH = pH_{cryst}$ range (*Controllable*) appeared successful if the assumption holds true that the Q-related *Observables* ($\overline{Q}_{cryst}$) can be used as proxy variables for the pH_{cryst}.. These methods use a protein's $\overline{Q}$ curve to generate probabilities for the pH_{cryst} based upon the $\overline{Q}_{cryst}$ distribution of previously crystallized proteins. For some proteins, the pH search space for initial crystallization attempts is narrowed down to a pH unit or less. This was the case with the protein 1LRH. Conversely, there may be little to no reduction in the pH search space. However, using the pH ranges above the 10% threshold for the initial screens should increase the chances of generating a well-ordered protein crystal. It was not known whether there would be an increased chance of obtaining a 'hit' due to lack of available information.

Reducing the time spent in searching for crystallization conditions in areas that are not likely to produce crystals should increase the chances of yielding a high quality crystal. Without any prior knowledge, initial screens should attempt crystallization at a wide pH range of 4.0-9.0 every 0.5 pH units. Proteins have an increased tendency to denature outside of this pH range. Initial screens should be done with a sparse matrix screen varying all other ingredients, such as the salt and precipitant types and concentrations. With no prior information, this would increase the initial conditions searched. However, prior knowledge of a protein's $\overline{Q}$ distribution (estimated specific charge curve) as a function of solution pH may allow the researcher to remove many of the solution pH values examined. The prior knowledge could be encoded as Bayesian priors and used to generate probability distributions for various solution parameters. These probabilities can be further combined over multiple features as more data is collected and

analyzed to suggest regions in the crystallization search space more likely to produce well-ordered crystals.

Table C.2 (a) The distribution of pH values (buffers) searched by Crystal Screens 1 and 2 (Hampton Research; Aliso Viejo, CA). (b) Based on the pH_{cryst} range prediction, certain screen wells can be removed.

(a)

pH	Frequency	Percent
4.6	12	12.2
5.6	10	10.2
6.5	18	18.4
7.0	1	1.0
7.5	21	21.4
8.5	15	15.3
9.0	3	3.1
N/A*	18	18.4
Total	98	100.0
% of Total Experiments		

N/A = buffer pH not listed

(b)

Priors: All Proteins			Priors: Acidic Proteins		
1LRH	1IMV	1AVB	1LRH	1IMV	1AVB
-	-	-	-	-	-
10	-	10	10	-	10
18	18	18	18	18	18
-	1	-	-	1	1
-	21	-	-	21	21
-	-	-	-	15	-
-	-	-	-	3	-
18	18	18	18	18	18
46	58	46	46	76	68
46.9	56.1	46.9	46.9	77.6	69.4

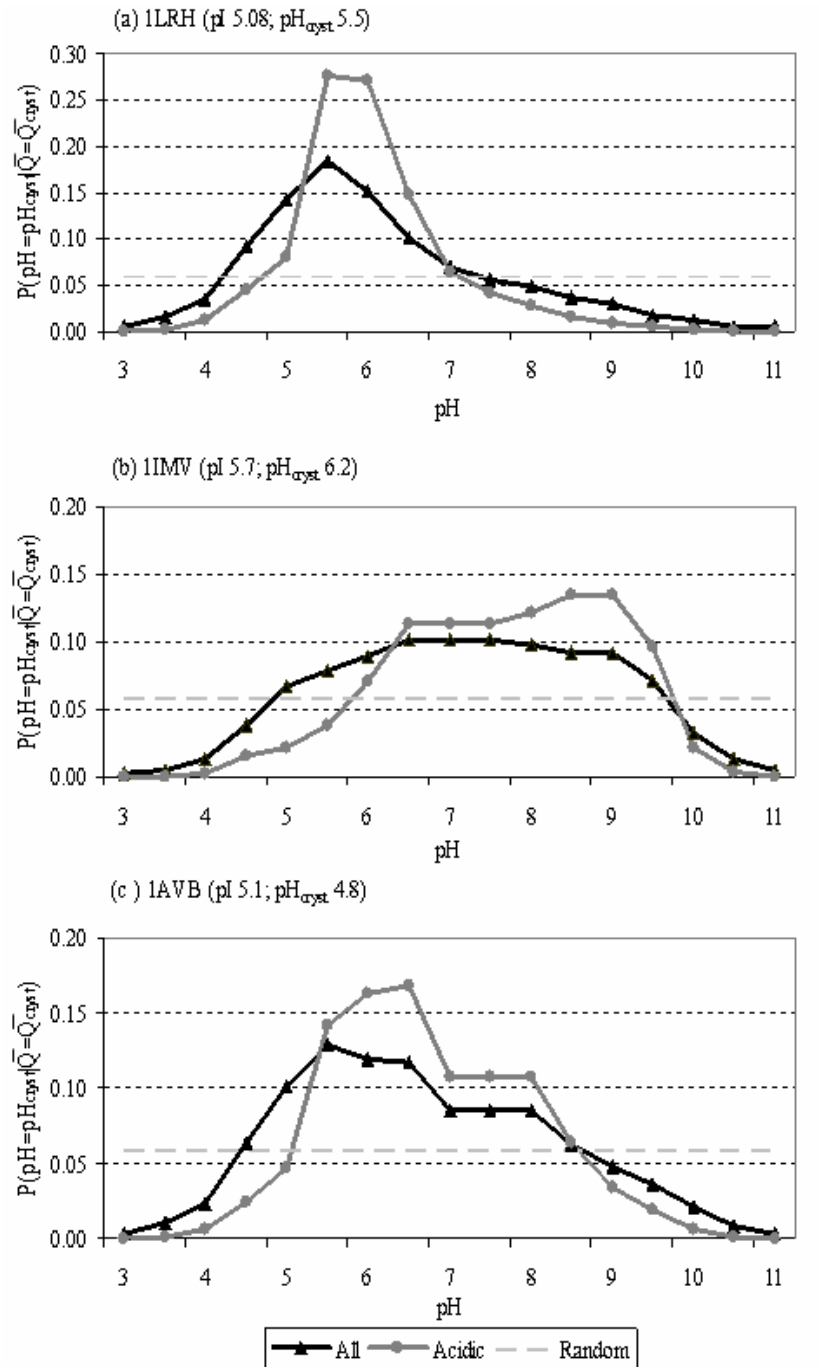


Figure C.3 An example application using three 'Acidic' test set proteins to predict the $P(pH = pH_{cryst} \mid \overline{Q} = \overline{Q}_{cryst}, nrPDB_{10.04.05})$, $P(pH = pH_{cryst} \mid \overline{Q} = \overline{Q}_{cryst}, pI_{est} = 'Acidic', nrPDB_{10.04.05})$, and $P(pH = pH_{cryst} \mid nrPDB_{10.04.05})$.

APPENDIX D: Curve Fitting

PRELIMINARY RESULTS OF CURVE FITTING

D.1 INTRODUCTION

Similar to Chapter 6, this section is based on the assumption (hypothesis) that the Q-related *Observables*, particularly $\overline{Q}_{cryst}$, can be used as a proxy variables for the pH_{cryst}. The preliminary results of using a cubic equation (Equation D.1) to fit the plot of the pI_{est} by $\overline{Q}_{cryst}$ (Figure D.1) is presented in this section. A high correlation (Spearman's rho) was observed between these two variables, 0.746, in Chapter 5 (Table 5.1). In order to give an estimate of the error, five-fold cross-validation was used. Five models (equations) were obtained by fitting a curve on four groups (4,288 proteins) and then testing the model on the remaining group of proteins (1,072 proteins). The resulting five equations were used to give an estimate of the error and averaged to obtain a final equation for estimating the $\overline{Q}_{cryst}$. The Charge Range Test (Section 4.8.3) was then used to evaluate the model fit and compare the results to the other methods discussed previously.

Equation D.1 $\overline{Q}_{cryst} = a * pI + b * pI^2 + c * pI^3 + d$

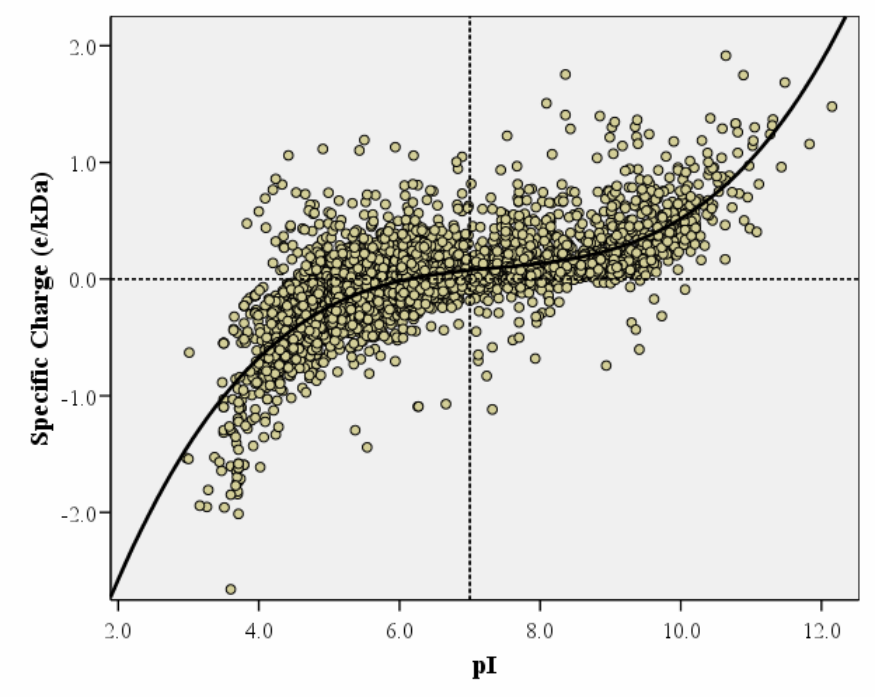


Figure D.1 Scatterplot of the pI_{est} versus the $\overline{Q}_{cryst}$ along with a cubic fit.

D.2 RESULTS

First, using five-fold cross-validation, five models were created using the random groupings to predict the $\overline{Q}_{cryst}$ of proteins not used in the creation of the model. The constants of the five cubic fit models are shown in Table D.1, while an example cubic fit using all of the data is shown in Figure D.1. In each model, one of the five random groups was left out and used to test the resulting model. The models were created by fitting the plot of the pI_{est} and $\overline{Q}_{cryst}$ using a cubic equation, which resulted in a mean r^2 of 0.547±0.007. The resulting model using the average of the five coefficients from Equation D.1 is shown in Equation D.2. Attempts were then made to use of this knowledge by using a target protein's pI_{est} to predict the most probable $\overline{Q}$ ranges for crystallization.

Equation D.2. $\overline{Q}_{cryst} = 2.455 * pI_{est} - 0.324 * (pI_{est})^2 + 0.015 * (pI_{est})^3 - 6.230$

Table D.1 Cubic fit variables for the fit of Equation D.1.

Fold*	a	b	c	d	r^2
Not 5	2.48	-0.33	0.015	-6.31	0.547
Not 4	2.45	-0.32	0.015	-6.21	0.552
Not 3	2.51	-0.33	0.015	-6.36	0.548
Not 2	2.45	-0.32	0.015	-6.22	0.550
Not 1	2.38	-0.31	0.014	-6.05	0.535
Mean	2.46	-0.32	0.015	-6.23	0.547
SD	0.05	0.01	<0.001	0.12	0.006

* Translates into modeling all groups, but the Group listed. For example, Not 5 translates into modeling Groups 1-4 and testing on Group 5.

Using the Charge Range Test, the effectiveness of the five models and their error are reported in Table D.2. This table shows that the cubic equation is correct in predicting the actual $\overline{Q}_{cryst}$ approximately 20% of the time. While 20% may not seem very accurate, it should noted that the success rate of obtaining a crystal suitable for diffraction studies has been listed as approximately 1-23% (Section 1.2). Because proteins will often crystallize over a range of pH values, the area around the prediction was also examined. When the $\overline{Q}$ range was expanded to ±0.1, 59% of the proteins $\overline{Q}_{cryst}$ are correctly predicted. When the range is expanded to the predicted $\overline{Q}_{cryst}$ value ±0.2, 78% of the test set protein's $\overline{Q}_{cryst}$ values fell within this range. An increase to 86% is observed when the range was expanded to ±0.3 *e*/kDa. The results from all five models and the validation group behaved similarly.

These results were then compared to those obtained earlier in the dissertation (Table D.3). The cubic fit models performed similarly to the $2Step_6$ and SOM_{14x2} results on the training set at all levels (Table D.3a), but slightly better (2-4%) on the test set at the Mean±0.1 level (Table D.3b). However, the other methods were not examined using five fold cross validation, which would allow for a more fair comparison between methods. Statistical measures could then be performed to determine whether the observed differences were statistically significant.

Table D.2 Charge Range Test results for the (a) modelled or (b) test set proteins of the five cubic fit models.

(a) Modelled

$\overline{Q}_{cryst}$	Not 5	Not 4	Not 3	Not 2	Not 1	Mean	SD
Predicted	18.9	19.6	19.4	19.2	20.1	19.4	0.4
Pred. ±0.1	58.5	57.9	58.0	58.4	60.0	58.6	0.9
Pred. ±0.2	77.6	77.4	77.1	78.1	78.2	77.7	0.5
Pred. ±0.3	86.3	86.5	86.1	86.1	86.4	86.3	0.2
Model r^2	0.547	0.552	0.548	0.550	0.535	0.547	0.006

(b) Test Set

$\overline{Q}_{cryst}$	Group 5	Group 4	Group 3	Group 2	Group 1	Mean	SD
Predicted	19.1	17.9	18.2	20.6	19.6	19.1	1.1
Pred. ±0.1	57.6	59.7	59.0	60.0	56.3	58.5	1.5
Pred. ±0.2	77.7	78.3	79.1	76.7	76.5	77.6	1.1
Pred. ±0.3	86.4	85.5	87.1	86.3	85.5	86.2	0.7

D.3 DISCUSSION

Again it should be mentioned that these results are based on the assumption that the $\overline{Q}_{cryst}$ variable can be used as a proxy variable for the pH_{cryst}. Although these results are preliminary, the cubic fit seemed to result in predictions that were as good as the clustering methods in Chapter 6 (2Step and SOMs). However, this method is much easier to calculate. Based on the pI_{est} value, which can be calculated based on the AA sequence prior to any crystallization attempts, Equation D.2 can be used to estimate initial values for $\overline{Q} = \overline{Q}_{cryst}$. Similar to the previous methods, the estimated titration curve can then be used to more intelligently select (or

remove) initial pH values that have a higher probability in resulting in crystallization (or remove those pH values with a lower probability).

Table D.3 Comparing the Cubic Fit results to the previous methods using the Charge Range Test.

(a) Training

$\overline{Q}_{cryst}$	pI_{est} Bins	Random Clusters	$2Step_6$	SOM_{14x2}	**Cubic Fit**
Mean ±0.1	54.0	40.4±2.4	57.2	59.6	**58.6±0.9**
Mean ±0.2	74.6	60.1±1.8	77.7	78.9	**77.7±0.5**
Mean ±0.3	84.8	73.1±1.5	87.4	87.7	**86.3±0.2**

(b) Test

$\overline{Q}_{cryst}$	pI_{est} Bins	Random Clusters	$2Step_6$	SOM_{14x2}	**Cubic Fit**
Mean ±0.1	49.0	37.5	54.4	56.0	**58.5±1.5**
Mean ±0.2	73.0	59.6	76.7	77.0	**77.6±1.1**
Mean ±0.3	83.1	71.8	87.6	87.7	**86.2±0.7**

APPENDIX E: Linear Regression and Neural Networks

PRELIMINARY RESULTS OF USING LINEAR REGRESSION AND NEURAL NETWORKS

E.1 INTRODUCTION

Similar to Appendix D, this section is also based on the assumption that the Q-related variables can be used as proxy variables for the pH_{cryst}. The preliminary results of using Linear Regression (LR) and Neural Networks (NN) to predict $\overline{Q}_{cryst}$ ranges based on the amino acid (AA) composition is presented in this section. Because the pI_{est} is based upon the charged AA composition (Arginine, Lysine, Histidine, Glutamic acid, Aspartic acid, and Tyrosine), a new hypothesis was tested: the composition of all twenty amino acids can predict the $\overline{Q}_{cryst}$ better than the pI_{est} alone (Appendix D). While LR assumes a linear relationship between variables, NN are able to handle non-linear problems.

Equation E.1 $Y = c + b_0 X_0 + b_1 X_1 + b_2 X_2$, where c= constant

Equation E.2 $\overline{Q}_{cryst} = c + b_A A_{comp} + b_C C_{comp} + \ldots + b_Y Y_{comp}$, where A_{comp} = Alanine composition, C_{comp} = Cysteine composition, etc.

Using SPSS 14.0, stepwise linear regression, where the input was the amino acid composition (percentage) of all 20 amino acids, was used to determine if the amino acid composition could give insight into the $\overline{Q}_{cryst}$.

Similar to modeling the pI_{est} and $\overline{Q}_{cryst}$ relationship with a cubic fit in Appendix D, these two methods discussed in this section predict a single $\overline{Q}_{cryst}$ value as an outcome. Because crystallization is thought to occur over a range of pH values (i.e. not a single pH value) the Charge Range Test was used to compare LR and NN to the previous methods. In order to give an estimate of the error, five-fold cross-validation was used. Therefore, five models (equations) were obtained by each method, modeling on four groups (4,288 proteins) and then testing the model on the remaining group of proteins (1,072 proteins). The resulting five equations were used to give an estimate of the error for estimating the $\overline{Q}_{cryst}$. The Charge Range Test (Section 4.8.3) was then used to evaluate the model fit and compare the results to the other methods discussed previously.

E.2 RESULTS

First, the zero-order correlations between the amino acid compositions and the estimated net charge variables were examined (Table E.1). Again, the random number was used as a control. Little correlation was observed between any AA and the pH_{cryst}, with the largest correlation being Arginine at 0.074. Much larger correlations were observed between the amino acids and the Q_{cryst}. To no surprise, the charged amino acids (Arginine, Aspartic acid, Glutamic acid, and Lysine) had a much higher correlation to the Q_{cryst}. However, both Histidine and Tyrosine had little correlation to the Q_{cryst}, <0.02. It is also interesting to note that Alanine had a relatively high correlation to the Q_{cryst}, -0.108, compared to the other amino acids. Only Aspartic acid, Glutamic acid, and Lysine had higher correlations with the Q_{cryst}.

In order to account for size differences among proteins, the $\overline{Q}_{cryst}$ was examined. The Spearman's correlations between the amino acid composition and the $\overline{Q}_{cryst}$ are shown in Table E.1. There wasn't much difference in the AA correlations with Q_{cryst} and $\overline{Q}_{cryst}$. While a few correlations were slightly reduced (greater than a 0.02 reduction), such as Alanine and

Methionine, only Arginine, Glutamic acid, and Aspartic acid increased their correlations by over 0.04. Again, no AA correlations with the Random Number approached statistical significance.

Table E.1 Spearman's rho correlations of amino acid composition and the Q -related *Observables*.

	pH_{cryst}	Q_{cryst}	$\overline{Q}_{cryst}$	Random Number
% Alanine	-0.002	**-0.108**	-0.088	0.006
% Arginine	0.074	**0.105**	**0.157**	-0.009
% Asparagine	-0.033	0.053	0.059	-0.006
% Aspartic Acid	0.021	**-0.308**	**-0.352**	0.004
% Cysteine	-0.016	0.019	0.007	-0.007
% Glutamic Acid	0.025	**-0.207**	**-0.252**	0.001
% Glutamine	0.019	0.005	0.007	0.002
% Glycine	-0.039	0.018	0.028	0.010
% Histidine	0.042	-0.016	0.001	0.003
% Isoleucine	0.021	0.004	0.021	-0.010
% Leucine	0.044	-0.042	-0.025	0.004
% Lysine	0.018	**0.241**	**0.254**	0.017
% Methionine	0.019	-0.043	-0.004	0.012
% Phenylalanine	0.034	-0.063	-0.070	-0.003
% Proline	-0.007	-0.003	0.012	-0.020
% Serine	-0.013	0.070	0.068	0.007
% Threonine	-0.043	-0.001	0.006	0.004
% Tryptophan	-0.012	0.007	0.018	-0.008
% Tyrosine	0.017	-0.010	-0.020	0.001
% Valine	-0.046	0.002	0.008	-0.011

After examining the correlations between the amino acid composition and the crystallization *Observables* (pH_{cryst} , Q_{cryst} , and $\overline{Q}_{cryst}$), attempts were made to predict $\overline{Q}_{cryst}$ using the amino acid composition. Similar to Appendix D, five-fold cross-validation was used to give an estimate of the error for the models. Five random groups of proteins were created with the model being developed on four, while being tested on the remaining group not used in the model creation. This results in five models for each method.

The LR equations of the five models are shown in Table E.2. In each model, one of the five random groups was left out and used to test the resulting model. The regression models were created by using the amino acid composition to predict the $\overline{Q}_{cryst}$. The resulting model using the average of the five coefficients from Equation E.1 is shown in Equation E.2. Attempts were then made to use of this knowledge by using a target protein's AA composition to predict the most probable $\overline{Q}$ ranges for crystallization.

Table E.2 The LR equations.

Amino Acid	LR1	LR2	LR3	LR4	LR5	Average
Cysteine	-0.962	-0.773	-0.478	-0.649	-0.753	-0.723
Aspartic Acid	-8.445	-8.456	-8.605	-8.668	-8.513	-8.537
Glutamic Acid	-7.723	-7.764	-7.940	-8.011	-7.964	-7.880
Glycine	0.000	0.401	0.000	0.000	0.000	0.080
Histidine	2.107	2.222	2.468	2.365	2.652	2.363
Isoleucine	-0.325	0.000	0.000	0.000	0.000	-0.065
Lysine	8.385	8.615	8.508	8.440	8.601	8.510
Leucine	-0.535	0.000	0.000	-0.334	0.000	-0.174
Asparagine	0.000	0.408	0.000	0.000	0.000	0.082
Arginine	8.076	8.229	8.052	7.949	8.038	8.069
Serine	0.000	0.000	0.313	0.000	0.443	0.151
Valine	0.427	0.683	0.747	0.612	0.504	0.595
Tryptophan	0.000	0.000	0.745	0.000	0.000	0.149
Tyrosine	-0.591	0.000	-0.538	0.000	-0.455	-0.317
Constant*	0.074	-0.100	0.000	0.000	0.000	-0.005

*Constants with a 0.000 did not reach statistical significance ($p<0.05$)

Using the Charge Range Test, the effectiveness of the five models and their error are reported in Table E.3. This table shows that the LR model is correct in predicting the actual $\overline{Q}_{cryst}$ approximately 21% of the time. While 21% may not seem very accurate, it should again be noted that the success rate of obtaining a crystal suitable for diffraction studies has been listed as approximately 1-23% (Section 1.2). Because proteins will often crystallize over a range of pH

values, the area around the prediction was also examined. When the $\overline{Q}$ range was expanded to ±0.1, 65% of the proteins $\overline{Q}_{cryst}$ are correctly predicted. When the range is expanded to the predicted $\overline{Q}_{cryst}$ value ±0.2, 82% of the test set protein's $\overline{Q}_{cryst}$ values fell within this range. An increase to 90% is observed when the range was expanded to ±0.3 *e*/kDa. The results from all five models and the validation group behaved similarly. Similar values were obtained using NN (Table E.4).

Table E.3 Charge Range Test results for the (a) modelled or (b) test set proteins of the five LR models.

(a) Modelled

$\overline{Q}_{cryst}$	Not 5	Not 4	Not 3	Not 2	Not 1	Mean	SD
Predicted	21.1	20.9	21.4	20.8	21.7	21.2	0.3
Pred. ±0.1	64.9	65.1	64.9	65.3	65.6	65.1	0.3
Pred. ±0.2	82.2	82.0	81.7	82.5	82.5	82.2	0.3
Pred. ±0.3	90.1	89.8	89.6	90.0	90.0	89.9	0.2
Model r^2	0.630	0.640	0.645	0.643	0.641	0.640	0.006

(b) Test Set

$\overline{Q}_{cryst}$	Group 5	Group 4	Group 3	Group 2	Group 1	Mean	SD
Predicted	21.1	20.6	21.1	21.9	19.8	20.9	0.8
Pred. ±0.1	65.4	64.5	65.3	67.0	61.8	64.8	1.9
Pred. ±0.2	82.4	81.9	83.9	81.2	80.3	81.9	1.3
Pred. ±0.3	89.1	89.8	91.3	89.9	88.6	89.8	1.0

These results were then compared to those obtained earlier in the dissertation (Table E.5). The LR and NN models performed 5-7% better than the previous results at the Mean±0.1 and Mean±0.2 levels (Table E.5a), but slightly better (2-4%) on the test set at the Mean±0.3 level (Table E.5b). However, most the other methods, including the $2Step_6$ and SOM_{14x2} results, were not examined using five fold cross validation, which would allow for statistical measures to be performed to determine the level of significance. However, the NN and LR were statistically

better (p-value <0.01) than the Cubic fit models as judged by a Student's t-test, which had similar values to those of unsupervised clustering ($2Step_6$ and SOM_{14x2}). However, the LR and NN results were not statistically different from each other at all levels. Therefore, due to the ease of use and interpretation, it would be recommended to use the LR models over the NN models.

Both of these methods also give the "importance" of the variables (amino acid composition). To no surprise, the charged amino acids were important in all models (10 in all). It should also be noted that the AA composition would be known for any target protein apriori to any crystallization attempts, while it's $\overline{Q}_{cryst}$ is unknown.

Table E.4 Charge Range Test results for the (a) modelled or (b) test set proteins of the five NN models.

(a) Modelled

$\overline{Q}_{cryst}$	Not 5	Not 4	Not 3	Not 2	Not 1	Mean	SD
Predicted	21.2	21.2	22.4	19.3	20.8	21.0	1.1
Pred. ±0.1	65.9	66.1	65.1	65.6	65.1	65.6	0.4
Pred. ±0.2	82.9	82.6	81.6	81.8	81.9	82.2	0.5
Pred. ±0.3	90.0	89.8	89.4	89.7	90.1	89.8	0.3
Model r^2	0.644	0.639	0.643	0.640	0.639	0.641	0.002

(b) Test Set

$\overline{Q}_{cryst}$	Group 5	Group 4	Group 3	Group 2	Group 1	Mean	SD
Predicted	20.7	21.7	22.0	20.0	20.9	21.1	0.8
Pred. ±0.1	62.2	66.3	65.0	65.2	65.0	64.8	1.5
Pred. ±0.2	81.1	81.1	83.8	81.9	81.6	81.9	1.1
Pred. ±0.3	89.1	89.0	91.6	89.3	89.4	89.7	1.1

Table E.5 Comparing the LR and NN results to the previous methods using the Charge Range Test.

(a) Training

$\overline{Q}_{cryst}$	Random Clusters	$2Step_6$	SOM_{14x2}	$\overline{Q}_{cryst}$ Predicted	Cubic Fit	LR	NN
Mean ±0.1	40.4±2.4	57.2	59.6	±0.1	58.6±0.9	65.1±0.3	65.6±0.4
Mean ±0.2	60.1±1.8	77.7	78.9	±0.2	77.7±0.5	82.2±0.3	82.2±0.5
Mean ±0.3	73.1±1.5	87.4	87.7	±0.3	86.3±0.2	89.9±0.2	89.8±0.3

(b) Test

$\overline{Q}_{cryst}$	Random Clusters	$2Step_6$	SOM_{14x2}	$\overline{Q}_{cryst}$ Predicted	Cubic Fit	LR	NN
Mean ±0.1	37.5	54.4	56.0	±0.1	58.5±1.5	64.8±1.9	64.8±1.5
Mean ±0.2	59.6	76.7	77.0	±0.2	77.6±1.1	81.9±1.3	81.9±1.1
Mean ±0.3	71.8	87.6	87.7	±0.3	86.2±0.7	89.8±1.0	89.7±1.1

E.3 DISCUSSION

Again it should be mentioned that these results are based on the hypothesis (interpretation) that the $\overline{Q}_{cryst}$ variable can be used as s proxy variable for the . Although these results are preliminary, the LR and NN models resulted in better predictions than all previous methods. An advantage of these models, is that the models allow for some interpretation, i.e. what amino acids are important for predicting the $\overline{Q}_{cryst}$.

Table E.6: The frequency of the amino acids in the LR and NN models.

Abbreviation	Amino Acid	LR Models (n = 5)	NN Models (n = 5)	# Times in Model
A (Ala)	Alanine	0	1	1
C (Cys)	**Cysteine**	**5**	**5**	**10**
D (Asp)	**Aspartic Acid**	**5**	**5**	**10**
E (Glu)	**Glutamic Acid**	**5**	**5**	**10**
F (Phe)	Phenylalanine	0	1	1
G (Gly)	Glycine	1	3	4
H (His)	**Histidine**	**5**	**5**	**10**
I (Ile)	Isoleucine	1	0	1
K (Lys)	**Lysine**	**5**	**5**	**10**
L (Leu)	Leucine	2	2	4
M (Met)	Methionine	0	0	0
N (Asn)	Asparagine	1	3	4
P (Pro)	Proline	0	3	3
Q (Gln)	Glutamine	0	0	0
R (Arg)	**Arginine**	**5**	**5**	**10**
S (Ser)	Serine	2	3	5
T (Thr)	Threonine	0	2	2
V (Val)	Valine	5	3	8
W (Trp)	Tryptophan	1	1	2
Y (Tyr)	Tyrosine	3	4	7

* Translates into modeling all groups, but the Group listed. For example, Not 5 translates into modeling Groups 1-4 and testing on Group 5.

BIBLIOGRAPHY

Akaike, H. (1983) Information measures and model selection. *Bull. Int. Stat. Inst.* **50**, 277-290.

Albeck, S., Burstein, Y., Dym, O., Jacobovitch, Y., Levi, N., Meged, R., Michael, Y., Peleg, Y., Prilusky, J., Schreiber, G., Silman, I., Unger, T., and Sussman, J.L. (2005) Three-dimensional structure determination of proteins related to human health in their functional context at The Israel Structural Proteomics Center (ISPC). This paper was presented at ICCBM10, *Acta Cryst.* D**61**, 1364-1372.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.* **25**, 3389-3402.

Andrade, M.A., Casari, G., Sander, C. and Valencia, A. (1997) Classification of protein families and detection of the determinant residues with an improved self-organizing map, *Biol. Cybern.* **76**, 441-450.

Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N., and Yeh, L.S. (2004) UniProt: the Universal Protein knowledgebase, *Nucleic. Acids Res.* **32**, D115-D119.

Asherie, N. (2004) Protein crystallization and phase diagrams, *Methods* **34**, 266-272.

Bahadur, R.P., Chakrabarti, P., Rodier, F., and Janin, J. (2003) Dissecting subunit interfaces in homodimeric proteins, *Proteins* **53**, 708-719.

Barlow, D.J. and Thornton, J.M. (1986) The distribution of charged groups in proteins, *Biopolymers* **25**, 1717-1733.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshell, M., Moxon, S., Sonnhammer, E.L.L., Studholme, D.J., Yeats, C., and Eddy, S. R. (2004) The Pfam protein families database, *Nucleic Acids Res.* **32**, D138-D141.

Bendtsen, J.D., Nielsen, H., von Heijne, G., and Brunak, S. (2004) Improved prediction of signal peptides: SignalP 3.0, *J. Mol. Biol.* **340**, 783-795.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000) The Protein Data Bank, *Nucleic Acids Res.* **28**, 235-242.

Bertone, P., Kluger, Y., Lan, N., Zheng, D., Christendat, D., Yee, A., Edwards, A., Arrowsmith, C.H., Montelione, G.T., and Gerstein, M. (2001) SPINE: an integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics, *Nucleic Acids. Res.* **29**, 2884-2898.

Bourne, P.E., Berman, H.M., McMahon, B., Watenpaugh, K.D., Westbrook, J.D., and Fitzgerald, P.M.D. (1997) Macromolecular crystallographic information file, *Methods Enzymol.* **277**, 571-590.

Bourne, P.E., Allerton, C.K J., Krebs, W., Li, W., Shindyalov, I.N., Godzik, A., Friedberg, I., Liu, T., Wild, D., Hwang, S., Ghahramani, Z., Chen, L., and Westbrook, J. (2004) The status of structural genomics defined through the analysis of current targets and structures, *Pac. Symp. Biocomput*. **9**, 375-386.

Braun, P, Hu, Y., Shen, B., Halleck, A., Koundinya, M., Harlow, E, and LaBaer, J. (2002) Proteome-scale purification of human proteins from bacteria, *Proc. Natl. Acad. Sci. USA* **99**, 2654-2659.

Brzozowski, A.M. and Walton, J. (2001) Clear strategy screens for macromolecular crystallization, *J. Appl. Cryst.* **34**, 97-101.

Bukrinsky, J.T. and Poulsen, J.C.N. (2001) pH, conductivity and long-term stability in the Crystal Screen solutions, *J. Appl. Cryst.* **34**, 533-534.

Burley, S.K. and Bonanno, J.B. (2002) Structuring the universe of proteins, *Annu. Rev. Genomics Hum. Genet.* **3**, 243-262.

Bussow, K., Quedenau, C., Sievert, V., Tischer, J., Scheich, C., Seitz, H., Hieke, B., Niesen, F.H., Gotz, F., Harttig, U., and Lehrach, H. (2004) A catalog of human cDNA expression clones and its application to structural genomics, *Genome Biol.* **5**, R71.

Cai, Y.–D, Li, Y.–X, and Chou, K.C. (2000) Using neural networks for prediction of domain structural classes, *Biochim. Biophys. Acta* **1476**, 1-2.

Canaves, J.M., Page, R., Wilson, I.A., and Stevens, R.C. (2004) Protein biophysical properties that correlate with crystallization success in Thermotoga maritima: maximum clustering strategy for structural genomics, *J. Mol. Biol.* **344**, 977-991.

Carter, C.W. Jr. and Carter, C.W. (1979) Protein crystallization using incomplete factorial experiments, *J. Biol Chem.* **254**, 12219-12223.

Carter, C.W. Jr., Baldwin, E.T., and Frick, L. (1988) Statistical design of experiments for protein crystal growth and the use of a precrystalline assay, *J. Cryst. Growth* **90**, 60-73.

Carter, C.W. Jr. (1999) *Crystallization of Nucleic Acids and Proteins: A Practical Approach*, edited by A. Ducruix and R. Giege, pp. 75-120. Oxford University Press, New York, NY.

Carugo, O. and Argos, P. (1997) Protein-protein crystal-packing contacts, *Protein Sci.* **6**, 2261-2263.

Chapman, B. and Chang, J. (2000) Biopython: Python tools for computational biology, *ACM SIGBIO Newslett.* **20**, 15-19.

Chayen, N.E. (1999) Recent advances in methodology for the crystallization of biological macromolecules, *J. Cryst. Growth* **198/199**, 649-655.

Chayen, N.E. (2002) Tackling the bottleneck of protein crystallization in the post-genomic era, *Trends Biotechnol.* **20**, 98.

Chayen, N.E. and Saridakis, E. (2002) Protein crystallization for genomics: towards high-throughput optimization techniques, *Acta Cryst.* D**58**, 921-927.

Chiu, T., Fang, D., Chen, J., Wang, Y., and Jeris, C. (2001) A robust and scalable clustering algorithm for mixed type attributes in large database environment. *Proc. 7th ACM SIGKDD International Conf. Knowledge Data Mining 2001*, 263-268.

Choi, D. –K., Ito, T., Tsukahara, F., Hirai, M., and Sakaki, Y. (1999) Developmentally-regulated expression of mNapor encoding an apoptosis-induced ELAV-type RNA binding protein, *Gene* **237**, 135-142.

Chothia, C. (1975) Structural invariants in protein folding, *Nature* **254**, 304-308.

Christendat, D., Yee, A., Dharamsi, A., Kluger, Y., Savchenko, A., Cort, J.R.,Booth, V., Mackereth, C.D., Saridakis, V., Ekiel, I., Kozlov, G., Maxwell, K.L., Wu, N., McIntosh, L.P., Gehring, K., Kennedy, M.A., Davidson, A.R., Pai, E.F., Gerstein, M., Edwards, A.M., and Arrowsmith, C.H. (2000) Structural proteomics of an archaeon, *Nature Struct. Biol.* **7**, 903-909.

Collins, B.K., Tomanicek, S.J., Lyamicheva, N., Kaiser, M.W., and Mueser, T.C. (2004) A preliminary solubility screen used to improve crystallization trials: crystallization and preliminary X-ray structure determination of Aeropyrum pernix flap endonuclease-1, *Acta Cryst.* D**60**, 1674-1678.

Congreve, M., Murray, C.W., and Blundell, T.L. (2005) Structural biology and drug discovery, *Drug Disc. Today* **10**, 895-907.

Conover, W.J. (1999). *Practical Nonparametric Statistics.* 3rd edition. John Wiley & Sons, New York, NY.

Couzin, J. (2005) Ten centers chosen to decode protein structures, *Science* **309**, 230.

Cox, M.J. and Weber, P.C. (1988) An investigation of protein crystallization parameters using successive automated grid searches (SAGS), *J. Crystal Growth* **90**, 318-324.

Cudney, B., Patel, S., Weisgraber, K., Newhouse, Y., and McPherson, A. (1994) Screening and Optimization Strategies for Macromolecular Crystal-Growth, *Acta Cryst.* D**50**, 414-423.

Dasgupta, S., Iyer, G.H., Bryant, S.H., Lawrence, C.E., and Bell, J.A. (1997) Extent and nature of contacts between protein molecules in crystal lattices and between subunits of protein oligomers, *Proteins* **28**, 494-514.

decode Genetics, Emerald BioStructures. Sturlugata 8 · IS-101 Reykjavik, Iceland.

DeLucas, L.J., Bray, T.L., Nagy, L., McCombs, D., Chernov, N., Hamrick, D., Cosenza, L., Belgovskiy, A., Stoops, B., and Chait, A. (2003) Efficient protein crystallization, *J. Struct. Biol.* **142**, 188-206.

Demoruelle, K., Guo, B., Kao, S., McDonald, H., Nikic, D., Holman, S., and Wilson, W. (2002) Correlation between the osmotic second virial coefficient and solubility for equine serum albumin and ovalbumin *Acta Cryst.* D**58**, 1544-1548.

Dock-Bregeon, A.-C., Moras, D. and Giege, R. (1999) *Crystallization of Nucleic Acids and Proteins: A Practical Approach*, edited by A. Ducruix and R. Giege, pp. 209-243. Oxford University Press Inc, New York, NY.

Doudna, J.A., Grosshans, C., Gooding, A., and Kundrot, C.E. (1993) Crystallization of ribozymes and small RNA motifs by a sparse matrix approach, *Proc. Natl. Acad. Sci. USA* **90**, 7829-7833.

Dyson, M.R., Shadbolt, P., Vincent, K.J., Perera, R.L., and McCafferty, J. (2004) Production of soluble mammalian proteins in Escherichia coli: identification of protein features that correlate with successful expression, *BMC Biotech.* **4**, 32.

Edwards, A.M., Arrowsmith, C.H., Christendat, D., Dharamsi, A., Friesen, J.D., Greenblatt, J.F., and Vedadi, M. (2000) Protein production: feeding the crystallographers and NMR spectroscopists, *Nat. Struct. Biol. Suppl.* **7**, 970-972.

Eisenstein, E., Gilliland, G.L., Herzberg, O., Moult, J., Orban, J., Poljak, R.J., Banerjei, L., Richardson, D., and Howard, A.J. (2000) Biological function made crystal clear - annotation of hypothetical proteins via structural genomics, *Curr. Opin. Biotech.* **11**, 25-30.

Farr, R.G., Perryman, A.L., and Samudzi, C.T. (1998) Re-clustering the database for crystallization of macromolecules, *J. Cryst. Growth* **183**, 653-668.

Ferran, E.A., Pflugfelder, B., and Ferrara, P. (1994) Self-organized neural maps of human protein sequences, *Prot. Sci.* **3**, 507-521.

Forsyth, W.R., Antosiewicz, J.M., and Robertson, A.D. (2002) Empirical relationships between protein structure and carboxyl pKa values in proteins, *Proteins* **48**, 388-403.

Fredericq, E. and Neurath, J. (1950) The interaction of insulin with thiocyanate and other anions. The minimum molecular weight of insulin, *J. Am. Chem. Soc.* **72**, 2684-2691.

Galkin, O. and Vekilov, P.G. (2001) Nucleation of protein crystals: critical nuclei, phase behavior, and control pathways, *J. Cryst. Growth* **232**, 63-76.

Gao, W., Li, S.–X., and Bi, R.–C. (2005) An attempt to increase the efficiency of protein crystal screening: a simplified screen and experiments, *Acta Cryst.* D**61**, 776-779.

George, A., Chiang, Y., Guo, B., Arabshahi, A., Cai, Z., and Wilson, W.W. (1997) Second virial coefficient as predictor in protein crystal growth, *Methods Enzymol.* **276**, 100-110.

George, A. and Wilson, W.W. (1994) Predicting protein crystallization from a dilute solution property, *Acta Cryst.* D**50**, 361-365.

Gill, R., Mohammed, F., Badyal, R., Coates, L., Erskine, P., Thompson, D., Cooper, J., Gore, M., and Wood, S. (2005) High-resolution structure of myo-inositol monophosphatase, the putative target of lithium therapy, *Acta Cryst.* D**61**, 545-555.

Gilliland, G.L. (1988) A biological macromolecule crystallization database: A basis for a crystallization strategy, *J. Cryst. Growth* **90**, 51-59.

Gilliland, G.L. (1997) Biological macromolecule crystallization database, *Methods Enzymol.* **277**, 546-556.

Gilliland, G.L., Tung, M., Blakeslee, D.M., and Ladner, J.E. (1994) Biological Macromolecule Crystallization Database, Version 3.0: new features, data and the NASA archive for protein crystal growth data, *Acta Cryst.* D**50**, 408-413.

Gilliland, G.L., Tung, M., and Ladner, J.E. (2002) The Biological Macromolecule Crystallization Database: crystallization procedures and strategies, *Acta Cryst.* D**58**, 916-920.

Goh, C.-S., Lan, N., Douglas, S.M., Wu, B., Echols, N., Smith, A., Milburn, D., Montelione, G.T., Zhao, H., and Gerstein, M. (2004) Mining the structural genomics pipeline: identification of protein properties that affect high-throughput experimental analysis, *J. Mol. Bio.* **336**, 115-130.

Golden, B.L., Podell, E.R., Gooding, A.R., and Cech, T.R. (1997) Crystals by design: a strategy for crystallization of a ribozyme derived from the Tetrahymena group I intron, *J. Mol. Biol.* **270**, 711-723.

Goulding, C.W. and Perry, L.J. (2003) Protein production in Escherichia coli for structural studies by X-ray crystallography, *J. Struct. Biol.* **142**, 133-143.

Green, A.A. (1931a) Studies in the physical chemistry of the proteins. VIII. The solubility of hemoglobin in concentrated salt solutions. A study of the salting out of proteins, *J. Biol. Chem.* **93**, 495-516.

Green, A.A. (1931b) Studies in the physical chemistry of proteins. IX. The effect of electrolytes on the solubility of hemoglobin in solutions of varying hydrogen ion activity with a note on the comparable behavior of casein. *J. Biol. Chem.* **93**, 517-542.

Gronwall, A. (1942) Studies on the solubility of lactoglobulin. I. The solubility of lactoglobulin in dilute solutions of sodium chloride at varying ionic strength and hydrogen ion activity, *Compt. Rend. Trav. Lab. Carlsberg* **24**, 8-20.

Guan, P., Hattotuwagama, C.K., Doytchinova, I.A., and Flower, D.R. (2006) MHCPred 2.0: an updated quantitative T-cell epitope prediction server, *Appl. Bioinformatics* **5**, 55-61.

Guo, B., Kao, S., McDonald, H., Asanov, A., Combs, L., and Wilson, W. (1999) Correlation of second virial coefficients and solubilities useful in protein crystal growth, *J. Cryst. Growth* **196**, 424-433.

Haas, C. and Drenth, J. (1999) Understanding protein crystallization on the basis of the phase diagram, *J. Cryst. Growth* **196**, 388-394.

Hampton Research. Aliso Viejo, CA 92656-3317.

Harris, L.J., Skaletsky, E., and McPherson, A. (1995) Crystallization of intact monoclonal antibodies, *Prot. Struct. Funct. Genet.* **23**, 285-289.

Haynes, C.A., Tamura, K., Korfer, H.R., Blanch, H.W., and Prausnitz, J.M. (1992) Thermodynamic properties of aqueous alpha-chymotrypsin solutions from membrane osmometry measurements, *J. Phys. Chem.* **96**, 905-912.

Heinemann, U., Frevert, J., Hofmann, K.–P., Illing, G., Maurer, C., Oschkinat, H., and Saenger, W. (2000) An integrated approach to structural genomics, *Prog. Biophys. Mol. Biol.* **73**, 347-362.

Hennessy, D., Buchanan, B., Subramanian, D., Wilkosz, P.A., and Rosenberg, J.M. (2000) Statistical methods for the objective design of screening procedures for macromolecular crystallization, *Acta Cryst.* D**56**, 817-827.

Hitscherich, C., Kaplan, J., Allaman, M., Wiencek, J., and Loll, P. (2000) Static light scattering studies of OmpF porin: implications for integral membrane protein crystallization, *Prot. Sci.* **9**, 1559-1566.

Holm, L. and Sander, C. (1998) Removing near-neighbour redundancy from large protein sequence collections, *Bioinformatics* **14**, 423-429.

Hsu, A.L., Tang, S.–L., and Halgamuge, S.K. (2003) An unsupervised hierarchical dynamic self-organizing approach to cancer class discovery and marker gene identification in microarray data, *Bioinformatics* **19**, 2131-2140.

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

Idicula-Thomas, S., Kulkarni, A.J., Kulkarni, B.D., Jayaraman, V.K., and Balaji, P.V. (2006) A support vector machine-based method for predicting the propensity of a protein to be soluble or to form inclusion body on overexpression in Escherichia coli, *Bioinformatics* **22**, 278-284.

Ikai, A. (1980) Thermostability and aliphatic index of globular proteins, *J. Biochem.* **88**, 1895-1898.

Iwata, S. (2003). *Methods and results in the crystallization of membrane proteins*, edited by S. Iwata, pp. 281-298. International University Line, La Jolla, CA.

Jancarik, J. and Kim, S.H. (1991) Sparse matrix sampling: a screening method for crystallization of proteins, *J. Appl. Cryst.* **24**, 409-411.

Janin, J. (1976) Surface area of globular proteins, *J. Mol. Biol.* **105**, 13-14.

Janin, J., Miller, S., and Chothia, C. (1988) Surface, subunit interfaces and interior of oligomeric proteins, *J. Mol. Biol.* **204**, 155-164.

Janin, J. and Rodier, F. (1995) Protein-protein interaction at crystal contacts, *Proteins* **23**, 580-587.

Jena Bioscience GmbH, Loebstedter Strasse 78, D-07749 Jena, Germany.

Jones, D.T. and Ward, J.J. (2003) Prediction of disordered regions in proteins from position specific score matrices, *Prot. Struct. Funct. Genet.* **53**, 573-5578

Judge, R.A., Johns, M.R., and White, E.T. (1996) Solubility of ovalbumin in ammonium sulfate solutions, *J. Chem. Eng. Data* **41**, 422-424.

Jurisica, I., Rogers, P., Glasgow, J.I., Fortier, S., Luft, J.R., Wolfley, J.R., Bianca, M.A., Weeks, D.R., and DeTitta, G.T. (2001) Intelligent decision support for protein crystal growth, *IBM Systems Journal* **40**, 394-409.

Kam, Z., Shore, H.B., and Feher, G. (1978) On the crystallization of proteins, *J. Mol. Biol.* **123**, 539-555.

Kantardjieff, K.A. and Rupp, B. (2004) Protein isoelectric point as a predictor for increased crystallization screening efficiency, *Bioinformatics* **20**, 2162-2168.

Kimber, M.S., Vallee, F., Houston, S., Necakov, A., Skarina, T., Evdokimova, E., Beasley, S., Christendat, D., Savchenko, A., Arrowsmith, C.H., Vedadi, M., Gerstein, M., and Edwards, A.M. (2003) Data mining crystallization databases: knowledge-based approaches to optimize protein crystal screens, *Prot. Struct. Funct. Genet.* **51**, 562-568.

Kingston, R.L., Baker, H.M., and Baker, E.N. (1994) Search designs for protein crystallization based on orthogonal arrays, *Acta Cryst.* D**50**, 429-440.

Kohonen, T. (2001). *Self-Organizing Maps.* 3rd Edition. Springer-Verlag, New York, NY.

Kyte, J. and Doolittle, R.F. (1982) A simple method for displaying the hydropathic character of a protein, *J. Mol. Biol.* **157**, 105-132.

Leavis, P.C. and Rothstein, F. (1974) The solubility of fibrinogen in dilute salt solutions, *Arch. Biochem. Biophys.* **161**, 671-682.

Lee, B. and Richards, F.M. (1971) The interpretation of protein structures: estimation of static accessibility, *J. Mol. Biol.* **55**, 379-400.

Lesley, S.A., Kuhn, P., Godzik, A., Deacon, A.M., Mathews, I., Kreusch, A., Spraggon, G., Klock, H.E., McMullan, D., Shin, T., Vincent, J., Robb, A., Brinen, L.S., Miller, M.D., McPhillips, T.M., Miller, M.A., Scheibe, D., Canaves, J.M., Guda, C., Jaroszewshi, L., Selby, T.L., Elsliger, M.–A., Wooley, J., Taylor, S.S., Hodgson, K.O., Wilson, I.A., Schultz, P.G., and Stevens, R.C. (2002) Structural genomics of the Thermotoga maritima proteome implemented in a high-throughput structure determination pipeline, *Proc. Natl. Acad. Sci. USA* **99**, 11664-11669.

Luan, C.–H., Qiu, S., Finley, J.B., Carson, M., Gray, R.J., Huang, W., Johnson, D., Tsao, J., Reboul, J., Vaglio, P., Hill, D.E., Vidal, M., DeLucas, L.J., and Luo, M. (2004) High-throughput expression of C. elegans proteins, *Genome Res.* **14**, 2102-2110.

Luft, J.R., Collins, R.J., Fehrman, N.A., Lauricella, A.M., Veatch, C.K., and DeTitta, G.T. (2003) A deliberate approach to screening for initial crystallization conditions of biological macromolecules, *J. Struct. Biol.* **142**, 170-179.

Mahony, S., Smith, T.J., McInerney, J.O., and Golden, A. (2004) Gene prediction using the Self-Organizing Map: automatic generation of multiple gene models, *BMC Bioinformatics* **5**, 23.

Majeed, S., Ofek, G., Belachew, A., Huang, C.–C., Zhou, T., and Kwong, P.D. (2003) Enhancing protein crystallization through precipitant synergy, *Structure* **11**, 1061-1070.

McPherson, A. (1995) Increasing the size of microcrystals by fine sampling of pH limits, *J. Appl. Cryst.* **28**, 362-365.

McPherson, A. (1999) *Crystallization of biological macromolecules*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

McPherson, A. (2004) Introduction to protein crystallization, *Methods* **34**, 254-265.

Miller, S., Janin, J., Lesk, A.M. and Chothia, C. (1987a) Interior and surface of monomeric proteins, *J. Mol. Biol.* **196**, 641-656.

Miller, S., Lesk, A.M., Janin, J., and Chothia, C. (1987b) The accessible surface area and stability of oligomeric proteins, *Nature* **328**, 834-836.

Molecular Dimensions Limited. Unit 4, Northfield Business Park, Northfield Road, Soham, Cambs., CB7 5UE, England.

Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures *J. Mol. Biol.* **247**, 536-540.

Nelson, D.L. and Cox, M.M. (2000). *Lehninger principles of biochemistry*, 3rd ed. Worth Publishers, New York, NY.

Noguchi, T. and Akiyama, Y. (2003) PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB) in 2003, *Nucleic Acids. Res.* **31**, 492-493.

Northrop, J.H. (1930) Crystalline pepsin: II. General properties and experimental methods, *J. Gen. Physiol.* **13**, 767-780.

Odahara, T., Ataka, M., and Katsura, T. (1994) Phase diagram determination to elucidate the crystal growth of the photoreaction center from Rhodobacter sphaeroides, *Acta Cryst.* D**50**, 639-642.

Oldfield, C.J., Ulrich, E.L., Cheng, Y., Dunker, A.K., and Markley, J.L. (2005) Addressing the intrinsic disorder bottleneck in structural proteomics, *Prot. Struct. Funct. Genet.* **59**, 444-453.

Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., and Thornton, J.M. (1997) CATH--a hierarchic classification of protein domain structures, *Structure* **5**, 1093-1108.

Page, R., Grzechnik, S.K., Canaves, J.M., Spraggon, G., Kreusch, A., Kuhn, P., Stevens, R.C., and Lesley, S.A. (2003) Shotgun crystallization strategy for structural genomics: an optimized two-tiered crystallization screen against the Thermotoga maritima proteome, *Acta Cryst.* D**59**, 1028-1037.

Page, R., Peti, W., Wilson, I.A., Stevens, R.C., and Wuthrich, K. (2005) NMR screening and crystal quality of bacterially expressed prokaryotic and eukaryotic proteins in a structural genomics pipeline, *Proc. Natl. Acad. Sci. USA* **102**, 1901-1905.

Page, R. and Stevens, R.C. (2004) Crystallization data mining in structural genomics: using positive and negative results to optimize protein crystallization screens, *Methods* **34**, 373-389.

Patrickios, C.S. and Yamasaki, E.N. (1995) Polypeptide amino acid composition and isoelectric point. II. Comparison between experiment and theory, *Anal. Biochem.* **231**, 82-91.

Peat, T.S., Christopher, J.A., and Newman, A. (2005) Tapping the Protein Data Bank for crystallization information, *Acta Cryst.* D**61**, 1662-1669.

Peng, K., Obradovic, Z., and Vucetic, S. (2004) Exploring bias in the Protein Data Bank using contrast classifiers, *Pac. Symp. Biocomput*. **9**, 435-446

Piazza, R. and Pierno, M. (2000) Protein interactions near crystallization: a microscopic approach to the Hofmeister series, *J. Phys.: Condens. Matter* **12**, A443-A449.

Radaev, S., Li, S., and Sun, P.D. (2006) A survey of protein-protein complex crystallizations, *Acta Cryst.* D**62**, 605-612.

Radaev, S. and Sun, P.D. (2002) Crystallization of protein-protein complexes, *J. Appl. Cryst.* **35**, 674-676.

Rapic-Otrin, V., McLenigan, M.P., Bisi, D.C., Gonzalez, M., and Levine, A.S. (2002) Sequential binding of UV DNA damage binding factor and degradation of the p48 subunit as early events after UV irradiation, *Nucleic Acids Res.* **30**, 2588-2598.

Retailleau, P., Ries-Kautt, M., and Ducruix, A. (1997) No salting-in of lysozyme chloride observed at low ionic strength over a large range of pH, *Biophys. J.*, **73**, 2156-2163.

Ries-Kautt, M. and Ducruix, A. (1997) Inferences drawn from physicochemical studies of crystallogenesis and precrystalline state, *Methods Enzymol.* **276**, 23-59.

Ries-Kautt, M. and Ducruix, A. (1999). *Crystallization of Nucleic Acids and Proteins: A Practical Approach*, edited by A. Ducruix and R. Giege, pp. 269-312. Oxford University Press Inc, New York, NY.

Rosenbaum, D.F. and Zukoski, C.F. (1996) Protein interactions and crystallization, *J. Cryst. Growth* **169**, 752-758.

Rupp, B. (2003) Maximum-likelihood crystallization, *J. Struct. Biol.* **142**, 162-169.

Rupp, B. and Wang, J. (2004) Predictive models for protein crystallization, *Methods* **34**, 390-407.

Saijo, S., Sato, T., Tanaka, N., Ichiyanagi, A., Sugano, Y., and Shoda, M. (2005) Precipitation diagram and optimization of crystallization conditions at low ionic strength for deglycosylated dye-decolorizing peroxidase from a basidiomycete, *Acta Cryst.* F**61**, 729-732.

Samudzi, C.T., Fivash, M.J., and Rosenberg, J.M. (1992) Cluster-analysis of the biological macromolecule crystallization database, *J. Cryst. Growth* **123**, 47-58.

Santesson, S., Cedergren-Zeppezauer, E.S., Johansson, T., Laurell, T., Nilsson, J., and Nilsson, S. (2003) Screening of nucleation conditions using levitated drops for protein crystallization, *Anal. Chem.* **75**, 1733-1740.

Saridakis, E. and Chayen, N.E. (2003) Systematic improvement of protein crystals by determining the supersolubility curves of phase diagrams, *Biophys. J.* **84**, 1218-1222.

Saridakis, E.E.G., Shaw Stewart, P.D., Lloyd, L.F., and Blow, D.M. (1994) Phase diagram and dilution experiments in the crystallization of carboxypeptidase G2, *Acta Cryst.* D**50**, 293-297.

Schwartz, R., Ting, C.S., and King, J. (2001) Whole proteome pI values correlate with subcellular localizations of proteins for organisms within the three domains of life, *Genome Res.* **11**, 703-709.

Schwarz, G. 1978. Estimating the dimension of a model. *Ann. Stat.* **6**, 461–64.

Scott, W.G., Finch, J.T., Grenfell, R., Fogg, J. Smith, T., Gait, M.J., and Klug, A. (1995) Rapid crystallization of chemically synthesized hammerhead RNAs using a double screening procedure, *J. Mol. Biol.* **250**, 327-332.

Sedzik, J. (1995) Regression analysis of factorially designed trials--a logical approach to protein crystallization, *Biochim. Biophys. Acta* **1251**, 177-185.

Segelke, B.W. (2001) Efficiency analysis of sampling protocols used in protein crystallization screening, *J. Cryst. Growth* **232**, 553-562.

Shaw, K.L., Grimsley, G.R., Yakovlev, G.I., Makarov, A.A., and Pace, C.N. (2001) The effect of net charge on the solubility, activity, and stability of ribonuclease Sa, *Protein Sci.* **10**, 1206-1215.

Shaw Stewart, P.D. and Khimasia, M. (1994) Predispensed gradient matrices - a new rapid method of finding crystallization conditions, *Acta Cryst.* D**50**, 441-442.

Shieh, H.–S., Stallinos, W.C., Stevens, A.M., and Stegman, R.A. (1995) Using sampling techniques in protein crystallization, *Acta Cryst.* D**51**, 305-310.

Shih, Y.C., Prausnitz, J.M., and Blanch, H.W. (1992) Some Characteristics of Protein Precipitation by Salts, *Biotech. Bioeng*. **40**, 1155-1164.

Shimura, H., Scholossmacher, M.G., Hattori, N., Frosch, M.P., Trockenbacher, A., Schneider, R., Mizuno, Y., Kosik, K.S., and Selkos, D.J. (2001) Ubiquitination of a new form of alpha-synuclein by parkin from human brain: Implications for Parkinson's disease, *Science* **293**, 263-269.

Sommer, M.O.A. and Larson, S. (2005) Crystallizing proteins on the basis of their precipitation diagram determined using a microfluidic formulator, *J. Synchroton. Rad.* **12**, 779-785.

Spraggon, G., Lesley, S.A., Kreusch, A., and Priestle, J.P. (2002) Computational analysis of crystallization trials, *Acta Cryst.* D**58**, 1915-1923.

SPSS, Chicago, IL.

Sumner, J.B. and Dounce, A.L. (1937) Crystalline catalase, *J. Biol. Chem.* **121**, 417-424.

Tardieu, A., Finet, S., and Bonnete, F. (2001) Structure of the macromolecular solutions that generate crystals, *J. Cryst. Growth* **232**, 1-9.

Teller, D.C. (1976) Accessible area, packing volumes and interaction surfaces of globular proteins, *Nature* **260**, 729-731.

The Gene Ontology Consortium, (2000) Gene Ontology: tool for the unification of biology, *Nature Genet.* **25**, 25-29.

Tran, T.T., Sorel, I., and Lewit-Bentley, A. (2004) Statistical experimental design of protein crystallization screening revisited, *Acta Cryst.* D**60**, 1562-1568.

Tsai, C.-J., Lin, S.L., Wolfson, H.J., and Nussinov, R. (1997) Studies of protein-protein interfaces: a statistical analysis of the hydrophobic effect, *Protein Sci.* **6**, 53-64.

Uehling, M.D. (2005) Need proteins? Just do it in Canada, *Bio-IT World*, November 2005.

Urquhart, B.L., Cordwell, S.J., and Humphery-Smith, I. (1998) Comparison of predicted and observed properties of proteins encoded in the genome of Mycobacterium tuberculosis H37Rv, *Biochem. Biophys. Res. Comm.* **253**, 70-79.

Valafar, H., Prestegard, J.H., and Valafar, F. (2002) Datamining protein structure databanks for crystallization patterns of proteins, *Ann. NY Acad. Sci.* **980**, 13-22.

Valdar, W.S.J. and Thornton, J.M. (2001) Conservation helps to identify biologically relevant crystal contacts, *J. Mol. Biol.* **313**, 399-416.

Van Bogelen, R.A., Schiller, E.E., Thomas, J.D., and Neidhardt, F.C. (1999) Diagnosis of cellular states of microbial organisms using proteomics, *Electrophoresis* **20**, 2149-2159.

Veesler, S. and Boistelle, R. (1999) *Crystallization of Nucleic Acids and Proteins: A Practical Approach*, edited by A. Ducruix and R. Giege, pp. 313-340. Oxford University Press Inc, New York, NY.

Vivares, D., Kaler, E.W., and Lenhoff, A.M. (2005) Quantitative imaging by confocal scanning fluorescence microscopy of protein crystallization via liquid-liquid phase separation, *Acta Cryst.* D**61**, 819-825.

Walter, T.S., Diprose, J.M., Mayo, C.J., Siebold, C., Pickford, M.G., Carter, L., Sutton, G.C., Berrow, N.S., Brown, J., Berry, I.M., Stewart-Jones, G.B.E., Grimes, J.M., Stammers, D.K., Esnouf, R.M., Jones, E.Y., Owens, R.J., Stuart, D.I., and Harlos, K. (2005) A procedure for setting up high-throughput nanolitre crystallization experiments. Crystallization workflow for initial screening, automated storage, imaging and optimization, *Acta Cryst.* D**61**, 651-657.

Westbrook, J., Feng, Z., Chen, L., Yang, H., and Berman, H.M. (2003) The Protein Data Bank and structural genomics, *Nucleic Acids Res.* **31**, 489-491.

Widmer, H. and Jahnke, W. (2004) Protein NMR in biomedical research, *Cell. Mol. Life Sci.* **61**, 580-599.

Wilson, W.W. (2003) Light scattering as a diagnostic for protein crystal growth-a practical approach, *J. Struct. Biol.* **142**, 56-65.

Wooh, J.W., Kidd, R.D., Martin, J.L., and Kobe, B. (2003) Comparison of three commercial sparse-matrix crystallization screens, *Acta Cryst.* D**59**, 769-772.

Wootton, J.C. (1994) Sequences with unusual amino-acid compositions, *Curr. Opin. Struct. Biol.* **4**, 413-421.

Yang, A.-S., Gunner, M.R., Sampogna, R., Sharp, K., and Honig, B. (1993). On the calculations of pK_as in proteins, *Prot. Struct. Funct. Genet.* **15**, 252-265.

Yee, A., Chang, X., Pineda-Lucena, A., Wu, B., Semesi, A., Le, B., Ramelot, T., Lee, G.M., Bhattacharyya, S., Gutierrez, P., Denisov, A., Lee, C.H., Cort, J.R., Kozlov, G., Liao, J., Finak, G., Chen, L., Wishart, D., Lee, W., McIntosh, L.P., Gehring, K., Kennedy, M.A., Edwards, A.M., and Arrowsmith, C.H. (2002) An NMR approach to structural proteomics, *Proc. Natl. Acad. Sci. USA* **99**, 1825-1830.

Zarranz, J.J., Alegre, J., Gomex-Esteban, J.C., Lezcano, E., Ros, R., Ampuero, I., Vidal, L., Hoenicka, J., Rodriguez, O., Atares, B., Llorens, V., Tortosa, E.G., del Ser, T., Munoz, D.G., and de Yebenes, J.G. (2004) The new mutation, E46K, of alpha-synuclein causes Parkinson and Lewy body dementia, *Ann. Neurol.* **55**, 164-173.

Zeelen, J.P., Hiltunen, J.K., Ceska, T.A., and Wierenga, R.K. (1994) Crystallization experiments with 2-enoyl-CoA hydratase, using an automated 'fast-screening' crystallization protocol, *Acta Cryst.* D**50**, 443-447.

Zhu, D.–W., Garneau, A., Mazumdar, M., Zhou, M., Xu, G.–J., and Lin, S.–X. (2006) Attempts to rationalize protein crystallization using relative crystallizability, *J. Struct. Biol.* **154**, 297-302.