

A High-Throughput Macromolecule Characterization System

By

©2013

Jae Hyun Kim

Submitted to the Bioengineering Graduate Program and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Chairperson C. Russell Middaugh, Ph. D

David B. Volkin, Ph. D

Stevin H. Gehrke, Ph. D

Kyle V. Camarda, Ph. D

Prajna Dhar, Ph. D

Date Defended: May 14th, 2013

The Dissertation Committee for Jae Hyun Kim

certifies that this is the approved version of the following dissertation:

A High-Throughput Macromolecule Characterization System

Chairperson

C. Russell Middaugh, Ph. D

Date approved: May 14th, 2013

Abstract

The size and complexity in structure of biopharmaceutical products makes them more susceptible to chemical or structural changes leading to lower potency or altered immunogenicity. Sustaining the stability of macromolecules becomes one of the greatest challenges in the development of biopharmaceutical products. The biophysical characterization of macromolecules is an essential step in stable formulation development. Structural changes of macromolecules in response to various environmental stresses or solution additives are measured using various techniques, and can then be analyzed using the empirical phase diagram (EPD).

The empirical phase diagram (EPD) is a colored representation of overall structural integrity and conformational stability of macromolecules in response to various environmental perturbations. Numerous proteins and macromolecular complexes have been analyzed by EPDs to summarize results from large data sets from multiple biophysical techniques. The current EPD method suffers from a number of deficiencies including lack of a meaningful relationship between color and actual molecular features, difficulties in identifying contributions from individual techniques, and a limited ability to be interpreted by color blind individuals. Three improved data visualization approaches are proposed as techniques complementary to the EPD. Experimental data sets can be visualized as (1) RGB colors using three-index empirical phase diagrams, (2) equiangular polygons using radar charts, and (3) human facial features using Chernoff face diagrams.

Recent development of high-throughput and multimodal spectrophotometers help rapidly collect the large volume of data that is required to create EPDs. Incompatible data formats of

various instruments and heterogeneous analysis software are, however, standing in the way of quickly organizing and analyzing such large volumes of data. It is essential to develop dedicated analysis software for such biophysical data to achieve high-throughput systems, in terms of both hardware and software, for biophysical characterization of macromolecules.

For this purpose, a web-based software framework called MiddaughSuite was developed in this work. The software was designed to easily handle data from various instruments, quickly analyze data using multiple mathematical functions, visualize data in the forms of graphs and diagrams including EPDs, radar charts and Chernoff face diagrams, and share data with other researchers.

Dedicated to
my wife Seungyeon and my son Ian

Acknowledgements

There are several people to whom I owe my sincere thanks and appreciation for helping me complete this dissertation. First, I offer my deepest thanks to my advisor, Dr. Russ Middaugh, for his continuous support, encouragement, and knowledge throughout the years. He truly inspired me to pursue this research and fed me with innovative new ideas. His passion for science has always motivated me to learn more in this field of study. It was my pleasure to find a way to implement his inspirations. Second, I would also like to give a special thank you to Dr. David Volkin for his valuable ideas about the visualization methods. He suggested a lot of applications for the new visualization methods and helped me develop the methods with constant feedback. Last, I offer my deepest thanks to Drs. Stevin Gehrke, Kyle Camarda, and Prajna Dhar for their time and valuable comments as the members of my dissertation committee.

I would like to express my sincere gratitude to Dr. Sangeeta Joshi and all of the members of the Macromolecule and Vaccine Stabilization Center. They spent a lot of time to test my software and gave me invaluable feedback. Without their help, I could not have completed MiddaughSuite software. Furthermore, they created a friendly and supportive working environment and I have enjoyed working with everyone.

My family has also made my education and research possible. My loving wife, Seungyeon, provided encouragement and support throughout my graduate work. No words can describe my appreciation for everything she has done for me. I love you so much. My dear son, Ian, always makes me happy. You are the greatest part of me!

Table of Contents

CHAPTER 1. INTRODUCTION.....	1
CHAPTER 2. IMPROVED DATA VISUALIZATION TECHNIQUES FOR ANALYZING MACROMOLECULE STRUCTURAL CHANGES.....	10
2.1 INTRODUCTION	10
2.2 MATERIALS & METHODS.....	12
2.2.1 <i>Materials</i>	12
2.2.2 <i>Construction of Empirical Phase Diagrams</i>	13
2.2.3 <i>Calculation of Structural Indices</i>	14
2.2.4 <i>Construction of Three-Index Empirical Phase Diagrams</i>	21
2.2.5 <i>Construction of Radar (Star) Diagrams</i>	23
2.2.6 <i>Construction of Chernoff Face Diagrams</i>	25
2.2.7 <i>Clustering Analysis of Phases</i>	26
2.3 RESULTS.....	28
2.4 DISCUSSION.....	37
2.5 REFERENCES	41
2.6 SUPPLEMENTARY FIGURES	44
CHAPTER 3. MIDDAUGHSUITE: A WEBSITE FOR BIOPHYSICAL CHARACTERIZATION OF MACROMOLECULES.....	51
3.1 OVERVIEW	51
3.2 USERS' REQUIREMENTS.....	52
3.2.1 <i>Quick and Convenient Data Analysis Practice</i>	52
3.2.2 <i>Laboratory Data Management</i>	53

3.2.3	<i>Share Data with Others</i>	53
3.2.4	<i>Ease of Use and Access</i>	54
3.3	FUNCTIONAL FEATURES	54
3.3.1	<i>Available as a website</i>	54
3.3.2	<i>Multidimensional Matrix</i>	54
3.3.3	<i>Uploading and extraction of data from various instruments</i>	57
3.3.4	<i>Analysis Functions</i>	60
3.3.5	<i>Visualizations</i>	60
3.3.6	<i>Share Data with Others</i>	62
3.4	IMPLEMENTATION	62
3.4.1	<i>Overview</i>	62
3.4.2	<i>Python Programming Language and Libraries</i>	64
3.4.3	<i>MiddaughSuite Client Page Implementation</i>	66
3.4.4	<i>MiddaughSuite Server Program Implementation</i>	67
	CHAPTER 4. THE WEB USER INTERFACE OF MIDDAUGHSUITE	69
4.1	OVERVIEW	69
4.2	ADMINISTRATION.....	69
4.3	MIDDAUGHSUITE MAIN SCREEN	71
4.4	UPLOAD MENU	73
4.4.1	<i>The Applied Photophysics Chirascan</i>	75
4.4.2	<i>The Photon Technology International (PTI) Fluorometer</i>	77
4.4.3	<i>The Avacta Optim 1000</i>	82

4.4.4 <i>The OLIS Protein Machine</i>	85
4.4.5 <i>User-Supplied Excel File</i>	88
4.5 DATA MENU.....	91
4.5.1 <i>Project List Panel</i>	92
4.5.2 <i>Tag Panel</i>	93
4.5.3 <i>Data List Panel</i>	94
4.5.4 <i>Project Detail Panel</i>	97
4.5.5 <i>Data Detail Panel</i>	99
4.6 FIGURE MENU	108
4.6.1 <i>Line Graph</i>	109
4.6.2 <i>Multiple Line Graph</i>	120
4.6.3 <i>Bar Graph</i>	122
4.6.4 <i>Contour Graph</i>	123
4.6.5 <i>Color Plots</i>	124
4.6.6 <i>Scatter Plot Matrix</i>	125
4.6.7 <i>Empirical Phase Diagram</i>	129
4.6.8 <i>Multiple Empirical Phase Diagrams</i>	134
4.6.9 <i>Radar Chart</i>	135
4.6.10 <i>Chernoff Face Diagram</i>	137
4.7 ANALYSIS MENU.....	138
4.7.1 <i>Element-wise Subtraction</i>	139
4.7.2 <i>Average</i>	140

4.7.3 Normalization and Normalization2D.....	141
4.7.4 Savitzky-Golay Filter.....	141
4.7.5 Spline Interpolation.....	143
4.7.6 Peak Picking by Mean Spectral Center of Mass	144
4.7.7 k-Means Clustering	145
4.7.8 Singular Value Decomposition.....	146
4.8 TABLE MENU	147
4.9 SEARCH MENU	148
CHAPTER 5. CONCLUSION	150

List of Figures

Figure 2.1 Experimental data for Bovine Serum Albumin (BSA) measured as a function of temperature at indicated pH values (A, B, C) and their corresponding structural indices (D, E, F). (A) CD signal at 222 nm, (B) Intrinsic fluorescence peak position, (C) Static light scattering at 295 nm, (D) Secondary Structure Index calculated from (A), (E) Tertiary Structure Index calculated from (B), and (F) Aggregation Index calculated from (C). A dashed line in (C) represents a cut-off value. Data in Figures 1A-C were published previously.¹⁰ 15

Figure 2.2 Experimental data for SP1732 measured as a function of temperature at indicated pH values (A, B, C), their intermediate structural indices (D, E, F) and the final structural indices (G, H, I). (A) CD signal at 222 nm, (B) Intrinsic fluorescence peak position, (C) Static light scattering at 295 nm, (D) Intermediate Secondary Structure Index calculated from (A), (E) Intermediate Tertiary Structure Index calculated from (B), (F) Intermediate Aggregation Index calculated from (C), (G) Secondary Structure Index, (H) Tertiary Structure Index, and (I) Aggregation Index. Dashed lines in (D, E, F) represent cut-off values. Data in Figures 2.2A-C were published previously.²⁴ 17

Figure 2.3 (A) Example of a radar chart with eight variables (experimental methods). All variables are set to a value of 1 with intervals of 0.2 (see text). (B) Chernoff face diagrams with seven variables. Each variable can vary from zero to one. Three faces are constructed with all variables set at 0, 0.5 and 1. 24

Figure 2.4 EPDs of the conformational stability of BSA as a function of pH and temperature. (A) Original EPD created using biophysical data in Figure 2.1A-C. (B) Three-Index EPD created using structural indices as shown in Figure 2.1D-F. 28

Figure 2.5 (A) Radar chart and (B) Chernoff face diagram of the conformational stability of BSA as a function of pH and temperature using data from the three structural indices from Figure 2.1D-F. Secondary and tertiary structure indices are inverted for both cases to represent the native state as (A) a dot and (B) a smiling face. Solid blue lines indicate an example of clustering results (see text). 30

Figure 2.6 (A) Three-Index EPD, (B) clustered Radar Chart, and (C) clustered Chernoff Face Diagram for the conformational stability of BSA (using data from three structural indices from Figure 2.1D-E). Each cluster is represented as a radar chart or Chernoff face diagram which averages the structural indices of all images that belong to the cluster. Six structural phases were observed empirically; 1: native state, 2: molten globular state, 3: aggregated, 4-6: structurally altered states due to low pH and/or high temperature without aggregation. 31

Figure 2.7 (A) EPD, (B) Three-Index EPD, (C) Radar Chart and (D) Chernoff Face Diagram for the protein antigen SP1732 as a function of temperature and pH. Figure 2.7A is created using

biophysical data in Figures 2.2A-C, while Figure 2.7B-D are created using structural indices shown in Figure 2.2G-I. Three structural regions are observed; 1: native state, 2: molten globular state, and 3: aggregated state. 32

Figure 2.8 (A) EPD, (B) Three-Index EPD, (C) Radar Chart and (D) Chernoff Face Diagram for the protein antigen HAC1 as a function of temperature and pH. Figure 2.8A is created using biophysical data in Supplementary Figures S1A-D, while Figures 2.8B-D are created using structural indices shown in Supplementary Figures S1E-H. Six structural phases are observed; 1: native state, 3: aggregated state, 2, 4-6: structurally altered state due to low pH and/or high temperature without aggregation. 35

Figure 3.1 Examples of Graphs and Diagrams Supported in MiddaughSuite 61

Figure 3.2 MiddaughSuite Architecture 64

Figure 4.1 MiddaughSuite Administration Page 70

Figure 4.2 MiddaughSuite Main Screen: (A) Before a user is logged in, (B) After logged in. 72

Figure 4.3 Upload Screen: (A) when uploading succeeds, (B) when uploading fails 74

Figure 4.4 (A) Chirascan (B) Chirascan-plus ACD (Automated Circular Dichroism) 75

Figure 4.5 Example of Chirascan data dimensions..... 76

Figure 4.6 PTI Fluorometer 77

Figure 4.7 Example of PTI *dsx* data dimensions 79

Figure 4.8 Example of PTI *ana* data dimensions..... 81

Figure 4.9 Avacta Optim 1000..... 82

Figure 4.10 Avacta Optim Data in Exported Excel 83

Figure 4.11 Example of Avacta Optim data dimensions ordered by temperature 84

Figure 4.12 Example of Avacta Optim data dimensions ordered by time 85

Figure 4.13 Olis Protein Machine 85

Figure 4.14 Example of Olis data directory structure. This figure is from the online brochure (http://www.olisweb.com/literature/pdf/Multiscan_brochure_low-res.pdf)..... 86

Figure 4.15 Example of Olis Zipped Directory data dimensions (ABS data)	87
Figure 4.16 (A) Example of an Excel template file (B) Uploaded data dimensions from (A).....	89
Figure 4.17 Data menu screen: (A) Project List Panel, (B) Tag Panel, (C) Data List Panel, (D) Project Detail Panel, (E) Data Detail Panel	91
Figure 4.18 The popup menu for the project list panel.....	93
Figure 4.19 Example of Multiple Tag Selection with AND operation.....	94
Figure 4.20 Search Filter.....	94
Figure 4.21 The popup menu for the data list panel	96
Figure 4.22 Project Detail Panel	98
Figure 4.23 Data Detail Panel.....	100
Figure 4.24 Dimension Field in the Data Detail Panel	101
Figure 4.25 The popup menu for the dimension field in the data detail panel.	103
Figure 4.26 The popup menu for the individual axis field in the data detail panel.	103
Figure 4.27 A popup window for the <i>Edit Values</i> menu.	104
Figure 4.28 A popup window for the <i>Remove Values</i> menu.	105
Figure 4.29 A popup window for the <i>Split Axis</i> menu.....	107
Figure 4.30 Figure menu screen. The screen consists of Menu, Options, and Graph Viewing Area.	109
Figure 4.31 Fluorescence spectra of emission wavelengths from 305 to 390nm as a function of temperature and pH: (A) Axes and values and (B) An example of a line graph at a specific pH value. The emission wavelength is assigned to the x axis and the temperature axis to line colors. (C) An example of a line graph. The emission wavelength is assigned to the x axis, the temperature axis to line colors, and the pH axis to line markers.	110
Figure 4.32 Data Selection in the Options panel	111
Figure 4.33 X Axis Selection in the Options panel	112

Figure 4.34 (A) The emission spectra and (B) the temperature melt of the same fluorescence measurement.	113
Figure 4.35 Error Bars in the Options panel.....	113
Figure 4.36 Line Color in the Options panel	114
Figure 4.37 (A) Line Marker in the Options Panel and (B) Pre-defined Markers in the MiddaughSuite.....	115
Figure 4.38 (A) “Use Same as Line Color” Option (B) Different markers are assigned to temperature values as well as line colors.....	116
Figure 4.39 (A) Line Style in the Options panel (B) Pre-defined line styles in the MiddaughSuite	117
Figure 4.40 (A) Manual Zoom in the Options panel (B) The line graph with specified range. The system calculates minimum of the y axis because it is not specified.	118
Figure 4.41 Same size (450×600) of figures with different size options.....	119
Figure 4.42 Group Selection in the Options panel.....	120
Figure 4.43 An example of Multiple Line Graph. Group axes selected in the Figure 4.42 are used.	121
Figure 4.44 An example bar graph.	122
Figure 4.45 Example contour plots with (A) number of levels = 10 and (B) number of levels = 30	123
Figure 4.46 An example of a color plot.	124
Figure 4.47 An example of a scatter plot matrix.	125
Figure 4.48 (A) Particle information in a sample organized in Excel file from MFI data (B) MiddaughSuite data of 4000 particles with 11 parameters at pH 4, 6, and 8.	126
Figure 4.49 (A) Sample Axis Selection and (B) XY Axis Selection in the Options panel	127
Figure 4.50 (A) Line Color and (B) Options in the Options panel.....	128
Figure 4.51 Examples of dot options.	129

Figure 4.52 Empirical Phase Diagram of BSA with 6 clusters and RGB components shown...	130
Figure 4.53 Options panel for EPD menu: (A) Axes Selection, (B) Color Selection, (C) Clustering Display and (D) RGB Components.....	131
Figure 4.54 Clustering information used in Figure 4.52.....	133
Figure 4.55 Multiple Empirical Phase Diagram menu: (A) Group Selection, (B) Layout in the Options panel and (C) An example of three EPDs created using this menu.	134
Figure 4.56 (A) Method Selection in the Options panel and (B) an example Radar chart.....	135
Figure 4.57 Clustering Display in the Options menu for Radar Charts and Chernoff Face Diagram.....	136
Figure 4.58 Method Selection in the Options panel for Chernoff Face Diagram.....	137
Figure 4.59 Analysis menu screen: (A) Functions list, (B) selected function name, (C) brief description of selected function, (D) “Execute this function” button, and (E) data, value range, and options panel for the selected function.....	138
Figure 4.60 (A) Example of Fluorescence Data for Buffer Subtraction and Average functions, (B) Operand Selection for Element-wise Subtraction. This function calculates A-B. (C) Arg Selection for Average. This function calculates the average and standard deviation of all selected values.	140
Figure 4.61 (A) Arg Selection for Normalization, (B) Arg Selection for Normalization2D, (C) Example of normalized fluorescence spectra, and (D) Example of normalized fluorescence spectra in which the relative intensity of the spectra as a function of temperature is preserved using Normalization2D.....	142
Figure 4.62 Data Selection for Spline Interpolation function.....	143
Figure 4.63 (A) Trp fluorescence spectra of Chymotrypsin at pH 7 (B) Peak Shift by Mean Spectral Center of Mass	144
Figure 4.64 Input panels for k-Means Clustering and Singular Value Decomposition.....	146
Figure 4.65 Table menu screen: (A) Sheet Selection, (B) Axes Selection, (C) “View” button, and (D) Table view and “Download as Excel file” button	148
Figure 4.66 Search menu screen: (A) Search panel, (B) Project Detail Panel, (C) Data List Panel, (D) Data Detail Panel, and (E) “Import This Project To Your Account” button.....	149

List of Tables

Table 2.1 The Relationship of Colors to Protein Structural Features in Three-Index Empirical Phase Diagrams.....	21
Table 3.1 An Example List of Experimental and Environmental Conditions for Intrinsic Fluorescence Spectroscopic Data	55
Table 3.2 A List of Supported Instrumentation and Their File Formats	57
Table 3.3 An Example Matrix Converted from a Photon Technology International Fluorometer File (.ana)	58
Table 3.4 A List of Analysis Functions in MiddaughSuite	59

Chapter 1. Introduction

Biopharmaceutical products have become one of the major driving forces for the growth of the pharmaceutical industry over the last two decades. The overall sales of all biologics approached USD 100 billion out of total USD 600 billion for the pharmaceutical industry in 2009.¹ It is also anticipated that the biologics market will continuously grow to up to 49% of the total drug sales by 2018.^{2,3} Currently, over 200 biopharmaceutical products are available in the market¹ with a large number of new arrivals each year. Since 2009, half a dozen new biologics license applications (BLAs) have been approved yearly by the US Food and Drug Administration (FDA)'s Center for Drug Evaluation and Research (CDER).⁴ In addition, the emergence of follow-on biologics - also known as biosimilars - is expected to be another growth factor for the entire pharmaceutical market, although the innovator drugs will suffer from a significant loss of revenue due to the expiry of their patents.^{2,3}

The success of the biopharmaceutical market demands more rapid development of candidate molecules. One of the greatest challenges in the development of biologics is to ensure their stability during all processes from manufacturing to administration. Due to their large size and complexity of structure, biopharmaceutical products are typically susceptible to degradation by a wide variety of physical and chemical stresses. These conformational or compositional changes of macromolecules may lead not only to lower potency but also to altered immunogenicity.⁵ To prevent or at least delay the degradation processes, pharmacologically inactive ingredients referred to as excipients are mixed with the macromolecule of interest. The goal of this “formulation” development is to find the right type and amount of (possibly multiple)

excipients that are most effective in protecting the macromolecule from the various types of degradations that may be present during the shelf-life of the product.

Although monitoring the stability of biologics with all possible combinations of excipients during actual storage is the most accurate method, the time and cost requirement typically discourages this approach. Instead, more extreme environmental conditions such as elevated temperature, suboptimal pH, and ionic strength are often chosen in an attempt to predict long-term storage stability.^{6,7} These conditions are selected to accelerate the major degradation pathways of the macromolecule. For example, a candidate formulation can be anticipated to have longer-term stability than other candidates if the structural changes of the candidate formulation occur at higher temperature than those of others. Since greater energy is needed to alter the structure of the macromolecule at higher temperature, the macromolecule is expected to be stable for a longer period of time. This prediction is, however, not always accurate. The short-term accelerated stability test is popular because of its time and cost effectiveness despite of its potential inaccuracy.

Even with short-term studies, the number of experiments required for a formulation study is still enormous considering the number of popular excipients and their combinations. Various design and optimization methods have been proposed to reduce the relevant search space. A detailed understanding of macromolecular behavior should be performed as the first priority before any design methods are applied. An initial biophysical characterization of a macromolecule attempts to draw a comprehensive picture describing how the macromolecule responds to environmental stress. This analysis should reveal conditions in which the macromolecule undergoes structural changes (e.g. unfolds), forms soluble or insoluble

aggregates, and becomes chemically modified. Based on the observed macromolecular behavior, more detailed experiments can be designed with narrowed condition to provide further information.

Observation of structural changes in a macromolecule is an essential part of characterization because preservation of macromolecular structure is crucial to maintain its functionality. X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy are the two major techniques used to determine macromolecular structure.⁸ Although these techniques can provide high resolution information about macromolecular structure, they are generally not appropriate for formulation studies. Crystallized samples are necessary for X-ray crystallography but successful crystallization is dependent on the nature of the sample. In addition, the crystal structure does not directly reflect the dynamic motions of proteins in pharmaceutically relevant solution states. For NMR spectroscopy, molecular weight and the need for isotopic labeling become limiting factors. Perhaps most importantly, these techniques are both too time-consuming and complex for regular and frequent use in macromolecular pharmaceutical analyses.

Another high-resolution technique for protein structural analysis is hydrogen deuterium exchange mass spectroscopy (HDX-MS).⁸⁻¹⁰ In this method, the macromolecular dynamics can be studied by monitoring the exchange rate of hydrogen with deuterium when the macromolecule is placed in deuterated water (D₂O). HDX-MS has become a useful tool for detecting a ligand binding or protein-protein interactions, to analyze post-translational modifications, to study comparability of follow-on biologics and for many other related topics.^{8,9} This technique, however, still requires considerable time for both experiments and data analysis.

Compared with the previously described techniques, classic biophysical methods have less discriminative powers due to their intrinsically lower resolution. Individual techniques such as UV absorbance, circular dichroism (CD), fluorescence, light scattering, Fourier transform infrared (FTIR) spectroscopy and differential scanning calorimetry (DSC) may not be able to detect small, subtle structural changes or even larger changes when the altered proteins are present in small amounts.⁸ Each method is also limited by its ability to only detect a few aspects of macromolecular behavior. For example, intrinsic Tryptophan fluorescence spectroscopy monitors changes in the tertiary structure, while CD and FTIR spectroscopy are more often used to investigate secondary structure. Static light scattering is a useful method that is sensitive to aggregate formation. Dynamic light scattering is especially useful for molecular size determination. UV absorbance spectroscopy is more often used to calculate concentration and more recently has been employed to monitor conformational changes.

Although the amount of information obtainable from each method is much less than higher resolution technologies such as X-ray crystallography, NMR spectroscopy and HDX-MS, a collection of orthogonal methods can greatly increase the quality of such information and lead to a better understanding of overall macromolecular behavior. For biophysical characterization of macromolecules, physical methods in combination with environmental perturbations such as temperature, pH, and solute additives can be used to produce a very comprehensive picture of a macromolecule. As the number of environmental perturbations and orthogonal physical methods are increased, the identification of internal system behavior (i.e., possible conformational states) can be achieved with higher accuracy. Obviously, the amount of experimental data increases drastically as the number of conditions and methods increases. A method to extract interpretable

macromolecular information from such large volumes of data is a key to the successful characterization of macromolecules for pharmaceutical use.

The empirical phase diagram (EPD) method^{11,12} summarizes experimental data and visualizes macromolecular behavior in the form of a colored diagram. It provides a global picture of structural changes in a macromolecule under a wide range of solution conditions as a kind of stress/response picture. The method utilizes singular value decomposition (SVD) and linear algebra to extract major patterns from a large collection of data. The EPD method requires a number of steps. First, the data should be extracted and stored from each instrument. Second, the data should be processed to correctly represent macromolecular behavior. Each physical method may require different types of processing. This processing may include buffer subtraction, smoothing, normalization, and peak picking. Third, processed data should be organized into one matrix that will be used as an input to SVD. Finally, the three largest components from SVD results are selected and visualized into a two-dimensional colored diagram.

Although the EPD method is useful for summarizing and visualizing patterns found in a large set of data, the use of color as a mean to display patterns inherently possesses several disadvantages. In the EPD, changes in colors represent alterations in structure of the macromolecules as inferred from experimental data. The color itself is randomly chosen; therefore, there is no explicit and common relationship between a color and the protein's structural state. In addition, this method is useless for people with color deficiencies in vision as well as the blind.

To overcome these disadvantages, three new data visualization approaches are introduced in Chapter 2. Instead of analyzing all experimental data together to construct an EPD, each

method used to determine secondary, tertiary, and quaternary structural changes is analyzed separately and mapped to a fixed color set. In this manner, a three-index EPD can show structural changes of macromolecules with defined colors. For example, the color yellow is always used to display the native state. Radar charts and Chernoff face diagrams are color-free diagrams used to represent EPDs. They utilize certain icons such as radar plots and facial expressions to display data under given conditions. They can accommodate a larger number of types of data explicitly in a diagram than the original EPD. In the initial EPDs, only three major components can be displayed as red, green, and blue (RGB) components in a color.

The biophysical characterization of macromolecules using EPDs based on classical biophysical techniques requires a large volume of experimental data. Recent development of high-throughput and multi-functional instruments enables such large-scale data collection while it greatly reduces overall cost, time and labor. High-throughput fluorometers, for example, can measure fluorescence and light scattering of 48 samples at once. Multi-functional spectrophotometers can employ multiple spectroscopic methods including near and far CD, UV absorbance, fluorescence, and light scattering simultaneously for up to 6 samples. Another type of enhanced instrumentation is the automated spectrophotometer. Although each sample is processed sequentially, up to 96 samples can be handled automatically without interruption. With the help of these advanced instruments, biophysical characterization extensive data concerning macromolecules can be obtained within a single day, with a minimal amount of sample and effort. Current data analysis procedures, however, cannot adequately support such data generation speed due to a lack of dedicated software.

Generally, instrument manufacturers provide their own software to process their experimental data. Since each instrument uses its own data format, the software, in many cases, has ability to export its data to more popular file formats. Their software often provides some functionality to analyze data. It usually requires significant effort on the part of individual users to gather and organize the experiments data from the various instruments. Because of the incompatibility of various data formats, data collection and organization procedures generally require significant labor and time as the volume of data increases. Without effective management of large volumes of data, the efficiency of data analysis and consequent visualization can also be drastically decreased.

To overcome such difficulties, a web-based software framework called MiddaughSuite was designed and developed in this work. In the Chapter 3, detailed requirements from users for the software are listed. The software should be able to easily handle data from various instruments, quickly analyze data using multiple mathematical functions, visualize data in the forms of graphs and diagrams, and share data with other researchers. The software was developed using the python programming language because there exists many publicly available python libraries for scientific computing. The architecture and python libraries are described in Chapter 3 as well.

In the Chapter 4, the web interface of MiddaughSuite is demonstrated in detail. The main six menus – *Upload*, *Data*, *Figure*, *Analysis*, *Table*, and *Search* – provide functionality identified in the Chapter 3. Using the dynamic HTML technology, MiddaughSuite can provide easy-to-use and an intuitive interface to users.

The high-throughput macromolecule characterization system is obtained by combining high-throughput multimodal spectrophotometers and MidaughSuite software. This coupling provides researchers with a seamless workflow from data collection to analysis. The coupled system makes it possible to obtain various types of EPDs rapidly, possibly within a single day, with a minimal amount of protein and effort.

References

- 1 Walsh, G. Biopharmaceutical benchmarks 2010. *Nature biotechnology* **28**, 917-924, doi:10.1038/nbt0910-917 (2010).
- 2 Harrison, C. Patent watch: Dangling from the patent cliff. *Nature reviews. Drug discovery* **12**, 14-15, doi:10.1038/nrd3924 (2013).
- 3 EvaluatePharma. World Preview 2018 — Embracing the patent cliff. (2012). <<http://www.evaluatepharma.com/worldpreview2018.aspx>>.
- 4 Mullard, A. 2012 FDA drug approvals. *Nature reviews. Drug discovery* **12**, 87-90, doi:10.1038/nrd3946 (2013).
- 5 Manning, M. C., Chou, D. K., Murphy, B. M., Payne, R. W. & Katayama, D. S. Stability of protein pharmaceuticals: an update. *Pharmaceutical research* **27**, 544-575, doi:10.1007/s11095-009-0045-6 (2010).
- 6 Randolph, T. W. & Carpenter, J. F. Engineering challenges of protein formulations. *AIChE journal* **53**, 1902-1907 (2007).
- 7 Liu, H., Gaza-Bulseco, G. & Sun, J. Characterization of the stability of a fully human monoclonal IgG after prolonged incubation at elevated temperature. *Journal of chromatography. B, Analytical technologies in the biomedical and life sciences* **837**, 35-43, doi:10.1016/j.jchromb.2006.03.053 (2006).
- 8 Berkowitz, S. A., Engen, J. R., Mazzeo, J. R. & Jones, G. B. Analytical tools for characterizing biopharmaceuticals and the implications for biosimilars. *Nature reviews. Drug discovery* **11**, 527-540, doi:10.1038/nrd3746 (2012).

- 9 Majumdar, R. *et al.* Minimizing Carry-Over in an Online Pepsin Digestion System used for the H/D Exchange Mass Spectrometric Analysis of an IgG1 Monoclonal Antibody. *J. Am. Soc. Mass Spectrom.* **23**, 2140-2148, doi:10.1007/s13361-012-0485-9 (2012).
- 10 Manikwar, P. *et al.* Correlating excipient effects on conformational and storage stability of an IgG1 monoclonal antibody with local dynamics as measured by hydrogen/deuterium-exchange mass spectrometry. *Journal of pharmaceutical sciences*, doi:10.1002/jps.23543 (2013).
- 11 Kuelto, L. A., Ersoy, B., Ralston, J. P. & Middaugh, C. R. Derivative absorbance spectroscopy and protein phase diagrams as tools for comprehensive protein characterization: a bGCSF case study. *Journal of pharmaceutical sciences* **92**, 1805-1820, doi:10.1002/jps.10439 (2003).
- 12 Maddux, N. R., Joshi, S. B., Volkin, D. B., Ralston, J. P. & Middaugh, C. R. Multidimensional methods for the formulation of biopharmaceuticals and vaccines. *Journal of pharmaceutical sciences*, doi:10.1002/jps.22618 (2011).

Chapter 2. Improved Data Visualization Techniques for Analyzing Macromolecule Structural Changes

2.1 Introduction

At the heart of the study of protein biochemistry lies understanding the interrelationships between the structure and function of these complex macromolecules. Structural aspects are best considered in the light of protein three-dimensional structures as determined by x-ray crystallography and nuclear magnetic resonance (NMR). In many cases, however, such structures are unavailable or suffer from experimental limitations such as their origin in the crystalline state, or the requirement for high concentrations and/or isotope labeling. Thus, physical methods of lower structural resolution such as circular dichroism (CD), fluorescence, light scattering, Fourier transform infrared spectroscopy (FTIR), and differential scanning calorimetry (DSC) are often used, especially in combination with environmental perturbations such as temperature, pH and solute additives (e.g. chaotropes) to evaluate macromolecule higher order structure and stability in solution.

As an intermediate approach, results from multiple lower resolution methods can be combined to produce a more information-rich characterization of protein structure. One such method is known as the empirical phase diagram (EPD). This approach consists of a colored map of macromolecule behavior in which the structure of a macromolecule (or its complexes), as a function of various solution and environmental conditions, is represented by vectors corresponding to individual measurements such as CD, FTIR spectra, intrinsic fluorescence, dye

binding, DSC and light scattering. Using singular value decomposition (SVD), the three largest contributions to a vector are obtained and reduced to color based on an RGB scheme.^{1,2} Although the actual colors are somewhat arbitrary, changes in color are useful since they represent structural changes. In addition, differences between EPDs can be directly compared (e.g. between mutants of the same protein) by processing all of the data together.³ By making measurements as a function of variables such as temperature, pH, concentration, ionic strength, and mechanical stress, the colored EPD diagram represents an information rich type of “stress/response” diagram that has been useful in a wide variety of applications including stabilization and formulation of protein therapeutics and a variety of macromolecular vaccines.¹⁻

10

The current EPD method suffers from a number of deficiencies that primarily reflect the use of color to represent the state of the protein’s structural integrity. This includes the lack of meaningful relationship between color itself and actual molecular features (in EPDs, structural changes are represented by color transitions^{1,2,9}) as well as limitations resulting from color deficiencies in vision and blindness which is possessed by a substantial portion of the human population. Red-green color vision defects are found in about 8% of males and 0.5% of females among people of Northern European origin, 5% among Chinese and Japanese populations, 4% or less among individuals of African origin.¹¹ Here we describe three alternative data visualization approaches to the current EPD methodology which not only overcome these difficulties, but may provide additional information leading to enhancement of the EPD macromolecular structure characterization tool.

In the first approach, we take advantage of the direct relationships between far UV CD spectra and protein secondary structure, intrinsic fluorescence spectra and protein tertiary structure changes, and light scattering measurements and quaternary structure (or aggregation state) to assign direct protein structural meaning to colors. In the second and third methods, we eliminate the use of the color all together through employment of radar (or star) charts¹² and Chernoff faces¹³⁻¹⁵. These techniques are popular iconic displays of multivariate data in which attribute values are mapped to the features of the icons.¹⁶ For radar plots, values from the multiple physical measurements are related to the spikes of equiangular polygons. For the Chernoff face approach, these same data sets are represented by the position and size of facial features in a model face. The resulting patterns exhibit characteristics of the data which can be recognized by pre-attentive perception.¹⁶ These data visualization methods have been used in many other fields including information visualization^{16,17}, computer science^{18,19}, biology²⁰, education²¹ and health care^{22,23}. In this work, we explore the utility of three different representations with six different proteins using temperature and solution pH as independent stress variables and discuss the advantages and disadvantages of each data visualization approach.

2.2 Materials & Methods

2.2.1 Materials

To demonstrate various visualization techniques, previously published data (e.g., intrinsic Trp and extrinsic ANS fluorescence spectroscopy, CD, and static light scattering) were used in this work including bovine serum albumin (BSA)¹⁰, aldolase¹⁰, and chymotrypsin¹⁰ as well as serine threonine kinase protein (SP1732)²⁴ and pneumococcal surface antigen A (SP1650)²⁴ from *S.*

pneumonia, and hemagglutinin from the H1N1 influenza virus (HAC1)²⁵. The EPDs were constructed using these previously published data and then compared to the three newly proposed data visualization techniques introduced in this work: the three-index empirical phase diagram, radar diagram, and Chernoff face diagram. All six of the protein macromolecules were dialyzed into a 20mM citrate phosphate buffer (pH 3 to 8), at a total ionic strength of 0.15 using appropriate amounts of NaCl. A detailed description of the experimental methods employed to generate the structural data used in this work are described elsewhere.^{10,24,25}

2.2.2 Construction of Empirical Phase Diagrams

The experimental data for each of the six proteins were previously obtained from multiple techniques (e.g., CD, fluorescence, and light scattering) as a function of pH and temperature.^{10,24,25} The combinations of those conditions are aligned to form an $m \times n$ input matrix, in which m is the number of experimental techniques and n is the number of condition combinations (e.g., number of pH values \times number of temperature measurements). Singular value decomposition of the input matrix is employed to produce a factorization of unitary orthonormal bases matrices and a diagonal matrix composed of singular values. The three largest singular values determine the major contributing factors in the form of orthonormal basis vectors. These factors are then mapped to a RGB color scheme and are visualized as an EPD. Thus, regions of the plot with similar color indicate similar physical states of the macromolecule. A detailed explanation of the calculations involved in construction of EPDs can be found elsewhere.¹⁻³

2.2.3 Calculation of Structural Indices

We define a structural index as the degree of corresponding structural change within a given range of environmental stress conditions. Secondary, tertiary and quaternary (aggregation) structure indices are used here to represent the state of a macromolecule. The structural index is allowed to vary from a value of zero to one while indicating the lowest to highest amount of the corresponding structure. As an example, native tertiary structure can be represented by a value of 1 while significantly conformationally altered tertiary structure by 0. Similarly, the aggregation index is defined to have a range from 0 to 1 where the number 0 implies no aggregation whereas a value of 1 indicates the maximum level of protein aggregation observed during an experimental study.

Each index is calculated in accordance with experimental data that reflect corresponding conformational or association changes in a given protein. There is no limitation to the choice of experimental variables, as long as the experiment is a measure of the amount of structure. Taking BSA as an example, intrinsic fluorescence peak position shifts was used as an indicator of tertiary structural change. Far UV CD spectroscopy at a specific wavelength (e.g. 222nm) was selected to monitor changes in secondary structure. Static light scattering experiments were employed to detect the amount of aggregation. If a certain experimental technique is not applicable to a specific macromolecule (e.g., fluorescence in a protein lacking Trp), alternative biophysical methods can be substituted (e.g., near UV absorption or CD for tertiary structural change, FTIR for secondary structure, etc.). To calculate a structural index, an initial interpretation of the original experimental data is necessary. If the experimental data shows a

monotonic change over the experimental variables, simple normalization (and optional inversion) would be sufficient to calculate the index from zero to one.

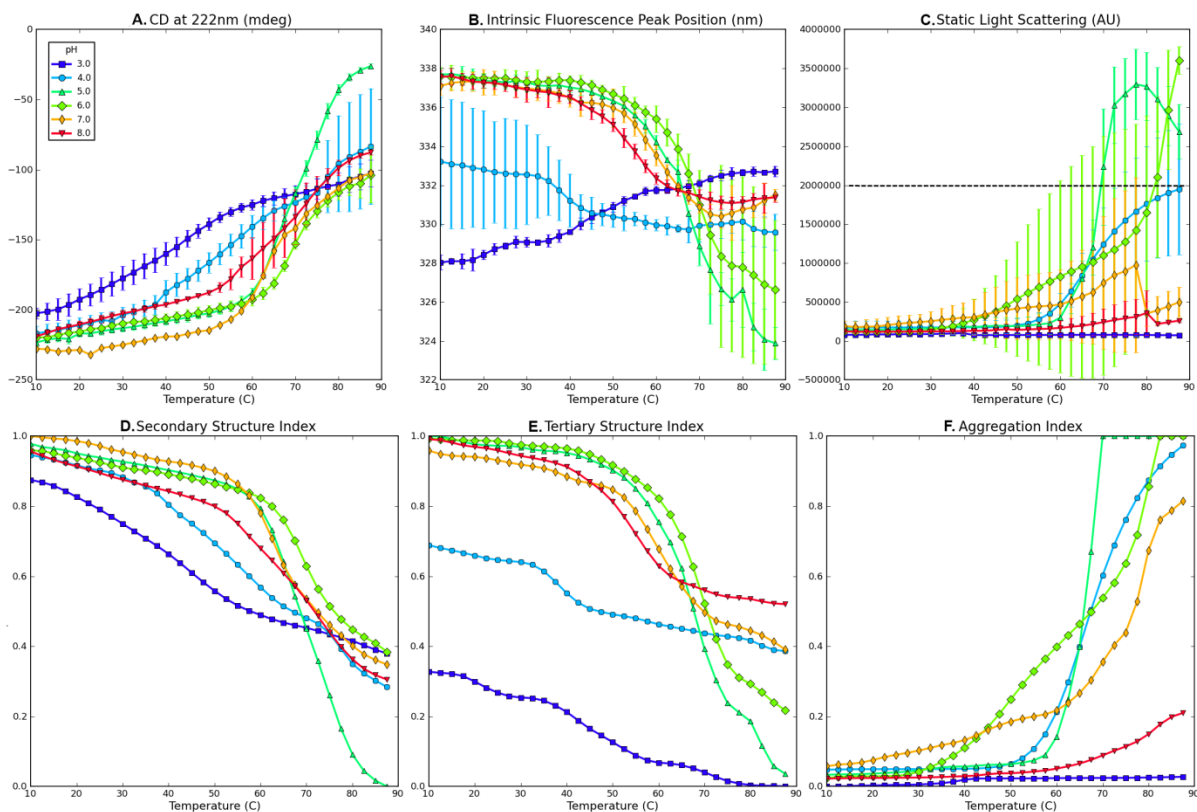


Figure 2.1 Experimental data for Bovine Serum Albumin (BSA) measured as a function of temperature at indicated pH values (A, B, C) and their corresponding structural indices (D, E, F). (A) CD signal at 222 nm, (B) Intrinsic fluorescence peak position, (C) Static light scattering at 295 nm, (D) Secondary Structure Index calculated from (A), (E) Tertiary Structure Index calculated from (B), and (F) Aggregation Index calculated from (C). A dashed line in (C) represents a cut-off value. Data in Figures 1A-C were published previously.¹⁰

To illustrate the basic procedure for preparing structural indices, the CD signal at 222 nm for BSA from pH 3 to 8 as a function of temperature (Figure 2.1A), decreased with distinct

transitions seen as the temperature increased, indicating a loss of protein secondary structure. The data define the range of possible changes for BSA's secondary structure over the given pH and temperature range as examined by CD. As seen in the Figure 2.1D, the secondary structure index of BSA was calculated by normalizing the inverted CD signals at a given pH and temperature with an increase in negative signal indicating a loss of structure.

The intrinsic fluorescence peak position shift in proteins is sensitive to the micro-environmental changes around tryptophan residues and was therefore selected to prepare a structural index for tertiary structure. Changes in the hydration state of indole side chains can produce either red or blue shifts corresponding to increases or decreases in surface exposures, respectively. This change can therefore be interpreted in terms of tertiary structural changes of the macromolecule. For example, the peak position of the intrinsic fluorescence spectra for BSA at pH 3 to 8 from 10 to 90°C is presented in Figure 2.1B. Unlike the previous CD data, the peak position shift at pH 3 occurs in a different direction than the shift at other pH values. Therefore, obtaining an index by normalization of the peak position shift data at all pH values could lead to misinterpretation of the tertiary structural index at pH 3, since it suggests that the tertiary structure is returning to a more native state as temperature increases. In this case, a comparison of the amount of deviation on each side of the native state value becomes more important than that of the direction. In such cases, we therefore invert the shift value to make it comparable to the behavior at other pH values. The same procedures are then applied to obtain the tertiary structure index as shown in Figure 2.1E. The aggregation index (Figure 2.1F) was determined from light scattering data (Figure 2.1C) as discussed below.

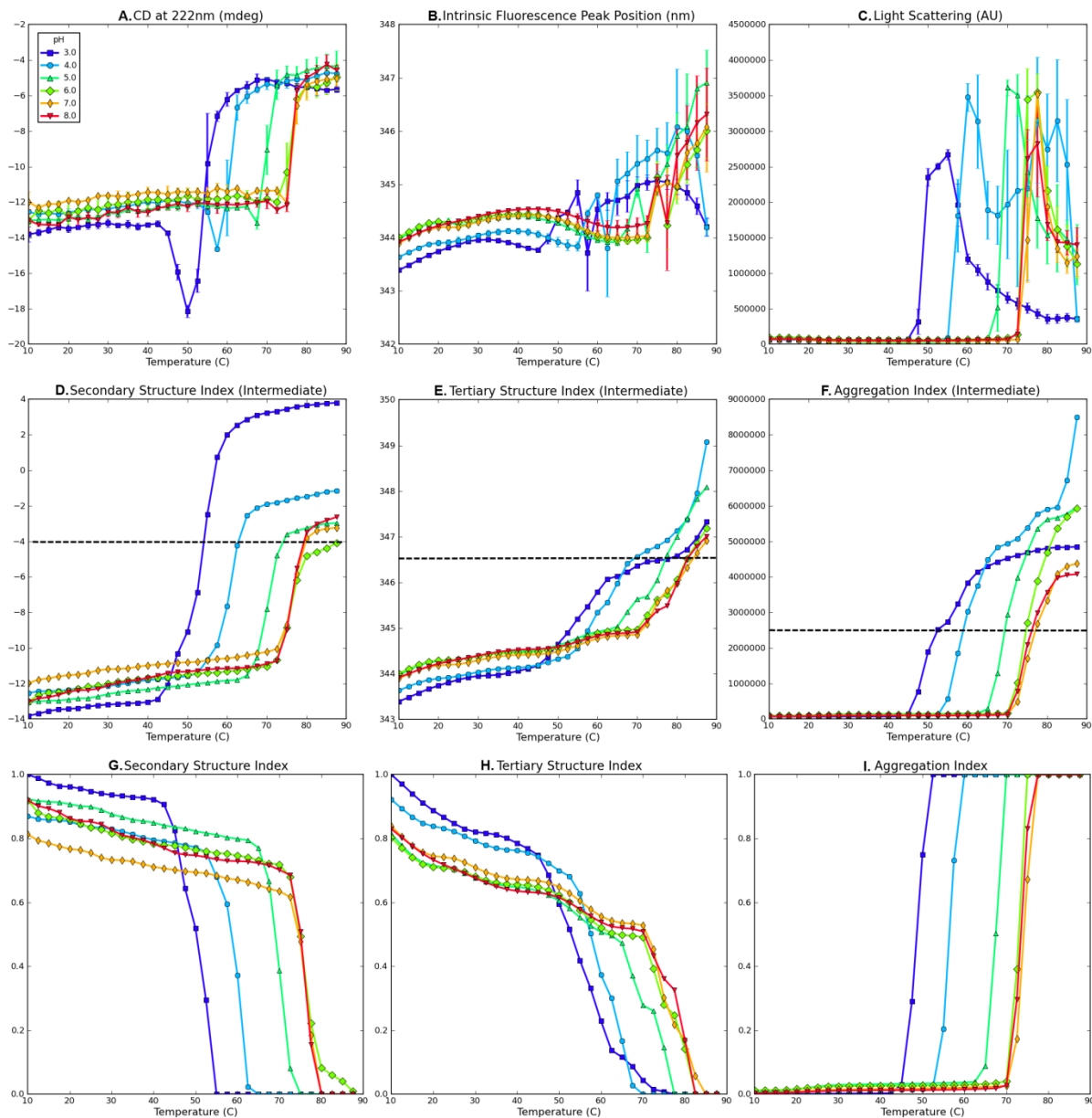


Figure 2.2 Experimental data for SP1732 measured as a function of temperature at indicated pH values (A, B, C), their intermediate structural indices (D, E, F) and the final structural indices (G, H, I). (A) CD signal at 222 nm, (B) Intrinsic fluorescence peak position, (C) Static light scattering at 295 nm, (D) Intermediate Secondary Structure Index calculated from (A), (E) Intermediate Tertiary Structure Index calculated from (B), (F) Intermediate Aggregation Index calculated from (C), (G) Secondary Structure Index, (H) Tertiary Structure Index, and (I) Aggregation Index. Dashed lines in (D, E, F) represent cut-off values. Data in Figures 2.2A-C were published previously.²⁴

It is common to obtain experimental data that does not manifest the slower, continuous changes observed with BSA. For example, the CD signal at 222nm for the protein antigen SP1732, as measured from pH 3 to 8 and 10 to 90°C, shows a sharp negative change that decreases in magnitude and converges as illustrated in Figure 2.2A. The sharp negative response may be induced by intermolecular interactions. The CD signal then disappears as the secondary structure of SP1732 becomes increasingly disrupted and aggregation sets in. Normalization of the inverted CD signal may not provide an accurate secondary structure index because higher values of the index at a negative peak do not necessarily indicate the secondary structure content in the native state. Thus, in this case, it is necessary to calculate the amount of the signal's deviation from its initial value. One method is to integrate the absolute value of the first derivative of the signal as described in equation (1). The constant C can be determined by the nature of the signal. The value of C should be either 1 or -1 for positive or negative correlation, respectively, between the signal and the amount of structure. In other words, a positive correlation ($C=1$) is defined when higher signal indicates less structure.

$$I(x) = f(x_0) + C \int |f(x)'| dx \quad (1)$$

where,

f : experimental measurement

x_0 : initial value

C : ± 1 , based on the nature of the signal

Figure 2.2D shows the result of applying equation (1) to data in Figure 2.2A with $C=1$. The CD signals are converted to increasing sigmoidal curves. The larger negative peak at pH 3, however, results in a much larger deviation from the initial value compared to other pH values. This result is inconsistent with the original CD signal in which the signals converge to indicate a particular level of unfolded structure at higher temperature. Normalization of individual pH values may adjust this inconsistency, but it will cause the difference in initial values to be lost. Rather, introducing an upper (or lower if $C= -1$)-bound cut-off might be a better option, because deviations beyond a certain level can be considered to be the same maximally unfolded (or structurally altered) state of the macromolecule. In addition, this approach conserves the initial values. The cut-off threshold value should be carefully chosen after the initial interpretation of the experimental data. Figure 2.2G shows the final secondary structure index after applying cut-off criteria to Figure 2.2D, followed by inversion and normalization as presented in equation (2). Equation (2) simply represents a combined procedure for cut-off, normalization, and inversion of the result from equation (1).

$$f_{index} = g\left(\frac{1 + C}{2} - C \frac{I_{max} - I(x)}{I_{max} - I_{min}}\right) \quad (2)$$

where

I_{max} : maximum or upper cut-off value of $I(x)$,

I_{min} : minimum or lower cut-off value of $I(x)$,

$$g(x) = \begin{cases} 1, & x > 1 \\ x, & 0 \leq x \leq 1 \\ 0, & x < 0 \end{cases}$$

The combination of equations (1) and (2) was applied to calculate the tertiary structure index and aggregation index from intrinsic fluorescence peak position shift and static light scattering data, respectively, as described in Figure 2.2H and I. The cut-off value for the aggregation index (Figure 2.2I) should be selected with the following considerations. It is generally observed that the intensity of the static light scattering signal decreases at higher temperature after it reaches a peak. This decrease in scattering intensity is usually caused by precipitation and settling of the sample, not by aggregation itself. The corresponding aggregation index, therefore, should reach and remain 1 after its peak value. Another consideration is the comparison among different environmental conditions (e.g. pH), when occasionally scattering signals are too excessive causing suppression of other signals. In this case, the choice of cut-off value will be subjective as to whether to use a lower cut-off value (to emphasize other relatively lower signals) or employ a higher cut-off value (to highlight more intense aggregation conditions). Figure 2.1C and F demonstrates the use of a lower cut-off value to exhibit aggregation at pH 4 and 7. Otherwise, these data might be overlooked because of higher signals at pH 5 and 6.

In summary, equations (1) and (2) are generalized equations to calculate structural indices from any sets of experimental data as illustrated in the two examples presented. Equation (1) converts any data set into a monotonically changing signal with its initial value, and the amount of deviation from its initial values, preserved. Equation (2) is then applied to the results from the equation (1) for inversion, normalization and cut-off. The constant C in equations (1) and (2) determines the direction of the signal. The I_{max} and I_{min} parameters in equation (2) determine the range of the signal for normalization and cut-off. It should be noted that

experimental noise in the measurements may affect the resulting structural index derived from equation (1), because small fluctuations in the first derivative will be accumulated and result in a gradual increase of the index. Thus, an appropriate smoothing algorithm such as the Savitzky-Golay smoothing filter²⁶ should be applied to the data to improve the signal to noise ratio, prior to application of equation (1).

2.2.4 Construction of Three-Index Empirical Phase Diagrams

The purpose of the three-index EPD is to present a colored diagram that not only displays the degree of change in a macromolecule's structure and association state in response to its environmental conditions, but the color itself can be related to changes in specific elements of protein structure. The degree of change is commonly studied using three structural levels: secondary structural change, tertiary structural change, and aggregation. The change in each aspect is defined by the previously introduced structural indices: Secondary structure Index (SI), Tertiary structure Index (TI), and Aggregation Index (AI).

Table 2.1 The Relationship of Colors to Protein Structural Features
in Three-Index Empirical Phase Diagrams

Color	Protein Structure
Yellow (Green + Red)	Native State
Red, Brown, or Pink (Red + Blue)	Molten Globular State
Blue	Aggregated State
Black	Extensively Unfolded State without Aggregation

The three-index EPD is constructed simply by mapping each structural index to an RGB color component. Since a color in an RGB scheme is expressed as a tuple of red, green, and blue components, we have assigned SI to red, TI to green, and AI to blue. Thus, a color produced by the summation of these three color components is mapped to a specific state of the target macromolecule. Table 2.1 lists the resultant colors and their interpretation in ideal cases. In the native state of a macromolecule, TI and SI would have a value of 1 and AI would be zero because the amount of tertiary structure and second structure would be highest and there will be no aggregation. This combination confers a yellow color to the native state because a range of index value from zero to one is assigned to a color gradation from black to the full color of the index respectively. In some cases, a red color will be observed as the amount of tertiary structure decreases but the secondary structure content remains high. This may also indicate a molten globular state in which tertiary structure change occurs prior to secondary structure alteration. A brown color, however, will more frequently appear because the amount of secondary structure will probably be reduced slightly compared to the native state, although it is still relatively higher than the fractional amount of tertiary structure. At the same time, some aggregation may be observed in the molten globular state, which forms a pink color. Once the macromolecule becomes fully unfolded, the amount of both tertiary and secondary structure becomes minimal and therefore the red and green color components decrease to black. If there is no aggregation, the resulting color will be black. This conformationally altered state attains a blue color if the macromolecule extensively aggregates.

As an optional feature, the individual RGB components can be provided alongside the three-index EPD. Since it is difficult to determine the amount of an RGB component with a

given color, the explicit display of its RGB components could be helpful in understanding the interpretation of a color. The three-index EPD accompanied with its RGB component diagrams, therefore, enable facile identification of changes at each level of structure in the protein.

2.2.5 Construction of Radar (Star) Diagrams

The radar diagram is a widely used graphical representation for multivariate data.^{12,17-23} It has many similar forms and names such as star glyph, star chart, and spider chart. The major idea for this data visualization approach is to arrange multiple axes in evenly spaced angles from the same starting point to form a polar coordinate system. Multivariate data, represented as n -dimensional vectors, is plotted on the n -axes and connected to each other to form a (filled) polygon.

The radar diagram used in this work is composed of multiple radar charts arranged in two dimensional Cartesian coordinates of environmental stress conditions. Each radar chart represents physical data measured at the given stress conditions. For simplicity, all polar axes in a radar plot are adjusted to a display value between zero and one in which zero is mapped to the axial starting point and one to the outer rim. Therefore, stability data should first be converted to the corresponding structural indices using equations (1) and (2). In each environmental condition (e.g. solution pH and temperature), the values in the associated data indices are mapped to points in the polar coordinates of the radar chart. The points are then connected to each other. To enhance the display the relative magnitude of these data, circular grids are placed every 0.2 interval between zero and one in a radar chart. Unlike the three-index EPD described above, the radar diagram can display any number of variables. Figure 2.3A shows an example of a radar chart with eight axes (e.g., using eight different experimental readouts). The radar diagram is

known to be well suited to identifying similarity and difference in patterns. The radar diagrams in this work are also able to aid in the recognition of a range of environmental conditions in which a macromolecule shows similar structural behavior. To better serve this purpose, two guidelines are proposed:

1. Some of the normalized data (i.e., structural indices) should be inverted so that the native state of a macromolecule will be displayed as a dot (or smallest area). The magnitude of signal changes will thus represent the amount of deviation from a native state.
2. Experimental methods should be grouped on the radar chart according to which structural feature is being measured, i.e. tertiary and secondary structure, aggregation, etc.

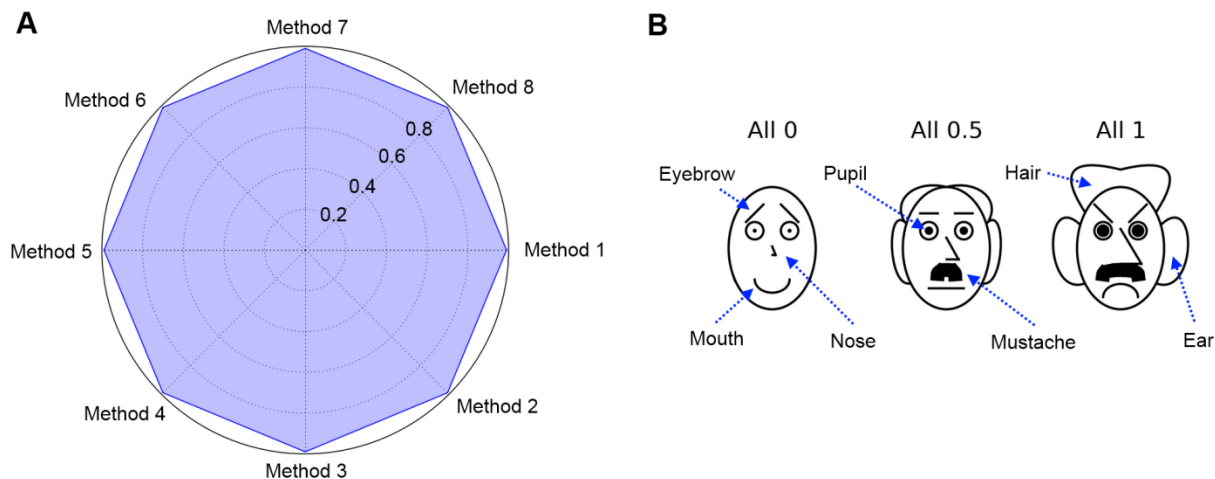


Figure 2.3 (A) Example of a radar chart with eight variables (experimental methods). All variables are set to a value of 1 with intervals of 0.2 (see text). (B) Chernoff face diagrams with seven variables. Each variable can vary from zero to one. Three faces are constructed with all variables set at 0, 0.5 and 1.

With regard to the first point, one of the most critical pieces of information explored in a protein characterization study is the environmental conditions under which a macromolecule starts to be structurally altered. If the native state is described as a dot, small transitions from the native state become more easily detectable. More importantly, the nature of the experimental method which detects a transition becomes readily evident. Secondly, the order of presentation of experiments in the radar plot is an important factor in intuitive pattern recognition. If there are multiple experiments that measure similar properties, they should be grouped together instead of undergoing random placement. For example, three tertiary structure sensitive measurements (e.g. intrinsic fluorescence peak position shift, UV absorbance second derivative peak positions, and near UV CD) can be placed in the positions of Method 1 to 3 in Figure 2.3A , two secondary structure measurements (e.g., CD and FTIR) in the positions of Method 4 to 5, and three aggregation measurements (e.g. static and dynamic light scattering, and optical density) in the positions of Method 6 to 8. Such a grouping of data should increase the visual interpretability by assigning the type of measurement to an appropriate angular placement in the radar diagram.

2.2.6 Construction of Chernoff Face Diagrams

H. Chernoff invented the Chernoff Face diagram as a multivariate data visualization technique.¹³⁻¹⁵ The key idea of this approach is to utilize the sensitive ability of human face recognition as an efficient tool to read and partition multivariate data represented as human faces. There is no restriction on how to map multivariate data to human facial features such as the shape, size and location of eyebrows, eyes, nose, mouth, ear, hair and face. Although the number of variables can be quite large, seven key facial features were implemented here for exploratory purposes as presented in Figure 2.3B.

The Chernoff Face diagram has the same format as the radar diagram except each plot at any coordinate is a Chernoff face instead of a radar chart. Since each facial feature is defined to have a parameter value ranging from zero to one, all physical data should be converted to the corresponding structural indices using equations (1) and (2). In each stress condition (e.g. pH and temperature), the values in the associated data indices are mapped to display associated facial features. In general, three or four facial variables are adequate to represent macromolecular conformational stability data as shown later in this work. The native state of a macromolecule is assigned to a smiling face for better recognition (along with a short nose combined with no hair or ears; see Figure 2.3B).

2.2.7 Clustering Analysis of Phases

One of the objectives of “stress/response” diagrams is to better understand macromolecular behaviors induced by environmental stresses. Macromolecular structural behavior observed by multiple experimental techniques is displayed in a two-dimensional environmental stress grid by color in the three-index EPD, an equiangular polygon in the radar diagram, and a human face in the Chernoff face diagram. All visualization techniques emphasize detection of similarity and outliers, which is suited to identifying boundaries where macromolecular structure initiates alteration. In many cases, however, changes in color, the shape of polygons and human facial characteristics may be too subtle for human visual perception to recognize a distinct boundary. Computational clustering algorithms can be helpful in determining such boundaries. A number of clustering algorithms have been developed for various types of problems.^{27,28} The development and performance evaluation of a certain clustering algorithm and its parameters are

out of the scope of this study. Rather, visualization of clustering results will be demonstrated for each of the three data visualization approaches.

The k -Means clustering algorithm^{27,28} was chosen for this study because of its popularity. The k -Means clustering algorithm is a widely used method to calculate k central points (or centroids) in which all samples belong to each cluster whose mean is the calculated centroid. The number of clusters k must be postulated, and therefore, various values of k are tried and selected after evaluation. The result is not always optimal because the algorithm tries to converge rapidly to a local optimum from random initial centroid locations. Therefore, the results can be manually correctable based on the interpretation of raw data. In addition, the same observation values under different environmental conditions are usually chosen as the same cluster by the algorithm, although actual interpretation may be different. In this case, clustering of indices rather than raw data might produce better results.

Once the clusters are obtained, they can be displayed on the three-index EPD, the radar diagram, and the Chernoff face diagram as line boundaries. For the radar and Chernoff face diagrams, a cluster can be represented as a single radar chart or a Chernoff face which displays averaged values of all images in the cluster. This *clustered* version of a radar chart or Chernoff face diagram provides a more compact view in summarizing the characteristics of each cluster. We generally find the clustered radar or Chernoff face diagrams provide the best summary of the data, as shown below.

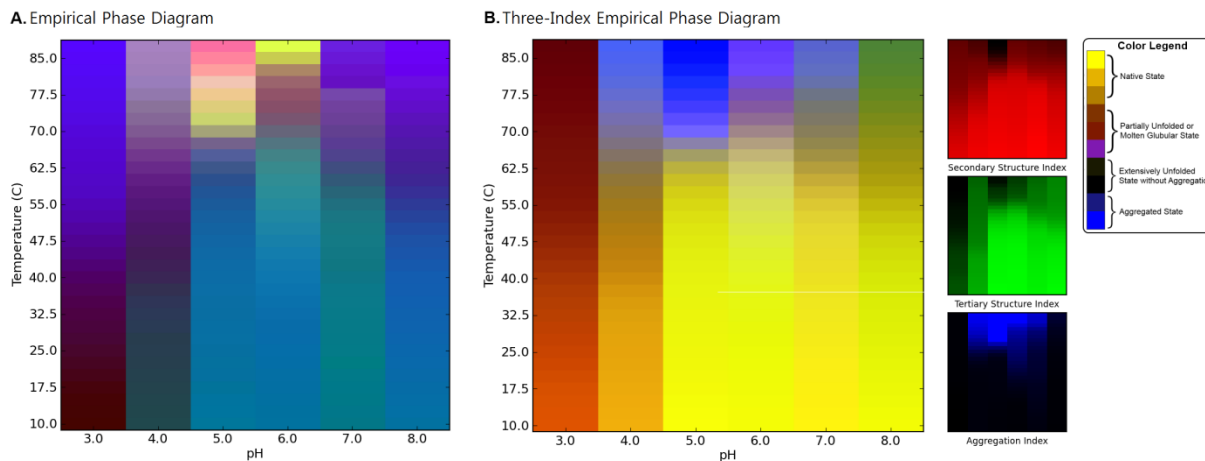


Figure 2.4 EPDs of the conformational stability of BSA as a function of pH and temperature. (A) Original EPD created using biophysical data in Figure 2.1A-C. (B) Three-Index EPD created using structural indices as shown in Figure 2.1D-F.

2.3 Results

Figure 2.4 compares two EPD data visualization approaches derived from the physical data of Figure 2.1 (effect of temperature and pH on BSA as measured by CD, fluorescence and light scattering; see methods section). Figure 2.4A shows an EPD constructed using the previously published data¹⁰ by the current EPD method. The newly developed three-index EPD displayed in Figure 2.4B, which is constructed using three structural indices calculated from the same experimental data as described in the Methods section. The two EPDs show very similar transitions in color patterns which indicate the structural changes of BSA in response to pH and temperature. In both cases, a region of pH and temperature with similar color represents related structural states defined by the specific experimental data. For example, the region from pH 5 to 8 below 60°C colored blue in Figure 2.4A and yellow in Figure 2.4B represents the native state of BSA. The native structure is altered due to lower pH or higher temperature as depicted by the

purple phase in Figure 2.4A or brown phase in Figure 2.4B. The aggregation of the protein is seen as the yellow/red region in Figure 2.4A or the blue region in Figure 2.4B from pH 4 to 6 above 70°C. A structural interpretation of the original EPD could not be achieved without direct reference to the underlying experimental data. In contrast, the three-index EPD depicts the type of structural change the protein undergoes by simple reference to color. For example, yellow or shades of yellow always reflect the native state of the protein, with no aggregation or change in the secondary and tertiary structure of the protein. The blue color represents an aggregated state. Brown and green colors define a structurally altered state with minimal or no aggregation. In short, the three-index EPD, when properly constructed, adds meaning to the colors in the EPD. The individual RGB components are also provided on the right side of Figure 2.4B for a better understanding of the three-index EPD. Since each RGB component represents its associated structural indices, the additional color plots on the right hand side in Figure 2.4B visually clarify the source of the transitions seen in the three-index EPD, in this case in terms of data from each analytical instrument (CD, fluorescence and static light scattering, respectively).

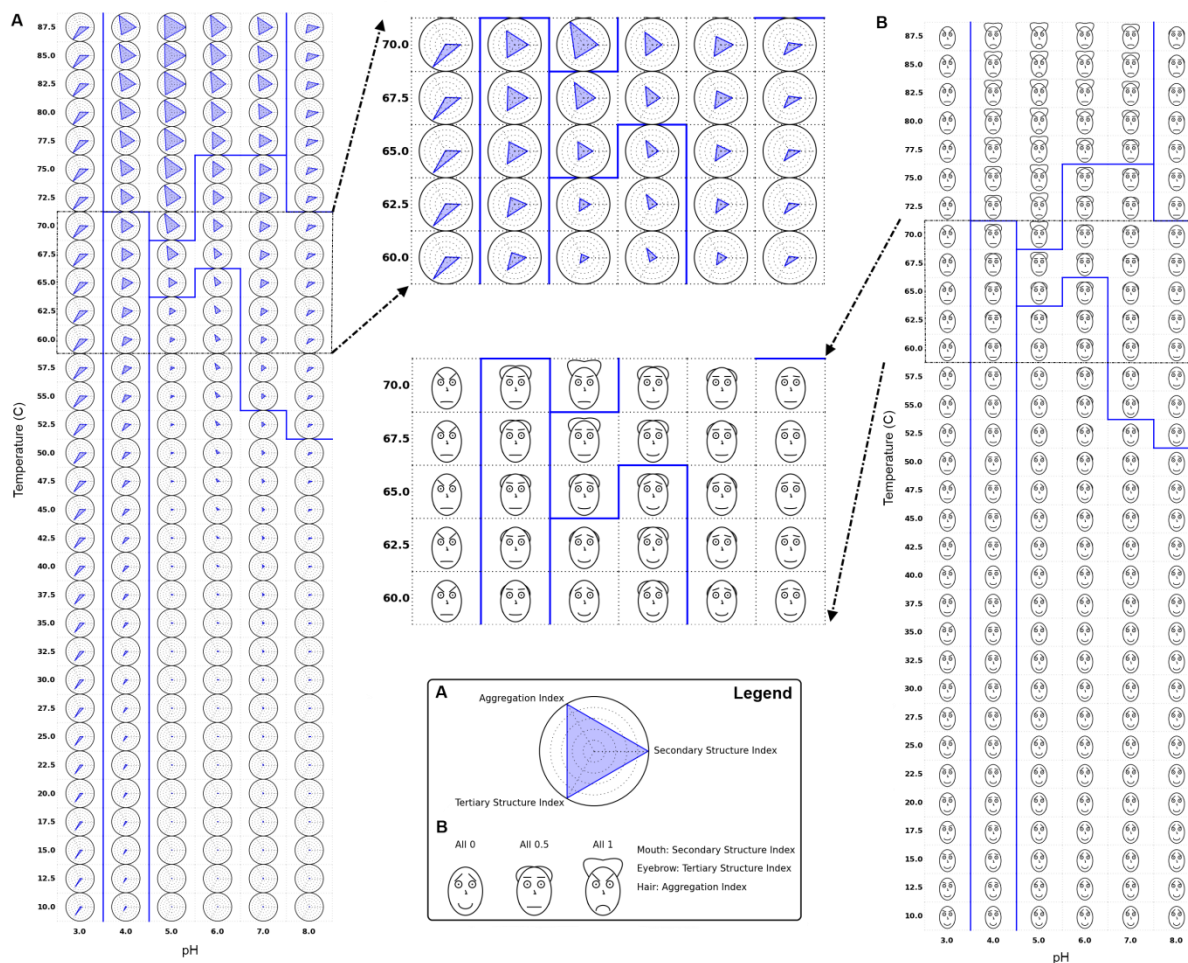


Figure 2.5 (A) Radar chart and (B) Chernoff face diagram of the conformational stability of BSA as a function of pH and temperature using data from the three structural indices from Figure 2.1D-F. Secondary and tertiary structure indices are inverted for both cases to represent the native state as (A) a dot and (B) a smiling face. Solid blue lines indicate an example of clustering results (see text).

The radar and Chernoff face diagrams of the same BSA stability data set (Figure 2.1; see methods section) are shown in Figure 2.5. In both cases, the secondary and tertiary structure indices are inverted to represent the native state either as a dot or a smiling face. These settings are used to more easily perceive small deviations from the native state. It is seen that all three of the data visualization approaches shown in Figure 2.4 and Figure 2.5 display the same structural transitions of BSA as a function of pH and temperature in terms of changes in either colors, shape of polygons, and facial features.

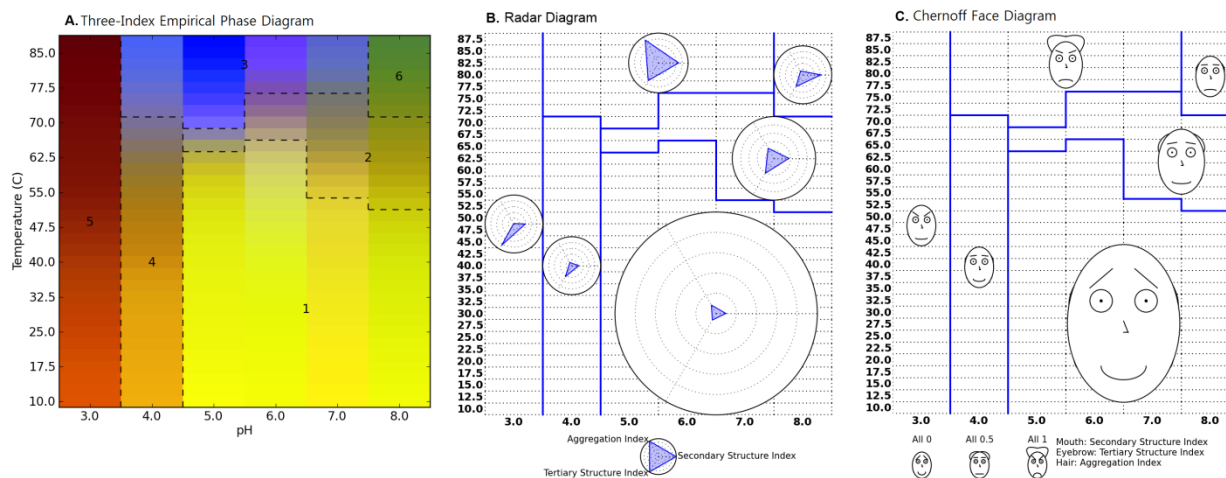


Figure 2.6 (A) Three-Index EPD, (B) clustered Radar Chart, and (C) clustered Chernoff Face Diagram for the conformational stability of BSA (using data from three structural indices from Figure 2.1D-E. Each cluster is represented as a radar chart or Chernoff face diagram which averages the structural indices of all images that belong to the cluster. Six structural phases were observed empirically; 1: native state, 2: molten globular state, 3: aggregated, 4-6: structurally altered states due to low pH and/or high temperature without aggregation.

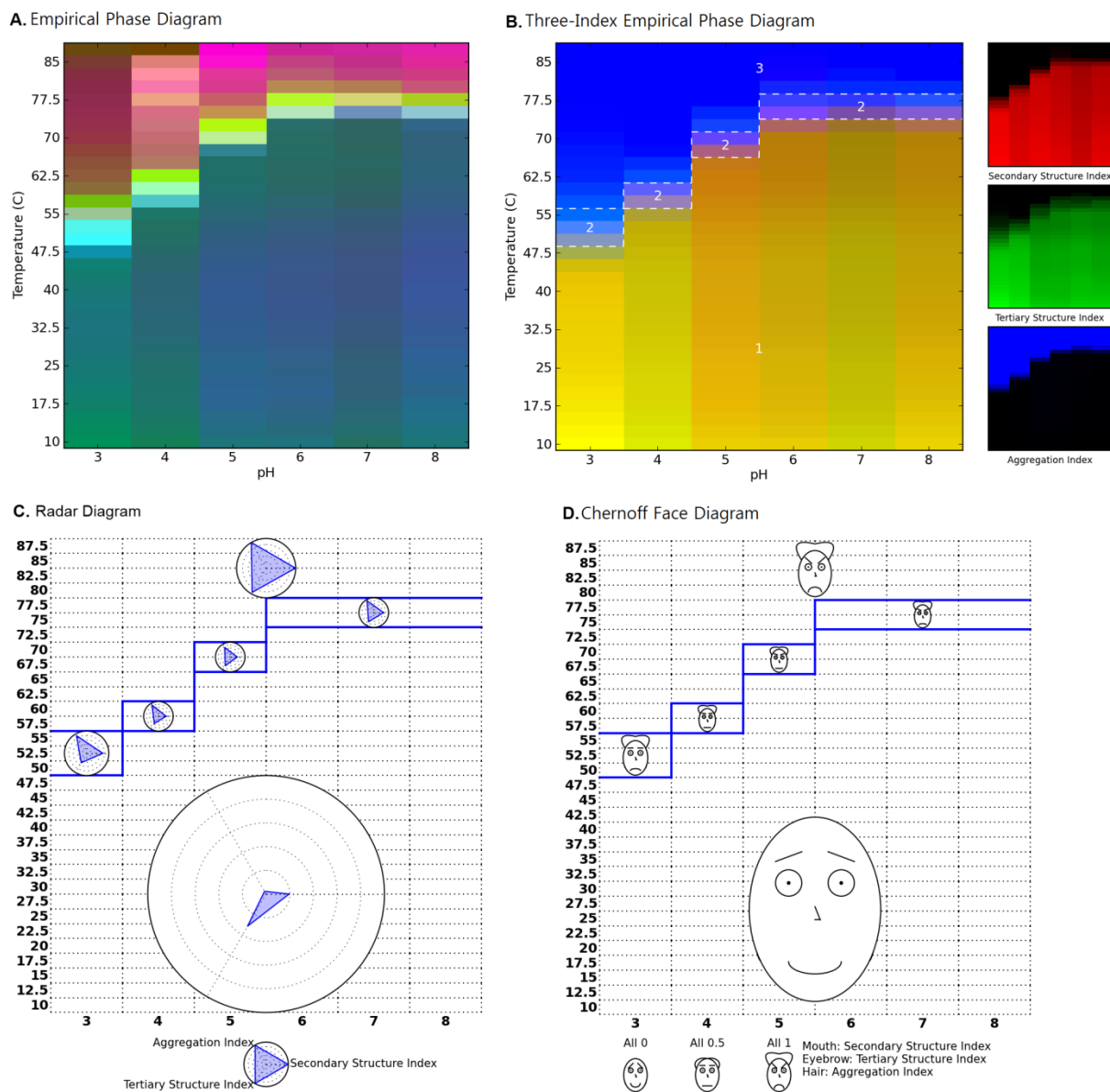


Figure 2.7 (A) EPD, (B) Three-Index EPD, (C) Radar Chart and (D) Chernoff Face Diagram for the protein antigen SP1732 as a function of temperature and pH. Figure 2.7A is created using biophysical data in Figures 2.2A-C, while Figure 2.7B-D are created using structural indices shown in Figure 2.2G-I. Three structural regions are observed; 1: native state, 2: molten globular state, and 3: aggregated state.

The solid blue lines in Figure 2.5 illustrate an example of clustering results. Clustering was performed using a k -Means clustering algorithm with $k=6$ and manually corrected afterwards based on interpretation of raw data. After several trials of different values of k , the k number was selected in which its result most closely matches our interpretation of raw data. One of the clustering results is displayed for all three diagrams in Figure 2.5 and Figure 2.6. In addition, Figure 2.6B and C show *clustered* radar and Chernoff face diagrams in which each cluster is represented by a single radar chart and Chernoff face. This single iconic plot is generated by an average value of the data inside a cluster and exhibits the characteristic of the cluster.

The original EPD and three different data visualization diagrams for the protein antigen SP1732 are shown in Figure 2.7A-D. The data from SP1732 provide a good example of a protein that clearly displays a molten globular state (see data in Figure 2.2 in Methods section). The EPD constructed using previously published data for SP1732²⁴ is shown in Figure 2.2A. SP1732 was found to be most stable at temperature below 45°C at low pH and below 70°C at pH 6-8.

Aggregation of SP1732 is observed in the pink region in the EPD. A molten globule state is also seen in the interface between these two “apparent” phases. It should be noted that “apparent” phases do not refer to equilibrium thermodynamic phases (i.e. no reversibility is implied), but rather simple, empirical representations of the physical behavior of the macromolecule.¹⁻³ As shown in Figure 2.7B, a small pink colored region between the native and unfolded state is also observed in the three-index EPD. The calculated structural indices used for the construction of the three-index EPD for the protein antigen SP1732 are shown in Figure 2.2G-I (see Methods section). A decrease in the CD signal is seen at pH 3 at about 40°C. This transition increases in temperature with increasing pH. At pH 6-8, this transition is observed only

at about 75°C. Similar trends are observed with intrinsic fluorescence and static light scattering data.

As discussed above, one of the limitations of the EPD is that the investigator needs to study the original data independently to better understand the origin of the colored regions. The three-index EPD remedies this shortcoming by manifesting the native state of the protein antigen SP1732 as a rich yellow color as seen in region 1 in Figure 2.7B. Aggregation along with unfolding of the protein appears blue in region 3. Molten globule states usually present as a red or a pink colored phase as seen in region 2. These indices are also mapped on to the radar and Chernoff face diagrams in Figure 2.7C and D, respectively. In the radar diagram, the native state of the protein antigen SP1732 is represented as blue triangles occupying minimal area. Transitions are seen as the blue triangle increasing its area at the corresponding angle. This is seen in region 3 of Figure 2.7C at high temperatures at all pH values, where we see an equilateral triangle that extends to the circumference of the circle indicating a total loss of structure and aggregation. Molten globule behavior appears as states with increases in aggregation and loss of tertiary structure as seen in region 2 of Figure 2.7C. In the Chernoff diagram, the indices for SP1732 are mapped to the mouth, eyebrow and hair of the Chernoff face. The native state appears as a “happy and bald” face in region 1 of Figure 2.7D, with the mouth, eyebrows, and the hair representing the secondary structure index, the tertiary structure index and the aggregation index, respectively. The molten globule state shows the presence of aggregates in the form of hair, and intermediate secondary and tertiary structures in the form of “poker faces” and horizontal mouth and eyebrows (Figure 2.7D).

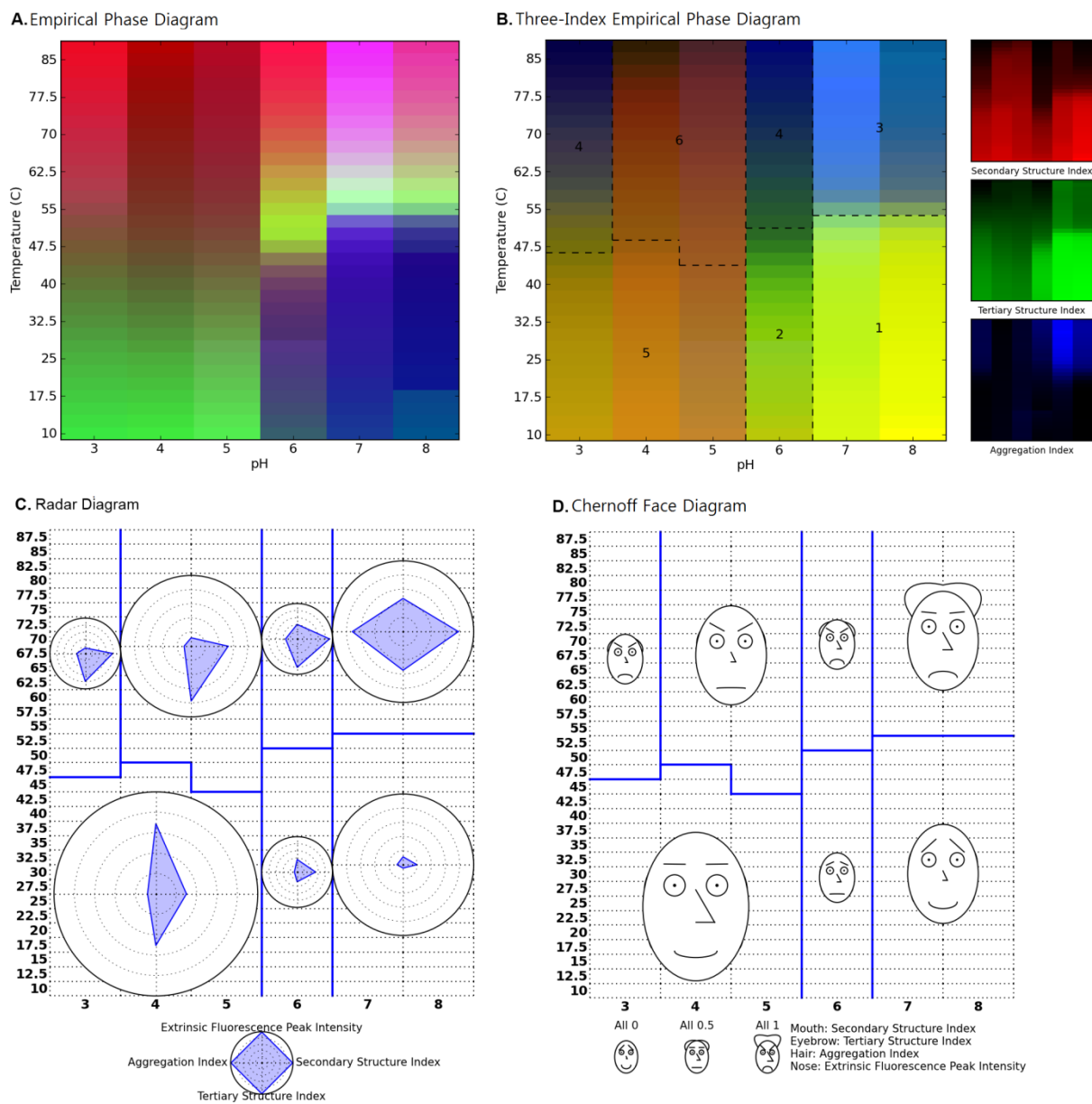


Figure 2.8 (A) EPD, (B) Three-Index EPD, (C) Radar Chart and (D) Chernoff Face Diagram for the protein antigen HAC1 as a function of temperature and pH. Figure 2.8A is created using biophysical data in Supplementary Figures S1A-D, while Figures 2.8B-D are created using structural indices shown in Supplementary Figures S1E-H. Six structural phases are observed; 1: native state, 3: aggregated state, 2, 4-6: structurally altered state due to low pH and/or high temperature without aggregation.

The original EPD and the three different data visualization diagrams for protein antigen HAC1 are shown in Figure 2.8. The previously published data²⁵ and newly constructed structural indices are included as Supplementary Figure S1. As shown in the EPD in Figure 8A, the native state of the HAC1 protein is seen at pH values of 7 and 8 at temperatures lower than 50°C. The protein is conformationally altered at pH 6 at temperatures below 40°C as reflected in the purple phase in Figure 2.8A. As the pH drops to 5 and below, the structure again changes and is represented by the green phase. Increasing temperature in all apparent phases is accompanied by unfolding or aggregation of the protein. As shown in Figure 2.8B, the pH-dependent conformational changes for HAC1 are more pronounced in the three-index EPD, in which regions 1 and 2 are more clearly differentiated. In addition, Region 1, which was identified as the native state, is again represented by a yellow color. The conformationally altered protein at pH 6 is shown by the appearance of green hinting at possible changes in the protein's secondary structure. The brown color seen at pH 5 and below indicates the presence of native secondary structure, but low levels of tertiary structure (Figure 2.8B).

One potential limitation of the three-index EPD is an inability to include data from more than three experiments. For example, no major change in color was observed at pH 4 and 5 with changes in temperature (Figure 2.8B). The changes observed in the EPD arise from the ANS data (Figure 2.8A and Supplementary Figure S1). The three-index EPD presumably does not show the changes because of the absence of the ANS data. This limitation can be offset by the construction of radar and Chernoff face diagrams. In these diagrams, the ANS data are mapped on the fourth arm in the radar diagram (Figure 2.8C) or on to the size of the nose in the face diagrams (Figure 2.8D). The native state of the protein antigen HAC1 is again shown by a

quadrilateral of minimum area in the radar diagram. In the Chernoff face diagram, this native state is shown with a happy face, a bald scalp and a small nose. As the protein conformation changes, the quadrilateral and the face changes in accordance with the changes observed in the data (e.g., angle changes in the radar diagram in Figure 2.8C and the presence of hair at high temperatures at pH 7 and 8 in Figure 2.8D reflecting aggregation of the protein).

Data obtained for several other proteins including aldolase¹⁰, chymotrypsin¹⁰ and SP1650²⁴ were also utilized in the construction of the original EPDs as well as the new data visualization methods (three-index EPDs, the radar diagrams and the Chernoff face diagrams). The previously obtained data and the newly constructed structural indices for aldolase, chymotrypsin, and SP1650 are shown in Supplementary Figures S2, S4 and S6, respectively. The various data visualization diagrams constructed for each of these proteins are shown in the Supplementary Figures S3, S5, and S7. Each of these versions of the four diagrams (for Aldolase, chymotrypsin and SP1650) demonstrates the same trends discussed above (for BSA, SP1732 and HAC1).

2.4 Discussion

Three new data visualization methods are presented in this work in the context of evaluating six different proteins in terms of conformational stability as a function of pH and temperature. The three-index EPD approach describes three aspects of macromolecular structure with a pre-defined color scheme. Given a range of environmental perturbations, the amount of tertiary and secondary structure and state of aggregation were measured by intrinsic fluorescence peak position, circular dichroism spectroscopy and static light scattering, respectively. After the

measurement, these data were interpreted and converted to structural indices representing the relative amount of structural change, which were then mapped to specific colors. As seen in several examples in this study, the color yellow represents the native state of a macromolecule and the color blue an aggregated state. A darker color, close to black, represents a maximally altered conformational state without any notable aggregation. These colors are considered the major indicators, while other colors (e.g., such as brown and green) represent partially altered states depending on the differentially reduced color levels of tertiary and secondary structure. For example, the colors red or pink are interpreted as indicators of molten globular states for the protein antigen SP1732 as seen on Figure 2.7B.

The assignment of color to the degree of structural change in the three-index EPD is achieved by displaying only three experimental data sets as measures of secondary, tertiary and quaternary structures. Currently, more than three types of experimental data can be summarized in the other approaches including the original EPD, the radar charts, and the Chernoff face diagram. In Figure 2.8, the difference among visualization techniques when employing four different data visualization methods to evaluate the same biophysical data sets are clearly shown for the protein antigen HAC1. In this case, extrinsic ANS fluorescence spectroscopy data (see Supplementary Figure S1) was not included in the three-index EPD because it provides a measure of the amount of dye binding to either the apolar surface(s) of a macromolecule or a positively charged binding site, neither of which can be simply related to macromolecular structure. The ANS data does, however, clearly highlight the structural transition regions from pH 6 to 8 around 55°C and distinguishes the region at pH 4-5 above 50°C in the EPD from the

same regions in the three-index EPD. A method to calculate structural indices from more experimental data is thus still needed and is under investigation.

The introduction of the structural index not only enhances the visualization and interpretability of the experimental data, but it also performs an initial analysis of the data. The curated index provides a sigmoidal melting curve that preserves the initial positions and slope changes of the original data. The alignment of changing directions and proper cut-offs permit the comparison of signals within the area of interest in a more straightforward manner. One additional analysis would be the subtraction of a tertiary structure index from a secondary structure index to observe a peak indicating the existence of molten globular states (data not shown). This peak is also seen as the color pink (or red) in the three-index EPDs. Another benefit of the structural indices is for use in computational methods, which includes clustering analyses. Clustering analyses may provide inaccurate partitioning, however, if the same value has different meaning. For instance, a decrease in the light scattering at higher temperature is often observed after it reaches a peak intensity value (e.g., see Figure 2.2C for data with the protein SP1732). The same level of light scattering data before and after the peak intensity may not have the same interpretation, but the clustering algorithms would identify the two regions as the same state. Thus, instead of using raw data, the use of an index for computational methods can maximize the accuracy of the results.

The two other data visualization approaches presented in this work, the radar chart and the Chernoff face diagram, have similar properties to one another. A specific shape or image of an icon (i.e., an equiangular polygon or a human face) is designed to reflect the characteristics of the underlying data and distinguish differences as a function of solution variables. Each type and

magnitude of selected experimental data is explicitly expressed in these diagrams. This result is in contrast to original EPDs which cannot identify the origin of the experimental technique and structural features which cause color changes. We currently consider these diagrams alternatives to color-based EPDs because they have the significant advantage that they can accommodate more different kinds of data. It is, however, difficult to read an exact value or to detect subtle changes in values with these two approaches, especially for Chernoff face diagrams. Both of the diagrams require more space to represent the same number of environmental stresses than EPDs; therefore, an individual shape or a face becomes too small to be easily recognized if the entire diagram is presented. Alternatively, the clustered versions of diagrams may become a useful tool to symbolize selected characteristics of macromolecules.

The utility of EPDs to summarize the effect of environmental stresses on protein conformational stability has recently been enhanced by the availability of multimodal spectrophotometers with multiple sample holders that can simultaneously or sequentially measure CD, fluorescence and UV absorption spectra as well as light scattering and turbidity as a function of temperature in an automated mode with a single sample.^{10,29} Thus, the multiple data visualization diagrams of the type described in this work can be obtained rapidly, within a single day is possible, with a minimal amount of protein and effort.

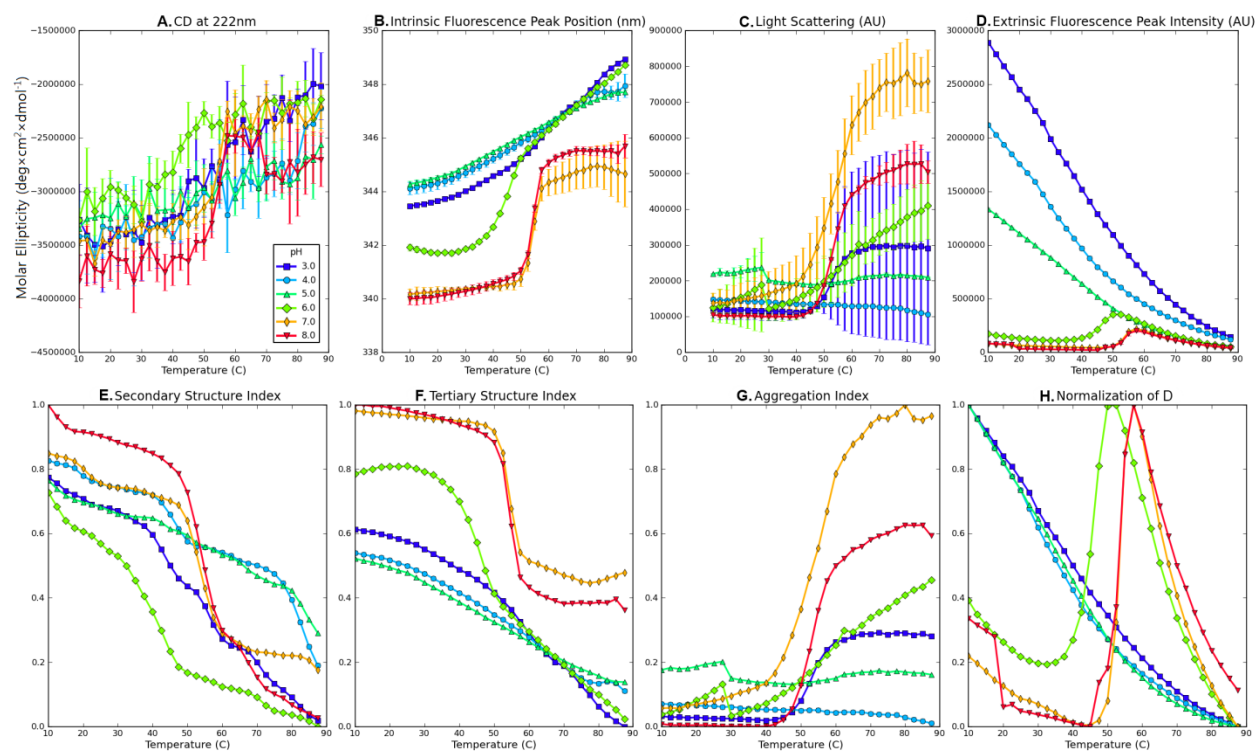
2.5 References

1. Kueltzo LA, Ersoy B, Ralston JP, Middaugh CR (2003) Derivative absorbance spectroscopy and protein phase diagrams as tools for comprehensive protein characterization: a bGCSF case study. *J Pharm Sci* 92:1805-20.
2. Maddux NR, Joshi SB, Volkin DB, Ralston JP, Middaugh CR (2011) Multidimensional methods for the formulation of biopharmaceuticals and vaccines. *J Pharm Sci* 100:4171-4197.
3. Alsenaidy MA, Wang T, Kim JH, Joshi SB, Lee J, Blaber M, Volkin DB, Middaugh CR (2012) An empirical phase diagram approach to investigate conformational stability of “second-generation” functional mutants of acidic fibroblast growth factor-1. *Protein Sci* 21:418-32.
4. Fan HH, Ralston J, Dibiasse M, Faulkner E, Middaugh CR (2005) Solution behavior of IFN-beta-1a: An empirical phase diagram based approach. *J Pharm Sci* 94:1893-1911.
5. Fan HH, Li HN, Zhang MY, Middaugh CR (2007) Effects of solutes on empirical phase diagrams of human fibroblast growth factor 1. *J Pharm Sci* 96:1490-1503.
6. Brandau DT, Joshi SB, Smalter AM, Kim S, Steadman B, Middaugh CR (2007) Stability of the Clostridium botulinum type a neurotoxin complex: An empirical phase diagram based approach. *Mol Pharm* 4:571-582.
7. Nonoyama A, Laurence JS, Garriques L, Qi H, Le T, Middaugh CR (2008) A Biophysical Characterization of the Peptide Drug Pramlintide (AC137) Using Empirical Phase Diagrams. *J Pharm Sci* 97:2552-2567.
8. Ramsey JD, Gill ML, Kamerzell TJ, Price ES, Joshi SB, Bishop SM, Oliver CN, Middaugh CR (2009) Using Empirical Phase Diagrams to Understand the Role of Intramolecular Dynamics in Immunoglobulin G Stability. *J Pharm Sci* 98:2432-2447.
9. Joshi SB, Bhambhani A, Zeng Y, Middaugh CR An Empirical Phase Diagram/High Throughput Screening Approach to the Characterization and Formulation of Biopharmaceuticals. In: Jameel F, Hershenson S, eds. (2010) *Formulation and Process Development Strategies for Manufacturing Biopharmaceuticals*. Wiley & Sons Publication, pp. 173-204.
10. Hu L, Olsen C, Maddux NR, Joshi SB, Volkin DB, Middaugh CR (2011) Investigation of protein conformational stability employing a multimodal spectrometer. *Anal Chem* 83:9399-405.
11. Deeb SS (2005) The molecular basis of variation in human color vision. *Clin Genet* 67:369-77.
12. Chambers JM, Cleveland WS, Kleiner B, Tukey PA (1983) *Graphical Methods for Data Analysis*. Wadsworth:158-162.

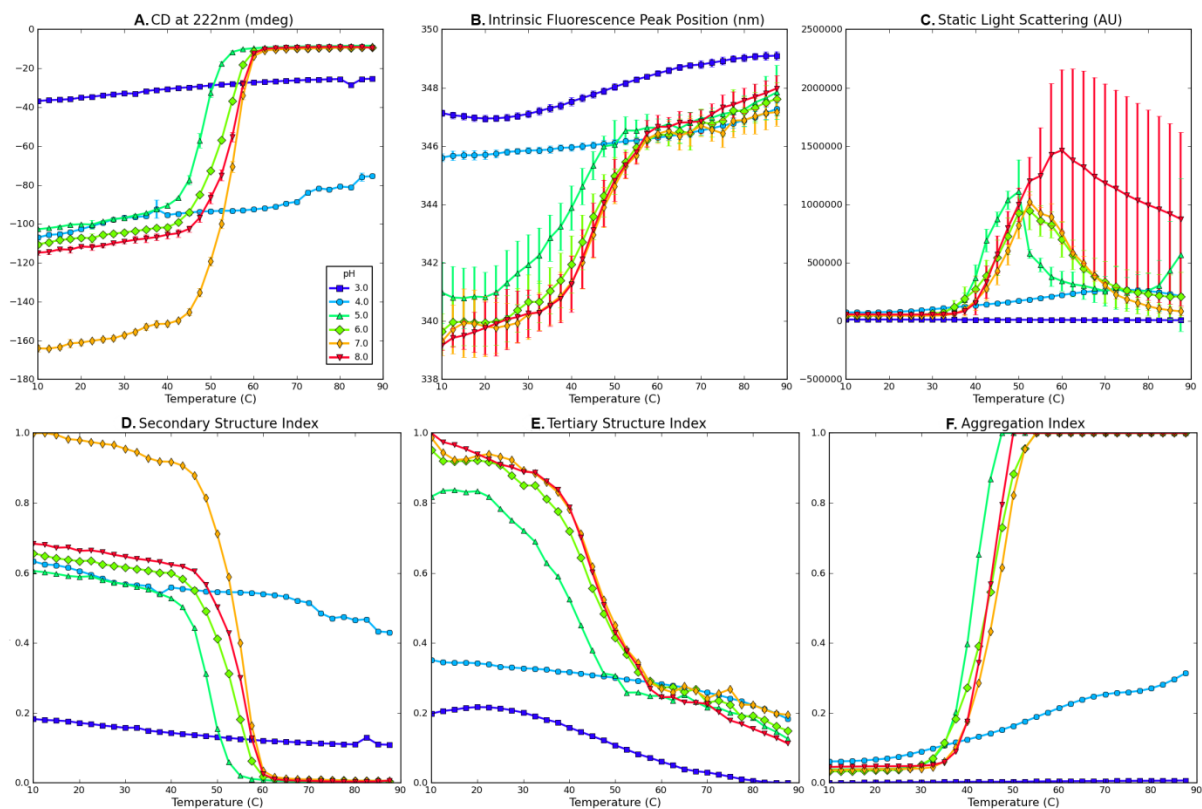
13. Chernoff H (1973) Using faces to represent points in k-dimensional space graphically. *J Am Stat Assoc* 68:361-8.
14. Chernoff H, Rizvi MH (1975) Error of Random Permutations of Features in Representing Multivariate Data by Faces. *J Am Stat Assoc* 70:548-554.
15. Everitt BS, Nicholls P (1975) Visual Techniques for Representing Multivariate Data. *J Roy Stat Soc* 24:37-49.
16. Keim DA (2002) Information visualization and visual data mining. *IEEE T Vis Comput Gr* 8:1-8.
17. Draper GM, Livnat Y, Riesenfeld RF (2009) A survey of radial methods for information visualization. *IEEE T Vis Comput Gr* 15:759-76.
18. Lee MD, Reilly RE, Butavicius MA An Empirical Evaluation of Chernoff Faces , Star Glyphs , and Spatial Visualizations for Binary Data. In: (2003) APVis '03 Proceedings of the Asia-Pacific symposium on Information visualisation. Vol. 24.
19. Gao J, Pattabhiraman P, Bai X, Tsai WT SaaS performance and scalability evaluation in clouds. In: (2011) Proceedings of 2011 IEEE 6th International Symposium on Service Oriented System (SOSE). IEEE, pp. 61-71.
20. Yokotani S, Nose T, Horiuchi Y, Matsushima A, Shimohigashi Y (2008) Radar chart deviation analysis of prion protein amino acid composition defines characteristic structural abnormalities of the N-terminal octa-peptide tandem repeat. *Protein Pept Lett* 15:949-955.
21. Funabiki Y, Kawagishi H, Uwatoko T, Yoshimura S, Murai T (2011) Development of a multi-dimensional scale for PDD and ADHD. *Res Dev Disabil* 32:995-1003.
22. Zhang Y, Hao Z, Wang R, Jin D A new method for the evaluation of gait pathology. In: (2007) Proceedings of the 1st international convention on Rehabilitation engineering & assistive technology in conjunction with 1st Tan Tock Seng Hospital Neurorehabilitation Meeting - i-CREATe '07. New York, New York, USA: ACM Press, p. 129.
23. Saary MJ (2008) Radar plots: a useful way for presenting multivariate health care data. *J Clin Epidemiol* 61:311-7.
24. Iyer V, Hu L, Liyanage MR, Esfandiary R, Reinisch C, Meinke A, Maisonneuve J, Volkin DB, Joshi SB, Middaugh CR (2012) Preformulation characterization of an aluminum salt-adjuvanted trivalent recombinant protein-based vaccine candidate against streptococcus pneumoniae. *J Pharm Sci*.
25. Iyer V, Liyanage MR, Shoji Y, Chichester JA, Jones RM, Yusibov V, Joshi SB, Middaugh CR (2012) Formulation development of a plant-derived h1n1 influenza vaccine containing purified recombinant hemagglutinin antigen. *Hum Vaccin Immunother* 8:455-466.

26. Savitzky A, Golay MJE (1964) Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Anal Chem* 36:1627-1639.
27. Xu R, Wunsch D (2005) Survey of clustering algorithms. *IEEE T Neural Networ* 16:645-78.
28. Andreopoulos B, An A, Wang X, Schroeder M (2009) A roadmap of clustering algorithms: finding a match for a biomedical application. *Brief Bioinform* 10:297-314.
29. Maddux NR, Rosen IT, Hu L, Olsen CM, Volkin DB, Middaugh CR (2012) An improved methodology for multidimensional high-throughput preformulation characterization of protein conformational stability. *J Pharm Sci* 101:2017-24.

2.6 Supplementary Figures

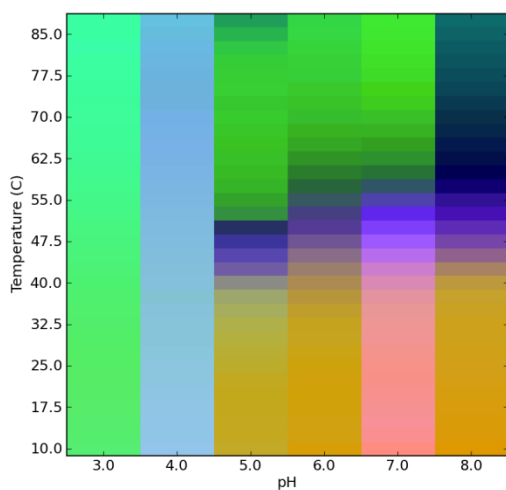


Supplementary Figure S1 Experimental data for from HAC1 measured as a function of temperature at indicated pH values (A-D) and their structural indices (E-H). (A) CD signal at 222 nm, (B) Intrinsic fluorescence peak position, (C) Static light scattering at 295 nm, (D) Extrinsic ANS fluorescence peak intensity, (E) Secondary Structure Index calculated from (A), (F) Tertiary Structure Index calculated from (B), (G) Aggregation Index calculated from (C), (H) Extrinsic ANS data normalized at each pH to emphasize transition peaks. Data in Figures S1A-C were published previously.²⁵

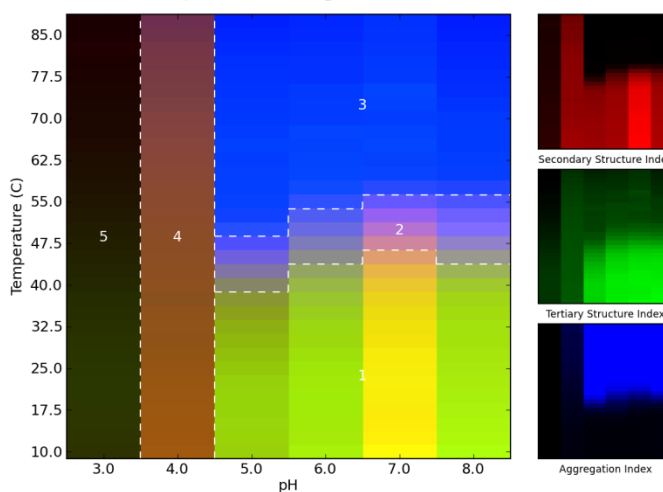


Supplementary Figure S2 Experimental data for Aldolase measured as a function of temperature at indicated pH values (A-C) and their structural indices (D-F). (A) CD signal at 222 nm, (B) Intrinsic fluorescence peak position, (C) Static light scattering at 295 nm, (D) Secondary Structure Index calculated from (A), (E) Tertiary Structure Index calculated from (B), (F) Aggregation Index calculated from (C). Data in Figures S2A-C were published previously.¹⁰

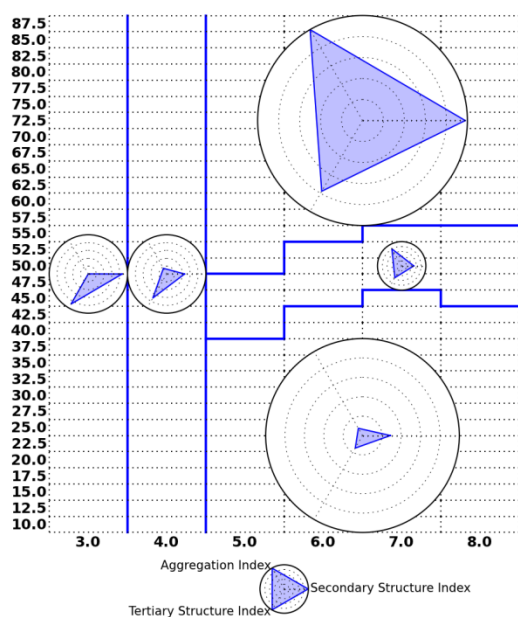
A. Empirical Phase Diagram



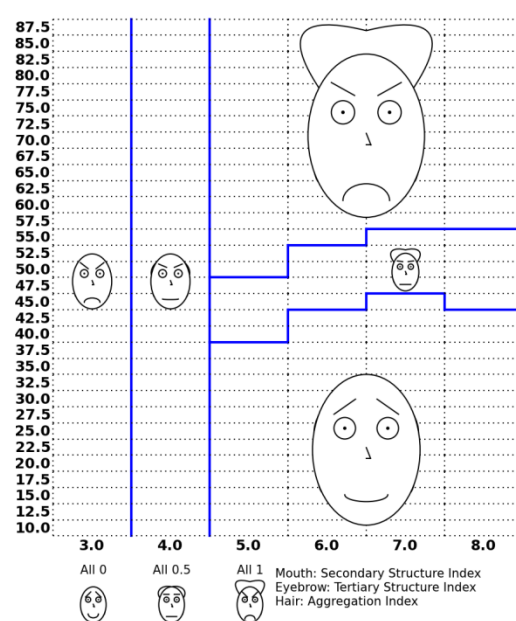
B. Three-Index Empirical Phase Diagram



C. Radar Diagram

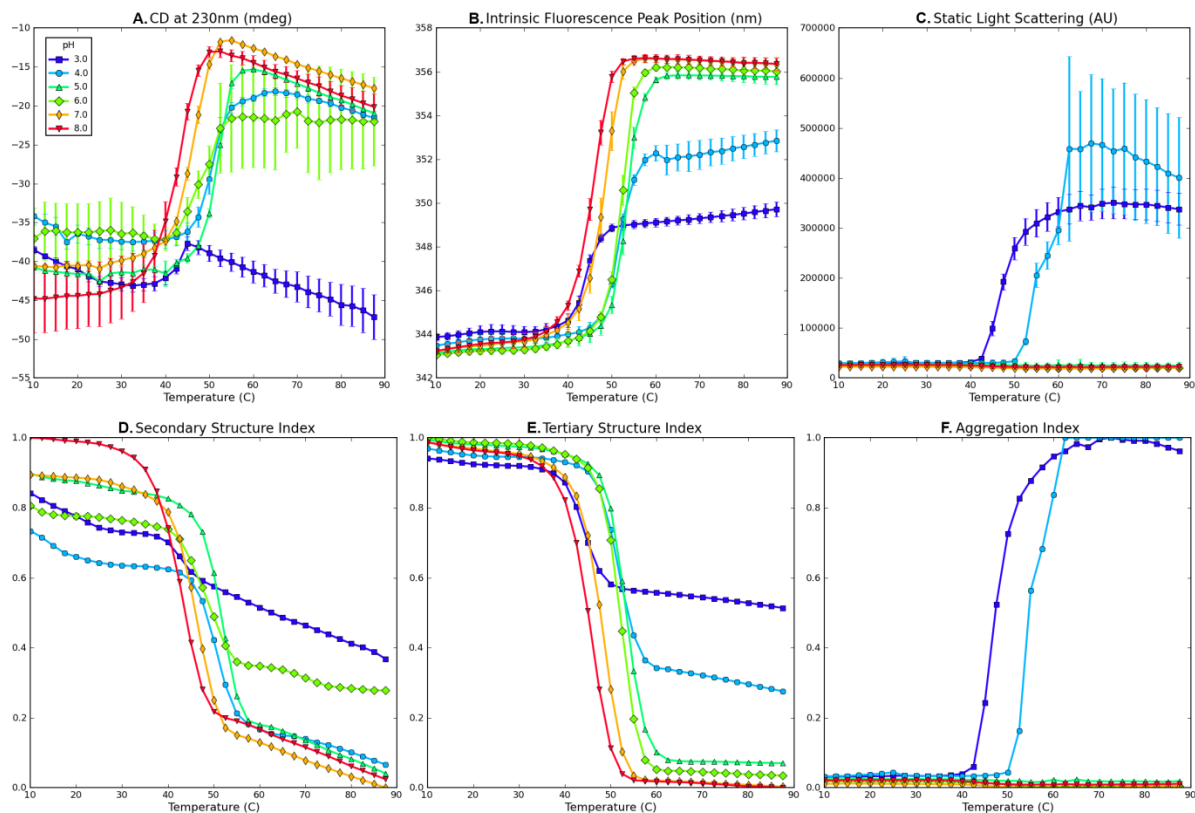


D. Chernoff Face Diagram



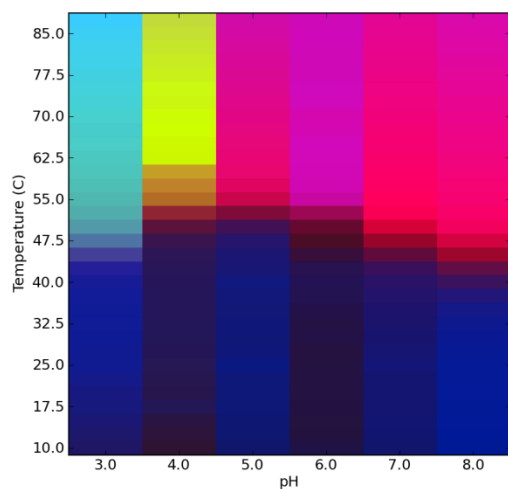
Supplementary Figure S3 (A) Empirical Phase Diagram, (B) Three-Index EPD, (C) Radar Chart and (D)

Chernoff Face Diagram for Aldolase as a function of temperature and pH. Five structural regions are observed; 1: native state, 2: molten globular state, 3: aggregated state, 4, 5: structurally altered state due to low pH without aggregation.

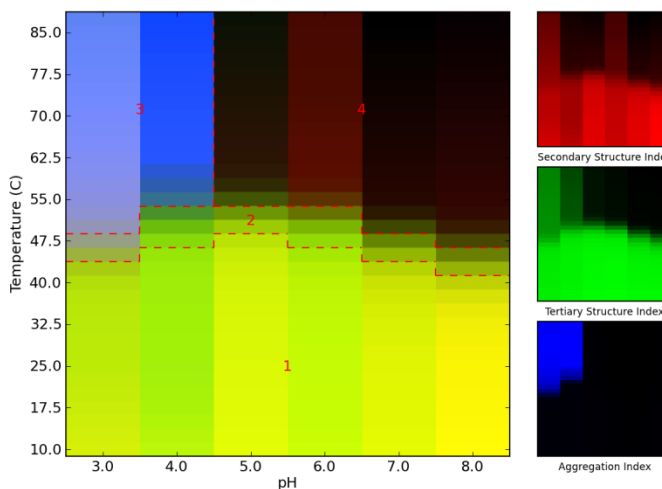


Supplementary Figure S4 Experimental data for Chymotrypsin measured as a function of temperature at indicated pH values (A-C) and their structural indices (D-F). (A) CD signal at 222 nm, (B) Intrinsic fluorescence peak position, (C) Static light scattering at 295 nm, (D) Secondary Structure Index calculated from (A), (E) Tertiary Structure Index calculated from (B), (F) Aggregation Index calculated from (C). Data in Figures S4A-C were published previously.¹⁰

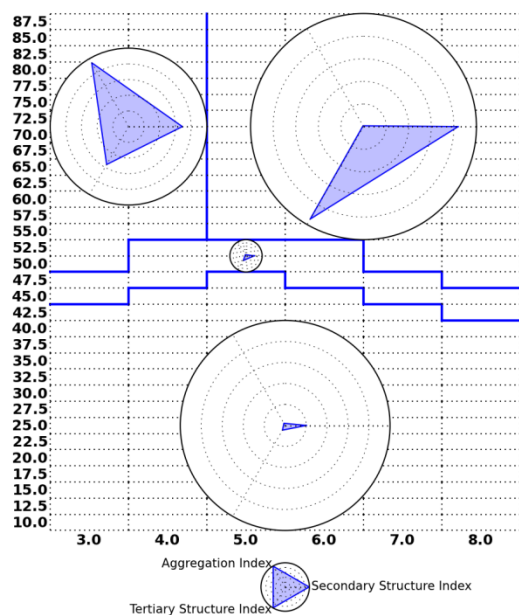
A. Empirical Phase Diagram



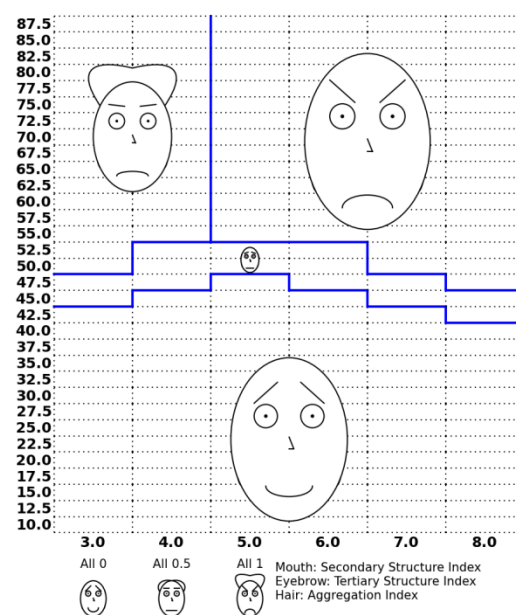
B. Three-Index Empirical Phase Diagram



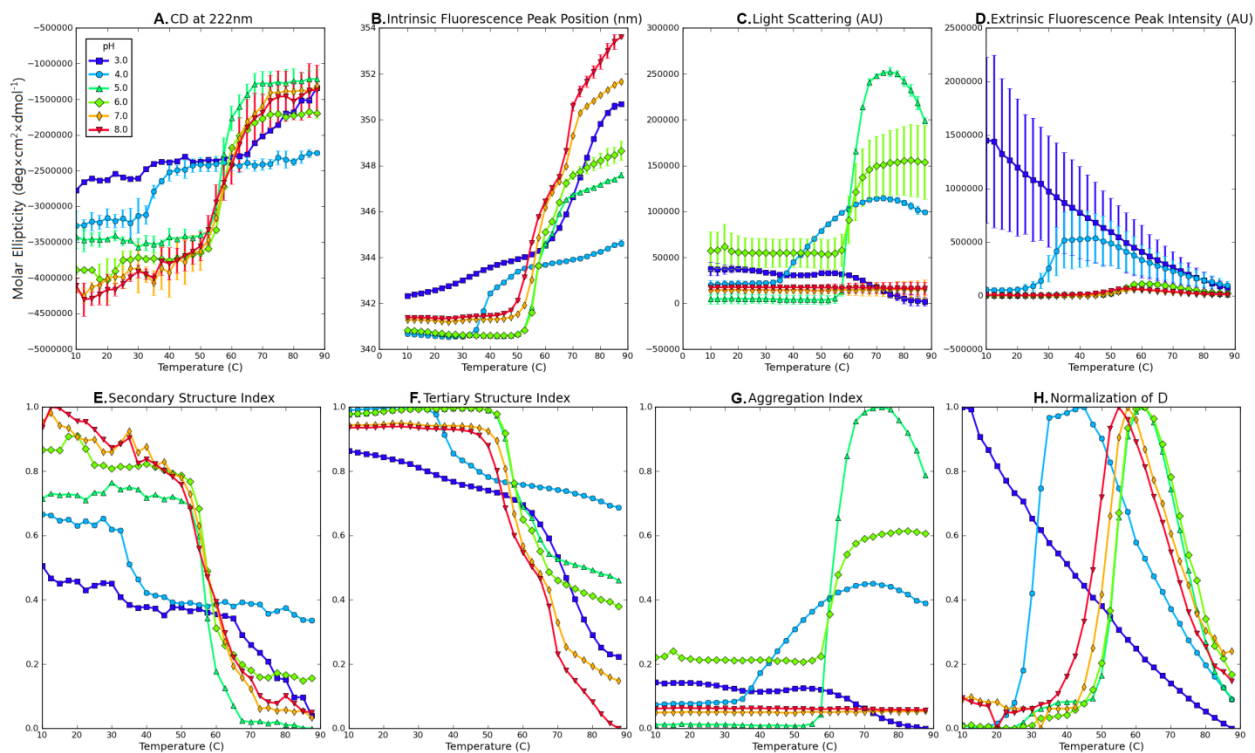
C. Radar Diagram



D. Chernoff Face Diagram

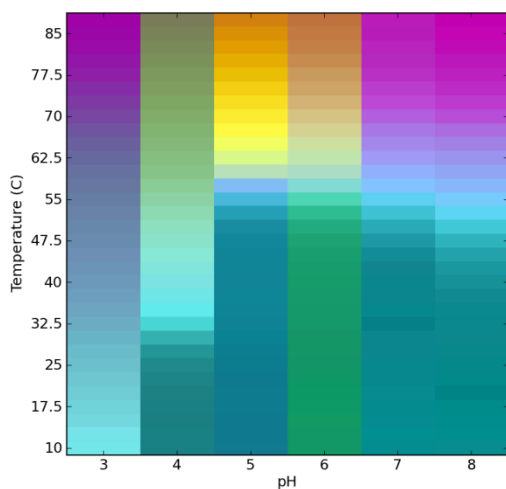


Supplementary Figure S5 (A) Empirical Phase Diagram, (B) Three-Index EPD, (C) Radar Chart and (D) Chernoff Face Diagram for Chymotrypsin as a function of temperature and pH. Four structural regions are observed; 1: native state, 2: molten globular state, 3: aggregated state, 4: structurally altered state due to high temperature without aggregation.

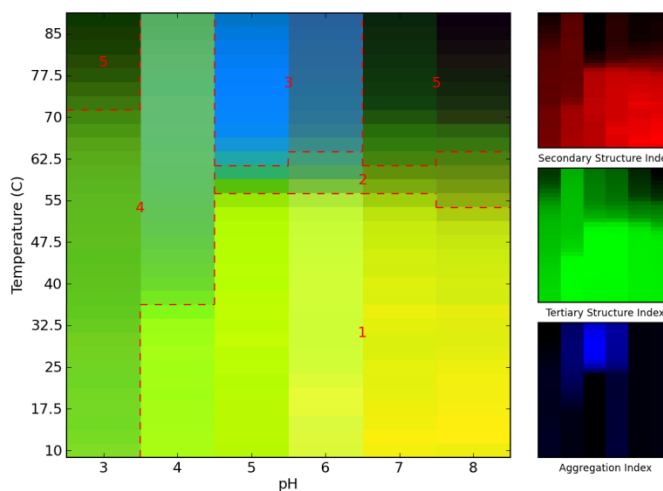


Supplementary Figure S6 Experimental data for SP1650 protein measured as a function of temperature at indicated pH values (A-D) and their structural indices (E-H). (A) CD signal at 222 nm, (B) Intrinsic fluorescence peak position, (C) Static light scattering at 295 nm, (D) Extrinsic ANS fluorescence peak intensity, (E) Secondary Structure Index calculated from (A), (F) Tertiary Structure Index calculated from (B), (G) Aggregation Index calculated from (C), (H) Extrinsic ANS data normalized at each pH to emphasize transition peaks. Data in Figures S6A-C were published previously.²⁴

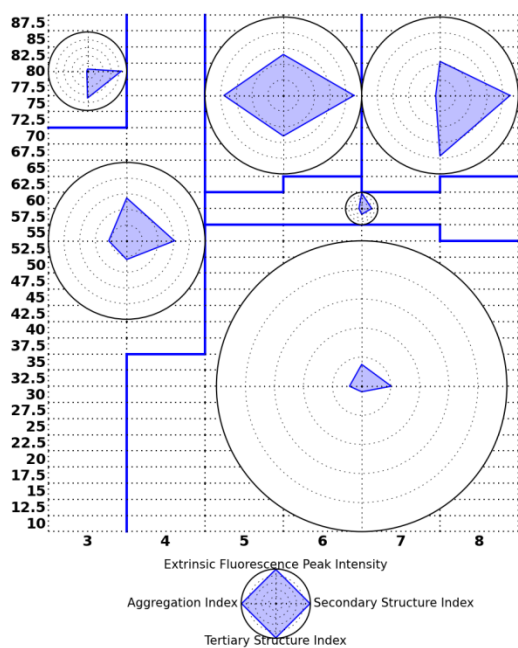
A. Empirical Phase Diagram



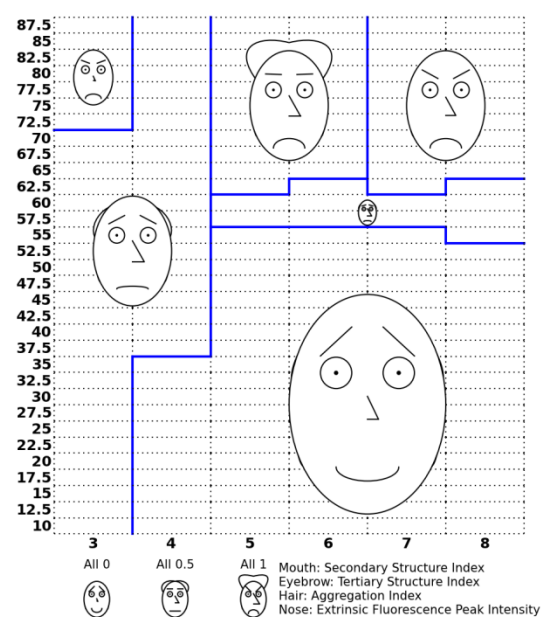
B. Three-Index Empirical Phase Diagram



C. Radar Diagram



D. Chernoff Face Diagram



Supplementary Figure S7 (A) Empirical Phase Diagram, (B) Three-Index EPD, (C) Radar Chart and (D) Chernoff Face Diagram for SP1650 protein as a function of temperature and pH. Five structural regions are observed; 1: native state, 2: molten globular state, 3: aggregated state, 4, 5: structurally altered state due to low pH without aggregation.

Chapter 3. MiddaughSuite: A Website for Biophysical Characterization of Macromolecules

3.1 Overview

Laboratory experiments used for the biophysical characterization of a macromolecule are performed in three steps: sample preparation, data acquisition, and data analysis. First, the macromolecule to be examined is prepared, usually in solution. A number of different buffers may be prepared to monitor their effects on the structure of the macromolecule. Second, various experiment techniques are used to measure structural changes in the macromolecule.

Environmental conditions such as temperature and pH may be varied during the experiment.

Finally, the data obtained are processed to visualize any structural changes of the macromolecule in response to various environmental conditions.

Each step takes a significant amount of time but the first two steps often little room for improvement. Sample preparation primarily involves human labor and the efficiency of data acquisition is totally dependent on the performance of an instrument. It is generally difficult to develop a new instrument with enhanced performance. The third step, however, can be much improved with the help of dedicated software. As dedicated software for data analysis of biophysical characterization data, MiddaughSuite was designed and implemented in this work.

3.2 Users' Requirements

MiddaughSuite, dedicated software for analysis of biophysical data, needs to fulfill the following requirements. The software should be able to easily handle data from various instruments, quickly analyze data using multiple mathematical functions, and visualize data in intuitive forms of graphs and diagrams. Furthermore, it would be desirable for the software to manage all experimental data produced in a laboratory because the data will be uploaded to the software for analysis. It should include a feature to encourage data sharing among researchers.

3.2.1 Quick and Convenient Data Analysis Practice

The reduction of time and cost in data analysis is the main objective of the software. Data analysis takes longer as the amount of data produced increases, especially if each step is not automated. The following is an example of a data analysis procedure:

- Step 1: Extraction of raw data from an instrument
- Step 2: Application of data analysis functions (i.e. peak tracing or singular value decomposition) to the raw data
- Step 3: Visualization of results in the form of various graphs and diagrams
- Step 4: Iteration of Step 2 and 3 until acceptable results are achieved.

In the first step, the extraction of raw data generally requires several steps involving human operations because a number of the instruments used in these studies produce data in various formats. Therefore, the data analysis software should be able to import multiple formats of data directly from instruments to avoid such inconveniences. In the second step, the analysis software should support batch work that applies the same analysis to multiple data at once. The

software should be able to produce various forms of graphs and diagrams. All of these functions should be provided seamlessly so that researchers can analyze data with minimal effort.

The software is primarily designed to aid researchers for faster and more convenient data analysis practices. The major advantage is a reduction in cost and time, thereby achieving higher productivity. In addition, the software can reduce human error by minimizing human involvement in the process of data analysis.

3.2.2 Laboratory Data Management

As the volume of data produced in a laboratory increases over time, a systematic method to manage the data becomes necessary. Traditional laboratory notebooks are widely used to record the results of experiments but are insufficient for managing the large volume of data now commonly encountered, especially in an electronic format. The software should be able to store all data in a central database and provide easy access to the data at any time. Since the software has the ability to import the data directly from instruments, the imported data is then stored and managed in the central database automatically. The software should store not only the raw data but also the results of data analysis and visualization. Additional information regarding the experiment procedures (or links to laboratory notebook references) is also required to be recorded along with the experimental data. In this way, all data that are produced in the laboratory can be stored in one place and managed electronically.

3.2.3 Share Data with Others

After all experiment data are stored and analyzed, the entire collection of records can be published electronically and made available for collaborating researchers. The software should

provide quick and easy browsing of published data. Subsequently, researchers should be able to download the published data into their account for further analysis.

3.2.4 Ease of Use and Access

The software should provide an intuitive and easy-to-use graphical user interface. The software should be operational in most computers in the laboratory. In addition, the entire repository of data should be easily maintained.

3.3 Functional Features

3.3.1 Available as a website

MiddaughSuite is developed as a website for ease of access and use. The accessibility of the software is maximized since MiddaughSuite supports almost all major internet browsers such as Google Chrome, Mozilla Firefox, Apple Safari and Microsoft Internet Explorer. There is no need to install any software into a computer because MiddaughSuite is not stand-alone software. A centralized web and data server make it easy to maintain the entire system. At the same time, users do not have any burden to maintain the system because all of the maintenance is performed in the backend.

3.3.2 Multidimensional Matrix

MiddaughSuite treats the experimental data as a form of multidimensional matrix since a multidimensional matrix is suitable for the storage of almost any kind of numerical measurements under various conditions. For example, intrinsic fluorescence spectroscopic data for a macromolecule can be measured in the conditions listed in Table 3.1. Since each condition variable is independent of each other, each condition can be represented as an axis in a 5-

dimensional matrix. The total number of fluorescence measurements in this example is 57,600 (= $6 \times 32 \times 1 \times 100 \times 3$), which is stored in one matrix.

Table 3.1 An Example List of Experimental and Environmental Conditions for Intrinsic Fluorescence Spectroscopic Data

Condition	Value Range	Number of Data Points
pH	3,4,5,6,7,8	6
Temperature	10 – 87.5°C, 2.5°C increment	32
Excitation Wavelength	295 nm	1
Emission Wavelength	300 – 399 nm, 1 nm increment	100
Run	1,2,3 (Triplicate Experiments)	3

There are several advantages to having a multidimensional data model. First, all related experimental data can be organized into a single multidimensional matrix ensuring completeness of data. Completeness of data is one of several key factors that assures the reliability of the experimental results. For example, you cannot organize your data as one multidimensional matrix if you only have two runs of fluorescence experiments (instead of three runs) for pH 8 in the example of Table 3.2. If some experiments are done twice while others three times, comparison between those results would not be valid. Another advantage of the multidimensional matrix data model is ease of maintenance since all related experimental data is merged into one matrix.

Storage of the experimental data as a multidimensional matrix makes it easy for researchers to view various aspects of their data simply by changing viewing angles of the matrix. This type of analysis is useful especially when researchers wish to find the relationship between experimental conditions and results. For example, researchers can easily visualize a line graph that shows an emission spectrum at pH 4 and 10°C in the example of Table 3.2. At the same time, researchers can observe a melting curve at a specific emission wavelength at pH 4 by changing the x axis of a line graph from 'Emission Wavelength' to 'Temperature.' Similarly, researchers can monitor the effect of pH by selecting the 'pH' axis.

If you have a complete set of data in the form of a multidimensional matrix, all of the remaining steps of data analysis are easier and faster. Typically, data analysis can be performed in two steps: data selection and analysis. The range of data can be simply specified by selecting an appropriate range of values in all axes in the matrix. A new matrix can be constructed using only selected values. Subsequently, the resulting matrix can be used as an input to the analysis function. The analysis function also provides its result as a multidimensional matrix. In this way, a major performance improvement is achieved compared to the case when the analysis function is applied to individual data scattered in unorganized files manually.

Suppose a researcher wishes to trace a peak of a fluorescence spectrum by applying the Mean Spectral Center of Mass (MSM) algorithm over the given pH and temperature range in Table 3.2. In this example, one can select a range of emission wavelengths from 320 nm to 360 nm, a range of pH from 3 to 8, and a range of temperature from 10°C to 87.5°C for three runs. By specifying the range of values, a new matrix is constructed and delivered to the MSM peak finding analysis function. The MSM function will return a resulting matrix containing a peak

position and intensity for all pH and temperature ranges for all three runs. This result can be averaged over three runs by applying the ‘Average’ function to the ‘Run’ axis. Finally, the researcher can draw a single line graph that describes a peak shift as the temperature increases for all pH values.

3.3.3 Uploading and extraction of data from various instruments

Experimental data files that are produced by multiple instruments can be uploaded directly to MiddaughSuite (See Table 2.2). This is a key feature of MiddaughSuite: reducing human involvement in the process of data analysis. The uploaded file is then converted to an appropriate multidimensional matrix automatically.

Table 3.2 A List of Supported Instrumentation and Their File Formats

Instrumentation	Experiments Type	File Format (Extension)
Applied Photophysics Chirascan	Circular Dichroism UV Absorbance Fluorescence (Optional)	Proprietary File Format (.dsx)
Photon Technology International Fluorometer	Fluorescence Light Scattering	Proprietary File Format (.ana)
Avacta Optim	Fluorescence Light Scattering	Exported Excel File (.xlsx)
The OLIS Protein Machine	Fluorescence Light Scattering Circular Dichroism UV Absorbance	Proprietary File Format (.ols) Compressed Directory (.zip)
User-supplied Excel File	Any	Excel File (.xlsx)

For example, the fluorescence spectra of four samples were measured using a Photon Technology International (PTI) Fluorometer that produces a resulting file in their proprietary file format (.ana). The resulting file can be directly uploaded to MiddaughSuite and then stored as a single multidimensional matrix described in Table 3.3.

Table 3.3 An Example Matrix Converted from a Photon Technology International Fluorometer File (.ana)

Dimension	Example Value Range
Sample	1,2,3,4
Temperature	10 – 87.5°C, 2.5°C increment
Excitation Wavelength	295 nm
Emission Wavelength	300 – 399 nm, 1 nm increment
Cycle	1,2

It should be noted that sample information is not contained in the experimental data file in many cases. Therefore, one should provide sample information such as sample name, buffer, pH, concentration and ionic strength as well as additional information including number of runs.

MiddaughSuite also provides a method for researchers to upload multiple instrument files as well as their information. To create a complete set of experimental data, it is necessary to run the experiment multiple times. Suppose one plans to collect fluorescence measurements under the conditions described in Table 3.1 using a 4-sample cuvette PTI Fluorometer. Since the fluorometer can only measure four samples at a time, at least six runs of the instrument are necessary to complete three runs at six pH values (i.e., 18 samples plus blank buffer samples).

After a researcher collects all six experimental files, the files should be compressed together as a single file with an information file. The information file contains file names and their sample information. Uploading the compressed experimental file automatically leads to construction of a single matrix that contains all experimental data as described in the information file.

Table 3.4 A List of Analysis Functions in MidaughSuite

Category	Number of Functions	Major Functions
Arithmetic	13	Addition, Subtraction, Multiplication, Division, Power, Average
Calculus	3	Derivative and Integration
Normalization	2	Normalization
Smoothing	1	Savitzky-Golay Filter
Interpolation	1	Spline Interpolation
Curve Fitting	2	Linear Regression, Polynomial Fitting
Spectrum Analysis	5	Peak Picking, Calculation of T_m/T_{onset} , Inner Filter Effect Correction, Structural Index Calculation
Clustering Analysis	3	k-Means Clustering, Hierarchical Clustering
Empirical Phase Diagram	1	Singular Value Decomposition

3.3.4 Analysis Functions

Analysis functions enable high performance data analysis because the functions are designed to take a multidimensional data matrix as an input argument. The input to analysis functions are typically defined by a range of axes values. After an analysis function is applied to the input matrix, the resulting matrices will be stored as new data matrices in MiddaughSuite. Table 3.4 shows a list of currently implemented functions in MiddaughSuite. Additional analysis functions can be added by requests from users.

For example, peak position shift analysis is widely used for analysis of intrinsic fluorescence data. Given an intrinsic fluorescence emission spectrum, a peak position is calculated using a specified algorithm such as a mean spectral center of mass (MSM) method. Changes in peak position are monitored as the environmental conditions are changed including temperature or pH alterations. The multidimensional matrix data in Table 3.1 is an effective example of this analysis because intrinsic fluorescence spectra in various conditions are organized into a single matrix. Although the MSM algorithm only requires one spectrum as an input, the peak position shift analysis function in MiddaughSuite can take the matrix as an input and apply the MSM algorithm to all ranges of temperatures, pH values, and Runs axes in the matrix at once. The result would also be a matrix that contains peak position values in all input conditions.

3.3.5 Visualizations

Visualization of data is an essential step in data analysis. Quick and easy visualization of a multidimensional matrix (or its range of axes values) helps researchers reduce the time needed for data analysis and documentation. MiddaughSuite supports a number of different types of

graphs and diagrams as demonstrated in Figure 3.1. It also supports a tabular view of a matrix that is downloadable as a Microsoft Excel file.

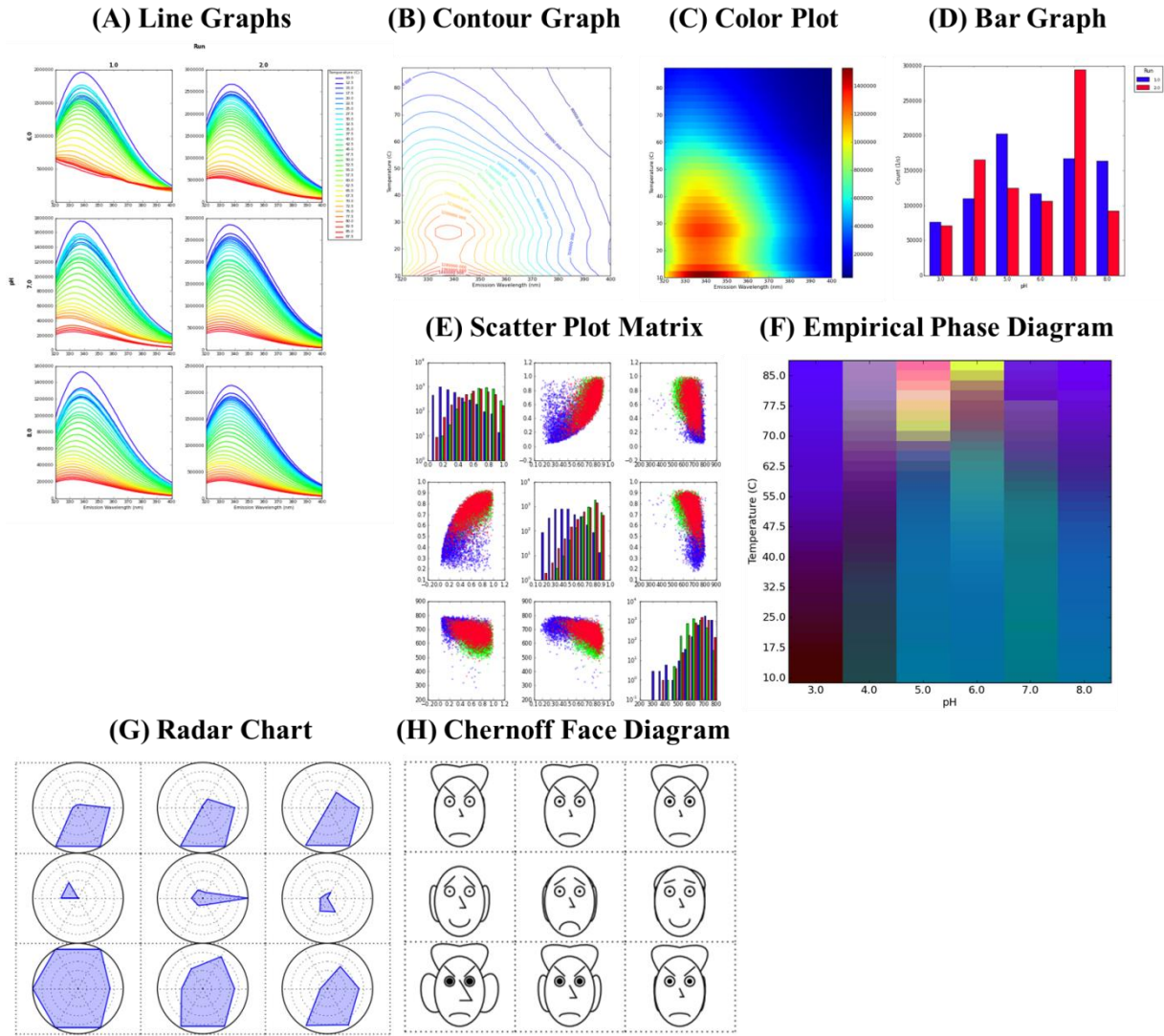


Figure 3.1 Examples of Graphs and Diagrams Supported in MiddaughSuite

3.3.6 Share Data with Others

An additional feature of MiddaughSuite is that it enables data sharing among registered users. The unit of data sharing is a *project* that is a collection of multidimensional data matrices. Generally, closely related data can be managed together in a project. The project and data matrices can contain additional information for detailed descriptions such as images, text, and files. This project can be marked with various sharing options such as *Private*, *Staff Only*, and *Public* indicating different levels of sharing. Other users can search available projects that are shared by others and import them into their accounts.

3.4 Implementation

3.4.1 Overview

MiddaughSuite is a website that was developed using the Python programming language. Python is a widely used programming language that has many beneficial features. One advantage is the availability of free libraries that support variety of functionality. Among others, two significant programming libraries, *SciPy* and *matplotlib* played a key role in the implementation of MiddaughSuite. *SciPy* is a Python library for scientific computation and *matplotlib* is used to plot various types of graphs and diagrams. In addition, a Python-based web framework called *Django* greatly reduces the time and effort to create a website. Finally, *Pyjs* makes it possible to develop an entire website with one programming language, Python, because *Pyjs* can convert Python codes to JavaScript-based web pages.

A website is composed of two parts: client pages and server programs. Client pages are the web pages that are loaded into clients' internet browsers when clients are connected to the

MiddaughSuite's webservice. Through these pages, users retrieve information from the server and perform tasks by sending requests to the server. Server programs handle various requests from the client pages and update client pages with the results. The communication between clients and the server is here based on both Hypertext Transfer Protocol (HTTP) and JavaScript Object Notation-Remote Procedure Call (JSON-RPC) Protocol. JSON-RPC over HTTP is a widely used communication technique especially for implementing dynamic web pages. A JavaScript-based page and JSON-RPC make it possible for the already-loaded web page to modify only a portion of it without reloading the entire page.

Figure 3.2 briefly describes the architecture for the MiddaughSuite website. As a typical setting for a web server, the *apache* HTTP server and *MySQL* database are chosen for their popularity. On top of the *apache* HTTP server, the *Django* web framework handles low-level transactions including HTTP, JSON-RPC and DB access. On top of *Django*, MiddaughSuite can be implemented with Python dealing with more abstracted input and output data. MiddaughSuite utilizes the database as primary data storage but it also makes use of file systems for storage of multidimensional data matrices. This storage of multidimensional matrices requires a specific Python library called *PyTables*.

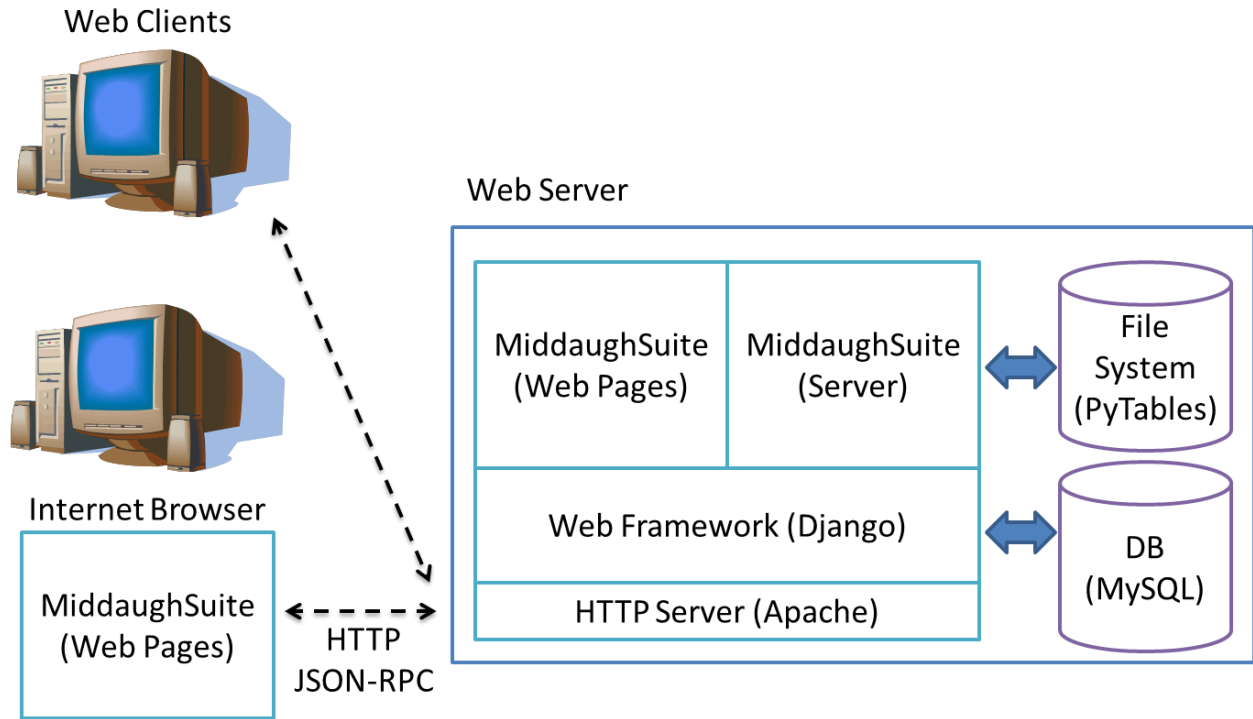


Figure 3.2 MiddaughSuite Architecture

3.4.2 Python Programming Language and Libraries

Python is a computer programming language developed by a Dutch computer programmer named Guido van Rossum in late 1980s. Python is known for its clear and expressive syntax emphasizing code readability. Python supports many features including dynamic data typing, modules, classes, exceptions, multiple inheritance, and automatic memory management. Python is widely used as a scripting language because Python interpreters are available in most computer platforms. The interpreters allow interactive execution of Python codes but they are traditionally known for their slow execution speed. Recent development of Just-In-Time (JIT) compiler and

related techniques help execute interpreted programming languages much faster. In addition, Python can be executed together with modules written in other languages such as C or C++.

One significant advantage of Python is its community support for standard and third party libraries. Its open-source policy, easy-to-use, powerful and clear syntax encourage its usage and development of supporting libraries in many application areas such as scientific computing and web applications. Therefore, Python was chosen for implementation of MiddaughSuite because MiddaughSuite is a web application that handles scientific computing.

SciPy is an open-source Python package for scientific computing. It depends on the NumPy library that handles multidimensional array manipulation. SciPy provides a number of routines in many categories including calculus, optimization, interpolation, clustering, signal processing, linear algebra, and statistics. Since these functions are based on NumPy, they are designed and implemented for multidimensional matrices. In addition, a large collection of routines make it convenient and take less time to develop analysis functions in MiddaughSuite.

Matplotlib is a python library for plotting graphs and diagrams. Supporting charts include line graphs, histograms, power spectra, bar charts, error charts, and scatter plots. More importantly, matplotlib can work with data from SciPy or NumPy in web application settings. Therefore, matplotlib can easily generate graphs and diagrams that MiddaughSuite would like to provide. The quality of figures is appropriate for use in publications.

PyTables is a python library for storage and management of extremely large amounts of data including multidimensional matrices. Although MiddaughSuite uses the MySQL database for its primary data storage of user and data information, multidimensional matrices are too large to be stored inside the conventional database. Therefore, Middaughsuite uses both MySQL and

PyTables for its data storage. Typical operations such as searching and managing user and data information are performed using MySQL. PyTables retrieves and updates matrix contents only when they are needed (e.g. applying analysis functions and plotting graphs).

3.4.3 MiddaughSuite Client Page Implementation

Current advancements in web-based technology enable dynamic web pages that can change their display based on their contents. This technique is essential for MiddaughSuite to provide an intuitive user interface as scientific computing software. Since MiddaughSuite handles multidimensional data matrices, users should be able to select an appropriate range of matrix axes and values as an input for analysis and visualization functions. In addition, such analysis or visualization functions often require additional parameters or need to cross-check other input matrices information. Therefore, the input web pages in MiddaughSuite should be able to suggest candidate values based on user actions of data selection. Candidate values are generally calculated using rules defined for each analysis or visualization function. Design and implementation of dynamic pages that perform complicated calculations and modify their contents at every user action take a significant amount of time and cost without any help from a systematic development environment.

JavaScript is a fundamental computer programming language that is widely used in the development of dynamic web pages. Generally, web pages contain multiple JavaScript functions that interact with the contents of the web pages. Every user actions on a web page including mouse clicks or keyboard strokes trigger such functions to modify their contents. Currently, major web browsers embed a *JavaScript* engine that interprets and executes JavaScript codes inside web pages locally in client computers. In addition, JavaScript is the only language with

which major web browsers share support. Despite its popularity, JavaScript is still known for its inconvenience especially in debugging and readability.

Pyjs is basically a *Python-to-JavaScript* translator with a pre-defined Graphical User-Interface (GUI) Widget Set. A widget is a pre-defined component such as text boxes or buttons. The purpose of *Pyjs* is to make it possible to develop dynamic web pages using Python language instead of *JavaScript*. Since Python is also used in the development of the MiddaughSuite web server, a unifying development environment can increase productivity. Once a web page is written in Python using provided widgets, *Pyjs* converts the Python code to JavaScript pages. During the conversion process, *Pyjs* automatically adds run-time support routines for easier debugging. *Pyjs* also handles all cross-browser issues caused by subtle difference in handling of JavaScript in each browser. Such convenience greatly reduces the time and cost to develop dynamic web pages.

The MiddaughSuite web page consists of only one page although it has many sections and menu items similar to ordinary web pages. Because MiddaughSuite is designed to be a dynamic web page, all contents are loaded in one page and displayed when needed. The benefit of this approach is that the loading of the page happens only one time in the beginning and displays content without any following delay.

3.4.4 MiddaughSuite Server Program Implementation

A web server is a computer that delivers web pages to the Internet browsers in client computers after users request web pages by typing the Internet address of the web server in the browser. A web page includes an HTML document and related images and other information. The Hypertext

Transfer Protocol (HTTP) is used to deliver the web pages. The software that handles HTTP and delivers web pages is also called a web server or an HTTP server.

The Apache HTTP Server is widely-used and freely available web server software. One key feature of Apache is its modular design for easier extension of functionality. Users can selectively install modules that provide additional functions. The *mod_wsgi* module is an apache module to support Python applications to communicate with the apache server via Web Server Gateway Interface (WSGI). The apache server with the *mod_wsgi* module enables the development of the web service using Python language.

Django is a Python Web framework that provides commonly used features for development of web sites in the Python language. The main features include Uniform Resource Locator (URL) processing and database access. URL is also known as the Internet address. Once users type an Internet address in the Internet browser, the web pages indicated by the Internet address are first located by the web server and then served. The implementation of a web site involves mapping between Internet addresses and their corresponding web pages. *Django* provides an easy way to establish such mapping. In many cases, the web pages require information stored in the database. The information includes user account and multidimensional data matrices (e.g. name, axes, and descriptions). *Django* also provides a simple method to access such information in the database. Overall, the utilization of the web framework greatly reduces the time and cost for development of MiddaughSuite.

Chapter 4. The Web User Interface of MiddaughSuite

4.1 Overview

The user interface of a software product determines its users' experience because the users interact with the software through the interface. The interaction includes the methods for the users to provide inputs to the software and to retrieve the corresponding outputs from the software. A well-designed interface should be able to control the information flow between the users and the software efficiently, thus leads to high-performance and easy-to-use software.

MiddaughSuite is web-based software. It provides a single webpage for users as a user interface. Through the webpage, users can upload, manage, analyze, visualize and share their data. Such variety of functionality was integrated into a single page by utilizing dynamic HTML and JavaScript technology. The single dynamic webpage provides a seamless and continuous working experience for the users without disturbance from the repeated loading of different web pages whenever a menu is clicked. The web user interface of MiddaughSuite is also designed to be simple but intuitive and consistent throughout all functionality. This chapter will provide a comprehensive guide for the web user interface of MiddaughSuite.

4.2 Administration

MiddaughSuite provides a separate administration webpage for the administrator to manage user accounts. MiddaughSuite is based on the *Django* web framework that provides the common functionality required for website development. The administration functionality is a part of the

Django framework. Using a pre-defined webpage within the *Django* framework (see Figure 4.1), the MiddaughSuite administrator can easily add new users and manage users information. The administration page can be further extended to include additional features specific for MiddaughSuite. Such features may include the data storage quota and other resource management for higher level administration.

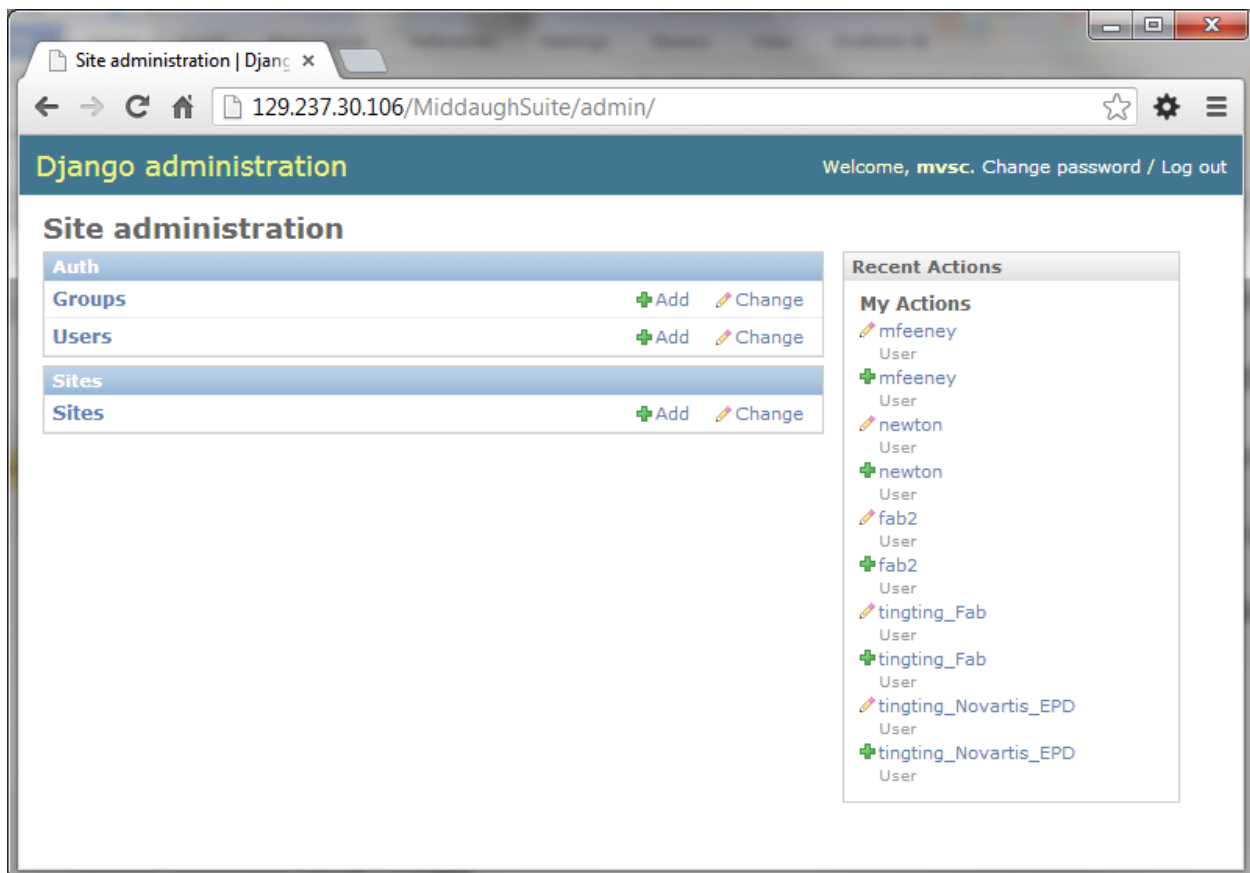



Figure 4.1 MiddaughSuite Administration Page

4.3 MiddaughSuite Main Screen

The MiddaughSuite web page consists of multiple components: header, left side bar, and main window. The header contains logo, user log-in, and main menu. The main menu includes *Home*, *Manual*, *Tutorial*, *Upload*, *Data*, *Analysis*, *Figure*, *Table* and *Search*.

- Home: The *Home* menu displays general information about MiddaughSuite. This is the initial page.
- Manual: The *Manual* menu displays documents that describe how to use MiddaughSuite. It includes a detailed description about each menu. MiddaughSuite provides help icons  in many screen components. If clicked, the corresponding help document in the *Manual* menu will be displayed.
- Tutorial: The *Tutorial* menu displays documents that demonstrate how to characterize a macromolecule from the experimental data. It includes procedures for creating the various empirical phase diagrams.
- Upload: The *Upload* menu displays the screen for uploading experimental data. See Section 4.4 for more details.
- Data: The *Data* menu displays the screen for managing uploaded data. See Section 4.5 for more details.
- Analysis: The *Analysis* menu displays the screen for analyzing uploaded data. See Section 4.7 for more details.
- Figure: The *Figure* menu displays the screen for visualizing uploaded data. See Section 4.6 for more details.

(A)

The screenshot shows the MidaughSuite website header with the logo on the left and a login form on the right. The login form includes the text "Please Log in:" followed by input fields for "username" and "password", and a "Log In" button. Below the header is a navigation menu with "Home", "Manual", and "Tutorial". A left sidebar contains a menu with "About", "Overview", "Account", "Browser", "Contact", and "Release Note". The main content area features the heading "What is MidaughSuite?" followed by two paragraphs of introductory text.

(B)

The screenshot shows the MidaughSuite website header with the logo on the left and a user greeting on the right: "Welcome, jayprimer." followed by "Log Out" and "Account" buttons. The navigation menu now includes "Home", "Manual", "Tutorial", "Upload", "Data", "Analysis", "Figure", "Table", and "Search". The left sidebar menu remains the same. The main content area has the heading "What is MidaughSuite?" followed by introductory text, and then three additional sections: "How to get an account?", "What kinds of browsers should I use?", and "Contact Information", each with its own paragraph of text.

Figure 4.2 MidaughSuite Main Screen: (A) Before a user is logged in, (B) After logged in.

- **Table:** The *Table* menu displays the screen for displaying numbers in table formats from the uploaded data. See Section 4.8 for more details.
- **Search:** The *Search* menu displays the screen for sharing uploaded data with other users. See Section 4.9 for more details.

The menus from *Upload* to *Search* are not available if a user is not logged in (See Figure 4.2). The content of the left side bar and the main window vary according to the selected main menu. For the document pages such as *Home*, *Manual* and *Tutorial*, the left side bar contains navigation panels that point at each section of document pages. For other pages, the left side bar may contain submenus or options.

4.4 Upload Menu

The upload menu provides a simple form (Figure 4.3) in the main window to upload a single file to MiddaughSuite. Using the form, a user can select the file from the local file system. The user should also select the appropriate file type from the drop-down menu. Clicking the upload button will initiate the uploading process. The uploading status will be displayed below the upload form. It may take up to several minutes depending on the file size and type. The uploading result will be notified as described in Figure 4.3 (A) and (B). If successful, the uploaded data becomes available through other menus such as *Data*, *Analysis*, *Figure*, *Table* and *Search*.

Currently, MiddaughSuite supports various file formats from multiple instruments that are available in the Macromolecule and Vaccine Stabilization Center at the University of Kansas. New file formats will be continuously supported as new instruments arrive.

(A)

Upload Data Analysis Figure Table Search

Select a file to upload: IpaD EPD Formatted.xlsx

Select a file type: ▼

Check if the selected file is a ZIP file.

Job	Detail	Status	Start	End
Uploading	IpaD EPD Formatted.xlsx (100%)	✓ Complete	4/1/2013 11:19:15	4/1/2013 11:19:15
Reading	IpaD EPD Formatted.xlsx	✓ Complete	4/1/2013 11:19:15	4/1/2013 11:19:16

The page at 129.237.30.106 says:

Complete!

(B)

Upload Data Analysis Figure Table Search

Select a file to upload: IpaD EPD Formatted.xlsx

Select a file type: ▼

Check if the selected file is a ZIP file.

Job	Detail	Status	Start	End
Uploading	IpaD EPD Formatted.xlsx (100%)	✓ Complete	4/1/2013 11:20:46	4/1/2013 11:20:46
Reading	[Exception] Invalid File Format	✗ Error	4/1/2013 11:20:46	4/1/2013 11:20:46

The page at 129.237.30.106 says:

Error occured during the operation.

Figure 4.3 Upload Screen: (A) when uploading succeeds, (B) when uploading fails

4.4.1 The Applied Photophysics Chirascan

The Chirascan is a circular dichroism (CD) spectrometer manufactured by Applied Photophysics Ltd (APL). It is a multi-functional spectrometer that can simultaneously measure CD, UV absorbance, and fluorescence (when equipped with a fluorescence accessory) and possesses temperature control (0-100°C). There are two different models available: The Chirascan and Chirascan-plus ACD (Automated Circular Dichroism) (See Figure 4.4). The difference between these two models is the sample holders. The original Chirascan has a 4-cell auto changer whereas the Chirascan-plus ACD has an auto-sampler that supports up to four 96-well plates.

(A)



(B)



Figure 4.4 (A) Chirascan (B) Chirascan-plus ACD (Automated Circular Dichroism)

The Chirascan software produces data in an APL proprietary file format with the file extension of “.dsx”. Using the *Upload* menu as described in the previous section, the *dsx* data file can be uploaded to MiddaughSuite with the file type “Applied Photophysics Chirascan File (.dsx).” Once the *dsx* file is uploaded, the data is organized as a multidimensional matrix and stored in the MiddaughSuite server. Figure 4.5 shows an example of Chirascan data dimensions.

Axis	
1. Method	[6] CircularDichroism, HT, Absorbance, Voltage, Count, Temperature
2. Sample	[3] pH4 0.4mg/ml chymotrypsin (1), pH4 0.4mg/ml chymotrypsin (2), pH4 0.4mg/ml chymotrypsin (3)
3. Temperature (C)	[32] 10, 12.5, 15, 17.5, 20, 22.5, 25, 27.5, 30, 32.5, ..., 65, 67.5, 70, 72.5, 75, 77.5, 80, 82.5, 85, 87.5
4. Emission Wavelength (nm)	[161] 200, 201, 202, 203, 204, 205, 206, 207, 208, 209, ..., 351, 352, 353, 354, 355, 356, 357, 358, 359, 360

Figure 4.5 Example of Chirascan data dimensions

- Method: The Chirascan measures using up to seven different methods including CD, UV Absorbance, Fluorescence, HT, Voltage, Count and Temperature based on the measurement setting. Typically CD, UV Absorbance, and Fluorescence data are used in protein analysis.
 - CD: Circular Dichroism
 - HT: Detector high voltage
 - Absorbance: UV absorbance is derived from the detector HT.
 - Fluorescence: Fluorescence measurement (optional)
 - Voltage: A record of the actual detector DC voltage achieved
 - Count: The number of 25us samples taken at every step in the scan
 - Temperature: Temperature at which the measurement was taken
- Sample: Sample name for each cuvette as typed in the Chirascan software. It is recommended that one writes down full conditions in the name filed before the experiment begins.

- Temperature (C): Temperature points in degrees Celsius at which the measurement was taken.
- Emission Wavelength (nm): The range of emission wavelengths in nanometer units.

4.4.2 The Photon Technology International (PTI) Fluorometer

Fluorometers from Photon Technology International, Inc. can be equipped with two detectors and temperature control. Thus it can measure fluorescence and light scattering simultaneously over a selected temperature range. As of 2012, the current version of PTI Fluorometer software produces data in their own proprietary file format with the file extension of “.gxz”. The older version of the PTI software is, however, still widely used and produces “.ana” files.

MiddaughSuite supports both file formats.

Using the *Upload* menu as described in the previous section, the data file can be uploaded to MiddaughSuite with the file type of either “Photon Technology International (PTI) Fluorometer File (.gxz)” or “Photon Technology International (PTI) Fluorometer File (.ana).” Once the data file is uploaded, the data is organized as a multidimensional matrix and stored in the MiddaughSuite server.



Figure 4.6 PTI Fluorometer

PTI .gxz File Format

Each *gxz* file consists of several groups of traces. A group of traces are usually organized into single multidimensional data. Therefore, one file can produce multiple multidimensional data. Each trace has automatically assigned names that contain certain amount of information. Some examples of trace names are listed as follows:

- 'S1 D1 375:420-570'
- 'S1 D2 375:420-570(1)'
- 'S2 D3 375:420-570 (2)'
- 'S1 ExCorr'
- 'S2 ExCorr(1)'
- 'S3 ExCorr (2)'
- 'S1 Temperature'
- 'S2 Temperature(1)'
- 'S3 Temperature (2)'
- 'S1 D1 375:420-570 [COR]'
- 'S1 D1 375:420-570 [COR](1)'
- 'S1 D1 375:420-570 [COR] (2)'

The first token S1 (to S4) indicates the position of the sample. If it is a background measurement, B1 (to B4) can be used instead. The second token D1 (to D2) indicates the detector number. The measurement from each detector, fluorescence or light scattering, is determined based on the instrumental configuration. '375:420-570' indicates the excitation wavelength and the range of emission wavelengths. The number in parenthesis at the end means

the repeated trace at (usually) a different temperature. 'ExCorr' traces are used internally to perform real-time background correction. The corrected trace has the '[COR]' suffix in its name. MidaughSuite takes advantage of the above information when it organizes the data. The newly uploaded PTI *gxz* data would have the dimensions similar to Figure 4.7.

Axis	
1. Wavelength (Nanometers)	[70] 310, 311, 312, 313, 314, 315, 316, 317, 318, 319, ..., 370, 371, 372, 373, 374, 375, 376, 377, 378, 379
2. Sample	[4] S1, S2, S3, S4
3. Detector	[1] D1
4. Excitation Wavelength (nm)	[5] 292, 295, 298, 301, 304
5. Temperature (C)	[27] 10, 15, 20, 25, 30, 35, 40, 42.5, 45, 47.5, 50, 52 ..., 67.5, 70, 72.5, 75, 77.5, 80, 82.5, 85, 87.5, 90
6. Data Type	[2] Corrected, Raw
7. Cycle	[1] 0.0
8. Group	[1] Detector1

Figure 4.7 Example of PTI *dsx* data dimensions

- Wavelength (Nanometers): Emission wavelengths
- Sample: The sample name for each cuvette position is specified as from S1 to S4. If a background is measured, B1 to B4 can be included.

- Detector: The detector number is recognized from the trace name. Values such as 'Detector 1', 'Detector 2', 'ExCorr', and 'Temperature' can be used in this field based on the tokens (e.g. D1, D2, ExCorr, and Temperature) in the trace name.
- Excitation Wavelength (nm): Excitation wavelength is recognized from the trace name where the excitation wavelength is specified.
- Temperature (C): Temperature at which the measurement was taken.
- Data Type: Traces with real-time background correction (e.g. trace names with '[COR]') are marked as 'Corrected'. Other traces are displayed as 'Raw'.
- Cycle: It is possible to conduct multiple temperature runs. It is common to measure another trace at 10°C or 20°C after the entire temperature melt is finished. In this case, 'Cycle' count is increased to differentiate each run. Please note that there is no data available other than at the 10°C or 20°C in the 2nd cycle. These empty data are filled with zeros.
- Group: The group name to which the traces for this data belong.

PTI .ana File Format

The *ana* file format is an older format than the *gxz* file format. The naming convention of the trace name is the same as the *gxz* file format except that it does not support any real-time background correction. The newly uploaded PTI *ana* data would have dimensions similar to Figure 4.8.

Axis	
1. Emission Wavelength (nm)	[86] 305.100006103516, 306.100006103516, 307.1000061035 ... 8.100006103516, 389.100006103516, 390.100006103516
2. Sample	[4] 0.5mg ml aldolase pH4 run 3 emission 2nm.ana S1, 0 ... 3, 0.5mg ml aldolase pH4 run 3 emission 2nm.ana S4
3. Method	[1] Detector 1
4. Excitation Wavelength (nm)	[1] 295
5. Temperature (C)	[32] 10, 12.5, 15, 17.5, 20, 22.5, 25, 27.5, 30, 32.5, ... , 65, 67.5, 70, 72.5, 75, 77.5, 80, 82.5, 85, 87.5
6. Cycle	[2] 1, 2

Figure 4.8 Example of PTI *ana* data dimensions

- Emission Wavelength (nm)
- Sample: The sample name for each cuvette position is specified as from S1 to S4. The uploaded filename is inserted prior to the sample name.
- Method: The detector number is recognized from the trace name. Values such as 'Detector 1' and 'Detector 2' are used in this field based on the tokens (e.g. D1 and D2) in the trace name.
- Excitation Wavelength (nm): Excitation wavelength is recognized from the trace name where the excitation wavelength is specified.
- Temperature (C): Temperature at which the measurement was taken.

- Data Type: Traces with real-time background correction (e.g. trace names with '[COR]') are marked as 'Corrected'. Other traces are displayed as 'Raw'.
- Cycle: It is possible to conduct multiple temperature runs. It is common to measure another trace at 10°C or 20°C after the temperature melt is completed. In this case, 'Cycle' count is increased to differentiate each run. Please note that in this instance there is no data available other than 10°C or 20°C in the 2nd cycle. These empty data are filled with zeros.

4.4.3 The Avacta Optim 1000

The Optim 1000 shown in Figure 4.9 is a high-throughput fluorometer from Avacta Analytical Ltd. The instrument can measure fluorescence and light scattering from up to 48 samples simultaneously with temperature control. Each sample only requires a small volume (around 9 uL) of macromolecule. Optim software supports export of the experimental data in Excel file format. Figure 4.10 shows an example of an exported Excel file. Using the *Upload* menu, the data file can be uploaded to MiddaughSuite with the file type of either “Avacta Optim Export Excel File (.xlsx)” or “Avacta Optim Export Excel File (.xlsx) : Time Series.” The difference between the two options is the way in which MiddaughSuite organizes data.



Figure 4.9 Avacta Optim 1000

	A	B	C	D
1	Wavelength (nm)	A1 @14.99°C @67.91s	B1 @14.99°C @71.31s	C1 @15.00°C @74.72s
2	248.2823931	330.587454	323.2410662	342.8314338
3	248.5456536	51.62585845	76.20960057	100.7933427
4	248.8088912	86.33258913	56.73284429	98.66581615
5	249.072106	86.57871081	54.42090394	116.2628402
6	249.3352979	59.50557543	86.77896417	193.3931202
7	249.598467	79.47928009	106.8002826	260.7913878
8	249.8616131	79.58452778	94.50662674	415.3317543
9	250.1247364	92.10255513	87.12403864	535.190523
10	250.3878367	84.68598782	112.0843956	460.7914043
11	250.6509142	89.74421723	87.25132231	388.891608

Figure 4.10 Avacta Optim Data in Exported Excel

The Optim uses specially designed sample arrays called MCAs. Each MCA has 16 sample wells labelled from A to P. The Optim can use up to three MCAs. Each MCA is assigned by a number from 1 to 3. Therefore, the total 48 sample positions are labelled from A1 to P1, A2 to P2, and A3 to P3. Both the temperature and the time at which the measurement is taken are also specified in the first row. The columns in the file are simply listed by the order in which the measurements are taken. This format makes it hard for users to find traces of each sample either by temperature or by time. There are two different ways for MiddaughSuite to organize the Optim data: by temperature or by time. The method is determined by the file type you choose when you upload your Optim file.

It should be noted that uploading and reading the optim file may take several minutes because its size becomes around 30 MBytes with 48 samples for temperature range from 15°C to 90°C at one degree intervals.

Avacta Optim Export Excel File (.xlsx)

MiddaughSuite organizes the optim data by temperature and ignores time information if “Avacta Optim Export Excel File (.xlsx)” is selected when uploaded. Although the actual temperature value in each sample may slightly vary, MiddaughSuite takes the first temperature value (e.g. the temperature value from A1) and uses it for all the other samples. The uploaded data would be similar to Figure 4.11.



Axis	
1. Sample [48]	A1, B1, C1, D1, E1, F1, G1, H1, I1, J1, K1, L1, M1 ... D3, E3, F3, G3, H3, I3, J3, K3, L3, M3, N3, O3, P3
2. Temperature (C) [63]	14.99, 16.24, 17.45, 18.64, 19.84, 21.03, 22.23, 2 ... 1.01, 82.2, 83.4, 84.61, 85.81, 87.01, 88.2, 89.42
3. Wavelength (nm) [1024]	248.282393077757, 248.545653571865, 248.8088912167 ... 3.533991042405, 503.769124891113, 504.004226606431

Figure 4.11 Example of Avacta Optim data dimensions ordered by temperature

Avacta Optim Export Excel File (.xlsx) : Time Series

MiddaughSuite organizes the optim data by time and ignores temperature information if “Avacta Optim Export Excel File (.xlsx) : Time Series” is selected when uploaded. Although the actual time value in each sample increases, MiddaughSuite takes the first time value (e.g. the time value from A1) and uses it for all the other samples for each round. The uploaded data would be similar to Figure 4.12.

Axis	
1. Sample [48]	A1, B1, C1, D1, E1, F1, G1, H1, I1, J1, K1, L1, M1 ... D3, E3, F3, G3, H3, I3, J3, K3, L3, M3, N3, O3, P3
2. Time (sec) [63]	67.91, 343.05, 618.15, 892.28, 1167.37, 1442.52, 1 ... 7, 16015.12, 16291.28, 16567.43, 16842.52, 17118.7
3. Wavelength (nm) [1024]	248.282393077757, 248.545653571865, 248.8088912167 ... 3.533991042405, 503.769124891113, 504.004226606431

Figure 4.12 Example of Avacta Optim data dimensions ordered by time

4.4.4 The OLIS Protein Machine

The Olis Protein Machine (or Olis Multi-Scan Spectrophotometer) is an instrument that can measure CD, UV absorbance, fluorescence and light scattering from six samples simultaneously with temperature control. Using the *Upload* menu, the data file can be uploaded to MiddaughSuite with the file type of either “The OLIS Protein Machine file (.ols)” or “The OLIS Protein Machine Zipped Directory (.zip).” Once the data file is uploaded, the data is organized as a multidimensional matrix and stored in the MiddaughSuite server.



Figure 4.13 Olis Protein Machine

The OLIS Protein Machine Zipped Directory (.zip)

The Olis software stores its data in the directory structure described in Figure 4.14. Each experimental session creates a directory named by its date and the experimental number. There are three methods subdirectories - ABS, CD, and EM. ABS is for UV absorbance. CD is for circular dichroism. EM is for fluorescence and light scattering. EM contains EM_R and EM_N subdirectories each of which indicates raw data and normalized data, respectively.

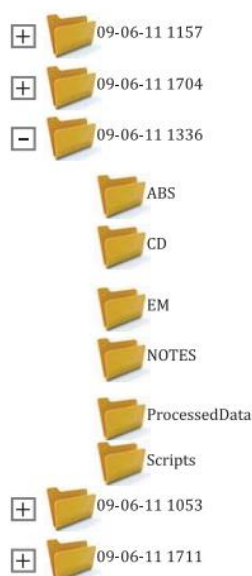
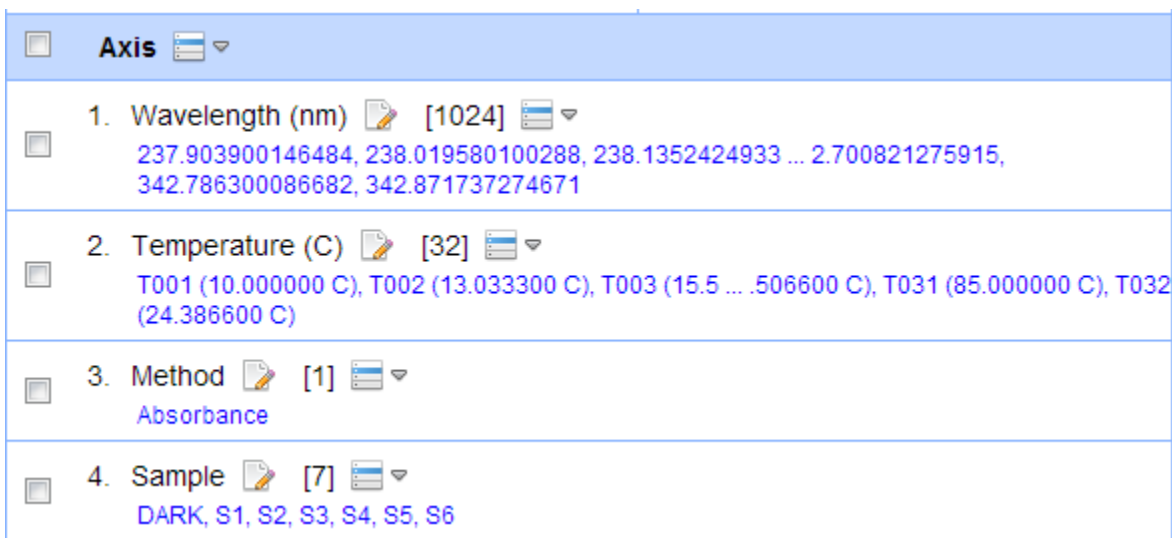


Figure 4.14 Example of Olis data directory structure. This figure is from the online brochure (http://www.olisweb.com/literature/pdf/Multiscan_brochure_low-res.pdf).

It is possible to upload the entire session directory. First, users should make a zip file that contains only the session directory (e.g. a zip file for the directory '09-05-11 1336'). Then users can upload the zipped directory file to MidaughSuite with the file type “The OLIS Protein Machine Zipped Directory (.zip)”. MidaughSuite produces four multidimensional data from a

single zipped directory file based on the measurement subdirectories (e.g. ABS, CD, EM_R, and EM_N). Figure 4.15 shows example dimensions of UV absorbance (ABS) data.



Axis	
1. Wavelength (nm)	[1024] 237.903900146484, 238.019580100288, 238.1352424933 ... 2.700821275915, 342.786300086682, 342.871737274671
2. Temperature (C)	[32] T001 (10.000000 C), T002 (13.033300 C), T003 (15.5506600 C), T031 (85.000000 C), T032 (24.386600 C)
3. Method	[1] Absorbance
4. Sample	[7] DARK, S1, S2, S3, S4, S5, S6

Figure 4.15 Example of Olis Zipped Directory data dimensions (ABS data)

- Wavelength (nm): Emission wavelengths
- Temperature (C): Temperature at which the measurement was taken. Both the measurement order and the actual temperature reading (e.g. “T001 (10.0 C)”) are displayed together. This helps identify the last round of room temperature trace (e.g. “T032 (24.4 C)”) after the temperature melt is completed. It is recommended that temperature values should be edited to numerical values with fixed step size for future manipulation such as visualization.
- Method: Absorbance, Circular Dichroism, or Fluorescence Emission based on the type of measurements
- Sample: Six samples are labeled from S1 to S6. (“DARK” indicates the background for Absorbance data)

The other types of data such as CD and fluorescence emission have similar dimensions. For fluorescence emission data, the additional axis called “Data” distinguishes “Raw” and “Normalized” data that are generated from the “EM_R” and “EM_N” directory.

The OLIS Protein Machine File (.ols)

Inside any directory folder in the Olis data tree, there exists multiple files each of which contains a single measurement at the given environmental condition. The files have the extension of “.ols”. It is possible to upload any single *ols* file instead of the entire zipped directory using the file type “The OLIS Protein Machine file (.ols).” The uploaded data has the same dimensions as the data from the zipped directory.

4.4.5 User-Supplied Excel File

Since MiddaughSuite does not support all file formats, it is necessary to provide a function to upload any arbitrary data prepared by users. Microsoft Excel is a widely used file format in which users can create three dimensional data (e.g. two dimensional tables in each sheet). For convenience, MiddaughSuite provides a template Excel file to be filled by users.

Figure 4.16 shows both an example of a template Excel file and its uploaded data dimensions. In the template Excel file, the table in each sheet must have the same size and labels for consistency. It would not be possible to construct three dimensional data from different shapes of multiple two dimensional data. In each sheet, the starting location of the table is fixed. Cell A3 is the title for the column and the cell B2 is the title for the row. In addition, only one table per sheet is acceptable.

(A)

	A	B	C	D	E	F	G	H
1								
2		pH						
3	Temperature (C)	3	4	5	6	7	8	
4	10	3049.783	4111.77	4779.28	6354.625	7637.3	7915.585	
5	12.5	3105.463	4264.53	4720.693	7072.415	7810.66	8398.69	
6	15	2869.59	3688.023	4523.373	6074.875	8092.72	7642.575	
7	17.5	3169.123	3736.807	4734.66	6011.19	7687.17	7536.395	
8	20	3079.007	3353.043	4355.763	5653.525	7216.785	7122.345	
9	22.5	3139.613	3338.013	4602.18	5801.325	6703.415	7203.86	
10	25	2990.16	3116.27	4156.33	5195.565	6093.56	6625.905	
11	27.5	3178.53	3076.65	4382.793	5601	6764.615	6493.86	
12	30	3464.553	3180.877	3799.377	4966.26	6176.775	6047.86	
13	32.5	3575.947	3603.357	3859.483	4829.635	5494.485	5372.72	
14	35	3583.433	3359.997	3977.92	3986.545	5057.07	4956.405	
15	37.5	3413.497	3323.18	4034.707	3852.09	4585.045	4664.2	
16	40	3388.88	3231.637	4233.36	4067.475	4679.975	4720.925	
17	42.5	3411.29	3201.87	3673.363	4396.69	4748.28	4494.955	
18	45	3450.623	3342.14	3719.957	5566.295	4827.685	5367.38	
19	47.5	3892.63	3763.987	3961.293	4179.76	4541.4	3937.75	
20	50	4571.71	3916.627	3943.47	4972.85	4820.46	4282.115	
21	52.5	9138.533	4876.95	4407.37	4984.17	4934.275	4261.145	
22	55	24259.69	6612.233	4818.837	5291.32	4953.33	4312.37	
23	57.5	57068.59	12182.52	5846.233	5980.42	5087.7	4305.395	
24	60	86975.07	31778.05	7882.02	6255.785	5064.84	3980.105	
25	62.5	101103.7	63860.99	11601.89	7630.315	5147.255	4166.9	
26	65	101493.8	109499.7	27156.18	11121.82	5673.375	4494.11	

(B)

Axis

- 1. Temperature (C) [31]
10, 12.5, 15, 17.5, 20, 22.5, 25, 27.5, 30, 32.5, ..., 62.5, 65, 67.5, 70, 72.5, 75, 77.5, 80, 82.5, 85
- 2. pH [6]
3, 4, 5, 6, 7, 8
- 3. Method [4]
CD Melt, FLU Intensity, FLU Peak, LS
- 4. Sample [1]
IpaD EPD

Figure 4.16 (A) Example of an Excel template file (B) Uploaded data dimensions from (A)

The name of a row or column can be changed arbitrarily. The number of rows and columns (and their labels) can also be changed as well. The only restriction is the identical shape of tables across multiple sheets and their starting locations.

In the example described in Figure 4.16, each sheet represents the experimental data obtained from multiple techniques such as light scattering, circular dichroism, fluorescence peak position shift and fluorescence peak intensity that are labeled as LS, CD Melt, FLU Peak and FLU Intensity, respectively. All of the experimental data are organized in a two dimensional data table with Temperature (row) and pH (column) axes. The row and column titles and the sheet names are converted to the first three axes (e.g. Temperature, pH and Method) in the uploaded data. The last axis will be the axis named “Sample” whose value comes from the name of the uploaded template file. The axis name “Method” and “Sample” are chosen instead of “Sheet” and “Filename” for convenience because the template is frequently used to upload biophysical data.

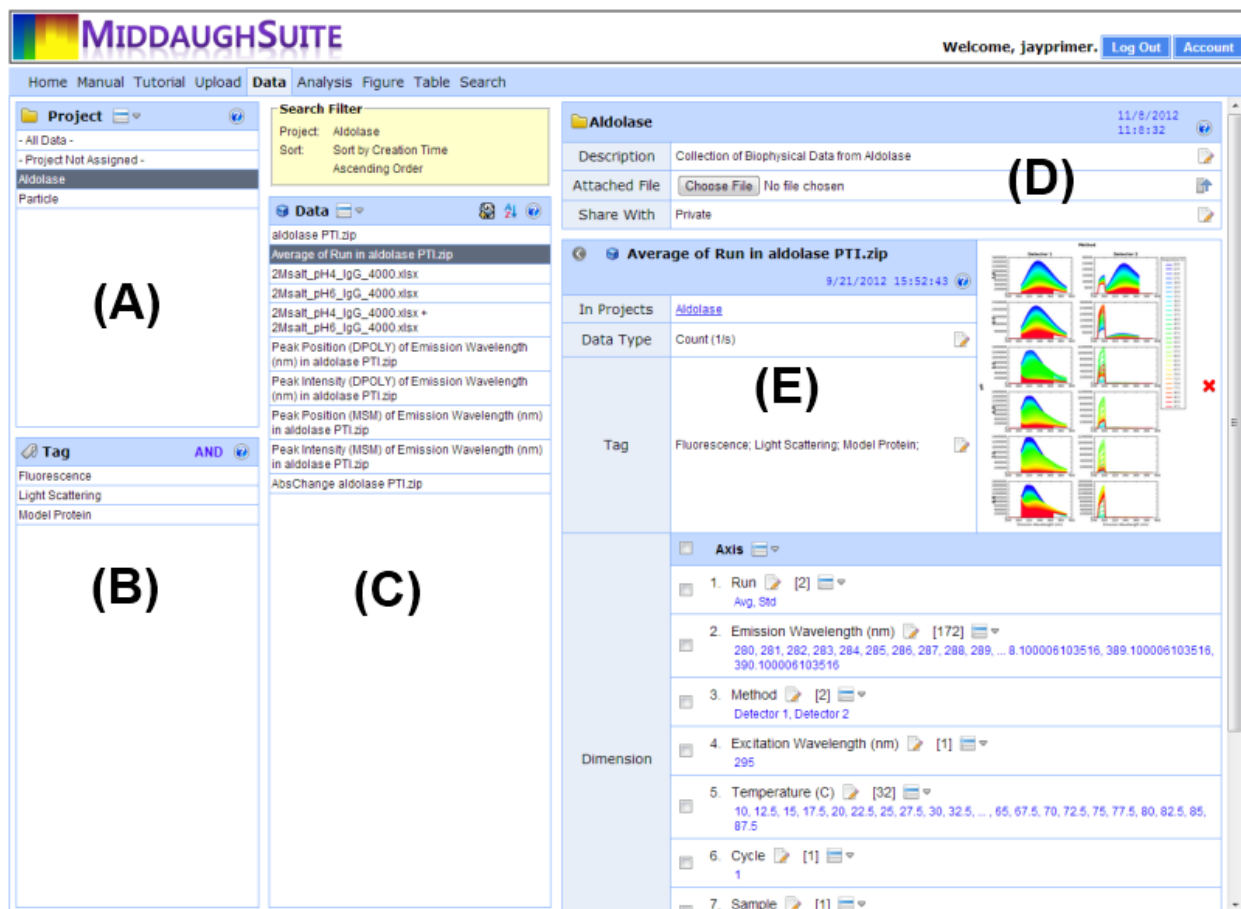


Figure 4.17 Data menu screen: (A) Project List Panel, (B) Tag Panel, (C) Data List Panel, (D) Project Detail Panel, (E) Data Detail Panel

4.5 Data Menu

The *Data* menu is designed to provide functionality for managing the uploaded multidimensional data. The management includes organization of data using projects and tags, viewing details, merging multiple data, changing shapes and editing dimension axes. These functions can be accomplished using the five panels depicted in Figure 4.17: project list panel, tag panel, data list panel, project detail panel and data detail panel.

Projects and tags help organize a large collection of data. Projects are similar to folders that can contain multiple data. A single piece of data can be contained inside the multiple projects. The project list panel lists the current projects and also provides menus for adding and deleting projects.

Tags are user-supplied keywords for the data. The list of tags appears in the tag panel and clicking a tag brings up the list of data to which the tag is assigned. Using projects and tags, users can easily browse their data. The searched list of data appears in the data list panel.

The project detail panel and the data detail panel display detailed information about the selected project and data. The information includes user-supplied descriptions, dimensions and history of changes. For projects, users can attach a single file. For data, users can set up a preview image. The data detail panel also provides menus for editing its values and the shapes of the multidimensional matrix.

4.5.1 Project List Panel

The project list panel displays the list of all projects in the user's account. Clicking any project item in the panel shows the list of data that belongs to the project. Data can be assigned to multiple projects but it does not mean that the data is copied into multiple different folders.

The project is implemented as a specific name tag attached to the data; therefore, the data can have multiple projects without duplication. There are two default items in the project list panel.

- **All Data:** Display all data in the user's account.
- **Project Not Assigned:** Display all data that do not belong to any projects.

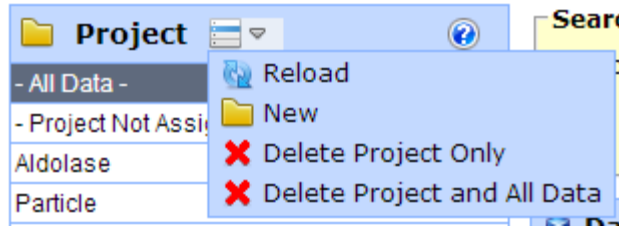






Figure 4.18 The popup menu for the project list panel

The project panel provides a popup menu for creation and deletion of projects as follows:

-  **Reload:** Reload the list of projects.
-  **New:** Create a new project. Enter a new project name in the popup window and press "Add" button.
-  **Delete Project Only:** Remove the selected project label from all data that belong to it. This operation does not delete any data.
-  **Delete Project and All Data:** Delete the selected project and all data in it. The data is not deleted if it belongs to other projects as well.

A popup menu in the data list panel provides functions to add data to a project or to remove data from a project.

4.5.2 Tag Panel

A tag is a user-defined search keyword for data. Data can contain multiple tags that can be viewed at the "Tag" section in the data detail panel. The tag panel displays the list of all tags found in the project selected in the project list panel. A tag is used to filter a list of data displayed in the data list panel.

Clicking a tag enables a tag search filter. Dragging the mouse with the first button clicked, or Using *Ctrl* or *Shift* keys in the keyboard with mouse clicks can select multiple tags. Multiple

tags are applied together by either "AND" or "OR" concatenations. For example in Figure 4.19, the system will find data that contains both “Fluorescence” and “Light Scattering” tags. The option icons **AND** and **OR** are toggled if clicked. In the previous example, clicking **AND** will change its condition to **OR** , and the system will search data that has either “Fluorescence” or “Light Scattering” tags. The icon **✖** will remove the tag search filter and the system will display all data in that project.

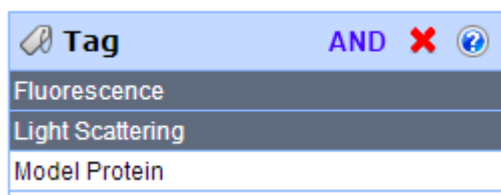


Figure 4.19 Example of Multiple Tag Selection with AND operation

4.5.3 Data List Panel

The data list panel displays data that meets the conditions specified in both the project list panel and the tag panel. Current search methods are displayed in the Search Filter section above the data list panel.

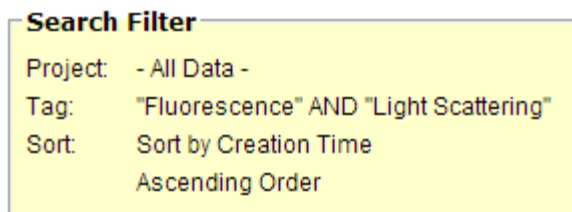





Figure 4.20 Search Filter



Figure 4.20 shows an example search condition. The system will search all data in the user's account, trying to find data that contains both "Fluorescence" and "Light Scattering" tags. Among the results, the data that is created earlier will be listed first. A detailed explanation concerning each condition is as follows:

- **Project:** Project selected in the project list panel. Only a single project can be selected.
- **Tag:** Tags selected in the tag panel. Multiple tags are concatenated by using "AND" or "OR" operations.
- **Sort:** There are options that determine the order of data displayed in the data list panel.

The options below are toggled when the icon in the title bar of the data list panel is clicked.

-  **Sort by Creation Time:** Data is sorted by the time it is created. The creation time cannot be changed once the data is created.
-  **Sort by Name:** Data is sorted by its name.
-  **Sort by Updated Time:** Data is sorted by its updated time. Whenever a user makes a change to the data, the updated time of the data is automatically changed to the current time. A user can check the last modified time at the title section of the data detail panel.

For each sort type option, you can select ascending or descending order.

-  Ascending Order
-  Descending Order

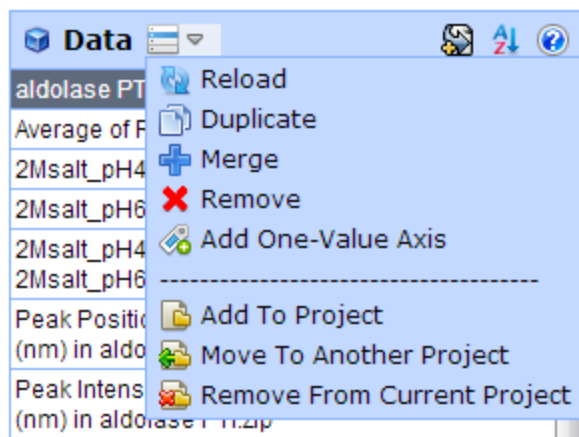










Figure 4.21 The popup menu for the data list panel

The data list panel provides a popup menu as described in Figure 4.21.

-  **Reload:** Reload the list of data.
-  **Duplicate:** Duplicate selected data. It is recommended that data should be duplicated before a user applies any irreversible functions. Currently MiddaughSuite does not support any roll-back functions.
-  **Merge:** Merge selected data and create a new merged data set. The merged data is first created to have union of all axes of the selected data. Initially the merged data is filled with zeros. Then, the content of each selected data is copied to the merged data. If overlapped, the content will be overwritten. If there is a region that is not covered by any data, then the region will remain zero.
-  **Remove:** Remove selected data from the user's account. This operation is currently irreversible.
-  **Add One-Value Axis:** The one-value axis is an axis that contains only one value. Creating a one-value axis is a useful tool for attaching additional information to the data.

For example, a one-value axis such as "Axis: Method, Value: Fluorescence" can be used to indicate that the data is generated using "Fluorescence Method", while it does not affect its internal data. This operation is simultaneously applied to all selected data.

-  **Add To Project:** Add selected data to the project that a user will select from the popup window. A user can even create a new project from the popup window and add selected data to the new project.
-  **Move To Another Project:** Move selected data to the project that a user will select from the popup window. It basically removes the data from the current project (selected in the project list panel) and adds the data to the new project. A user has to select a project in the project list panel first. The default items - "All Data" and "Project Not Assigned" - will not work properly in this operation.
-  **Remove From Current Project:** Remove selected data from the currently selected project. A user has to select a project in the project list panel first. The default items - "All Data" and "Project Not Assigned" - will not work properly in this operation.

4.5.4 Project Detail Panel

The project detail panel displays five attributes of the project that is currently selected in the project list panel as shown in Figure 4.22.








 Project Name		11/9/2012 14:30:0	
Description	Project Descriptions		
Attached File	<input type="button" value="Choose File"/> No file chosen		
Share With	Private		

Figure 4.22 Project Detail Panel

- **Project Name:** You can rename the project by clicking the  icon prior to the project name. You should reload the project list panel to properly reflect the new project name.
- **Last Modified Time:** Displays the last modified time of the other four attributes.
- **Description:** Any detailed information should be placed in this section. You can edit this section by clicking the  icon at the right end.
- **Attached File:** A project can store one file inside MiddaughSuite. It is recommended that one uploads a zipped file that contains all documents and raw data files. You can upload a file by clicking the  icon on the right. You can also remove the attached file by clicking the  icon at the right end. Clicking the uploaded file name will download the attached file.
- **Share With:** You can set the sharing level of the current project by clicking  on the right. There are three levels of sharing: *Private*, *Staff Only*, and *Public*.
 - **Private:** Do not share this project with others. This is the default option.
 - **Staff Only:** Share this project with the staff members who use the MiddaughSuite. Only users that are indicated as staff or a superuser in the MiddaughSuite system can search the project with this option.



- **Public:** Share this project with all users of MiddaughSuite. The *Public* option does not mean the public of the Internet. It means open to all users who are registered in the MiddaughSuite system.

4.5.5 Data Detail Panel

The data detail panel shows specific attributes of the data currently selected in the data list panel. In case of multiple selections, each field displays values that are in common among selected data. Any axis that has no common values will not be displayed.

The example in Figure 4.23 shows a multidimensional matrix that consists of six axes. The total size of the matrix is $81 \times 6 \times 4 \times 32 \times 1 \times 1 = 62208$. There are several one-valued axes (e.g. *Sample* and *Method*). These one-valued axes are typically used to tag additional information. The one-value axis plays an important role when multiple data are merged.

Data in MiddaughSuite stores multidimensional numerical data together with various types of attributes including data name, data type, last modified time, preview image, project information, update history, additional text description, and the attached file.

- **Data Name:** The name of data can be edited by clicking the  icon prior to the data name. The data name displayed in the data list panel will not be changed until it is reloaded.
- **Last Modified Time:** Displays the last modified time of the other data attributes.
-  **Previous Button:** Move to previously selected data. Up to 10 previously selected data are memorized and navigated back.

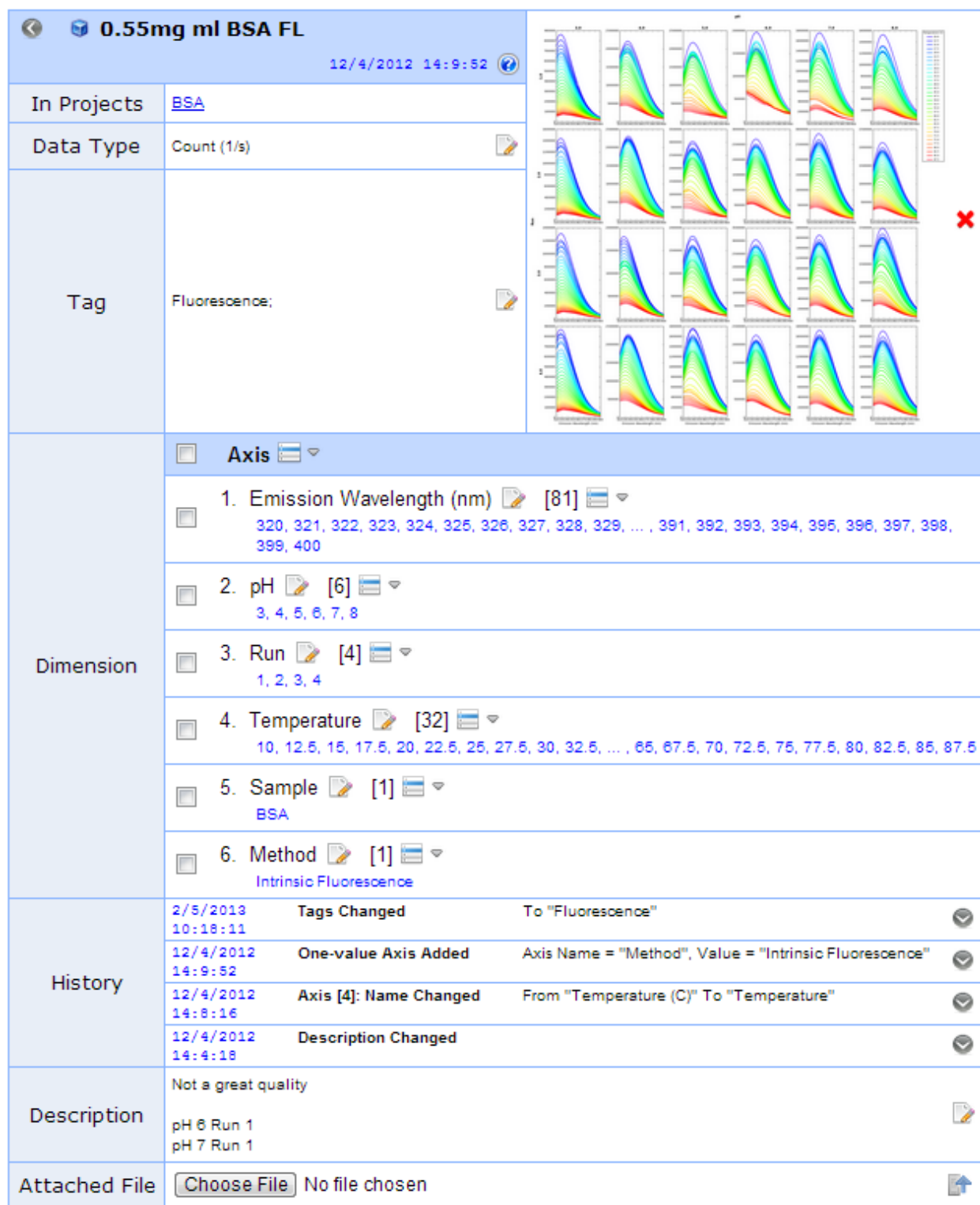
















Figure 4.23 Data Detail Panel

- **Next Button:** Move to next data. This button is enabled only when it is moved back using the previous button.
- **Preview Image:** You can upload an image as a preview image by clicking the  icon at the right end. Also, you can remove the preview image by clicking the  icon on the right. Clicking the preview image will display the image in its full-size as a popup window. You can close the popup window by clicking outside of the window.
- **In Projects:** Displays the list of projects to which this data belongs. You can click the project name link to navigate to it.
- **Data Type:** Displays the label for the data content. You can edit the label by clicking the  icon on the right. The units should be given in parenthesis at the end of the label. (e.g. Count (1/s))
- **Tag:** Displays the list of tags attached to this data. You can edit the label by clicking the  icon at the right end. You have to separate tags using semicolons. (e.g. Fluorescence; Light Scattering; Model Proteins;) Tags are shown in the tag panel for a quick search.
- **Dimension:** Displays all axes and their values in this data as displayed in Figure 4.24.

	Axis No.	Axis Name (Unit)	Number of Values	Values (Preview)
<input type="checkbox"/> Check Box for Menu	2.	Emission Wavelength (nm) 	[172] 	8.100006103516, 389.100006103516, 390.100006103516
<input type="checkbox"/>	3.	M D		280, 281, 282, 283, 284, 285, 286, 287, 288, 289, 290, 291, 292, 293, 294, 295, 296, 297, 298, 299, 300, 301, 302, 303, 304, 305, 305.100006103516, 306, 306.100006103516, 307, 307.100006103516, 308, 308.100006103516

Tooltip: All Values

Figure 4.24 Dimension Field in the Data Detail Panel

- **Axis No.:** Numbering of axes in the data. The order has no meaning.
- **Axis Name (Units):** The name and units of the axis. It is generally displayed in the diagrams and graphs. You can edit the name and unit by clicking the  icon. When the name is edited, the units should be specified in parenthesis at the end.
- **Number of Values:** This field displays the number of values on the axis.
- **Values (Preview):** The values in the axis are displayed in the color blue. Only the first and last few values can be showed if they are excessive in number.
- **Tooltip:** The mouse pointer over the value preview will display a tooltip that shows all values.
- **Checkbox:** A checkbox is used to select axes for the  *Add One-Value Axis* and  *Merge* menus in the Dimension Menu.
- **History:** This field displays the history of operations performed on this data. More detailed information will be displayed if the field is expanded by clicking the  icon. Clicking the  icon will hide the detailed history.
- **Description:** Any specific information should be placed in this section. You can edit this section by clicking the  icon on the right.
- **Attached File:** Data can store one file inside MiddaughSuite. You can upload a file by clicking the  icon on the right. Also, you can remove the attached file by clicking the  icon on the right. Clicking the uploaded file name will start downloading the file.

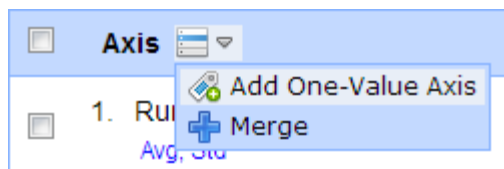


Figure 4.25 The popup menu for the dimension field in the data detail panel.

Dimension Menu

The dimension section in the data detail panel provides a popup menu (Figure 4.25) to manipulate multidimensional axes and their values.

- **Add One-Value Axis:** A one-value axis is an axis that contains only one value. Creating a one-value axis is a useful tool with which attach additional information to the data. For example, a one-value axis such as “Axis: Method, Value: Fluorescence” can be used to indicate that the data is generated using “Fluorescence Method”.
- **Merge:** Merge the selected axes and create a merged axis. The newly merged axis will replace the selected axes. Using the check boxes in each axis, you can select multiple axes to be merged.

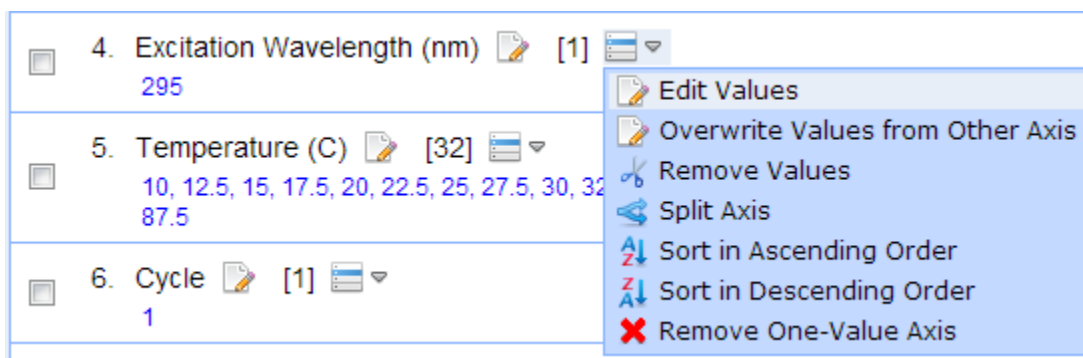


Figure 4.26 The popup menu for the individual axis field in the data detail panel.

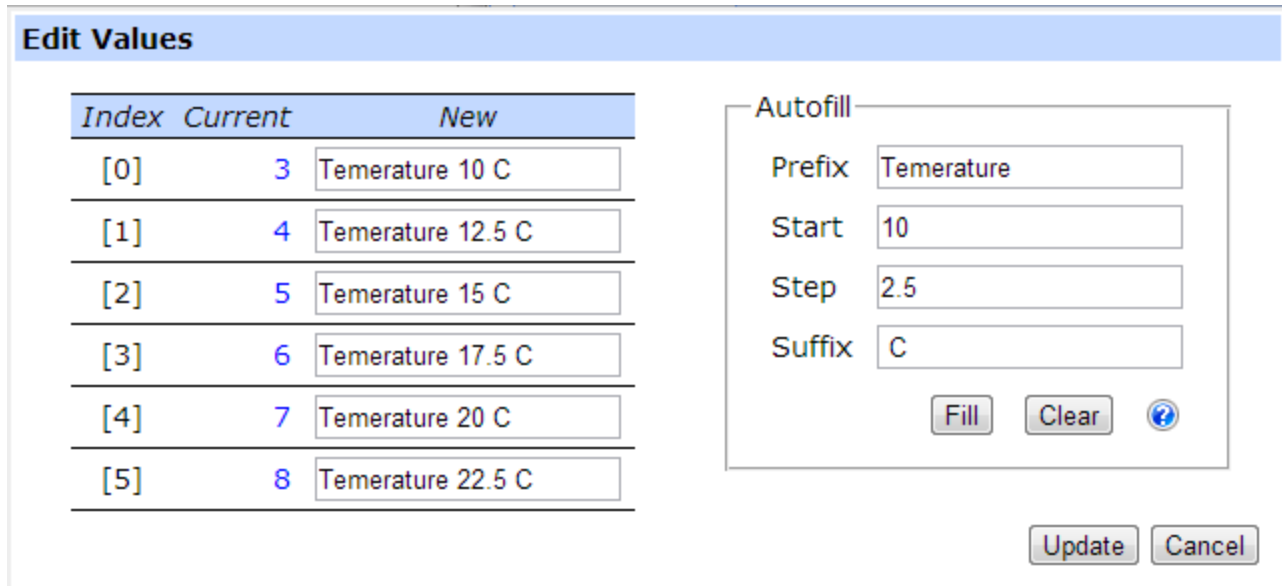






Figure 4.27 A popup window for the *Edit Values* menu.

Individual Axis Menu

Each axis has its own menu as displayed in Figure 4.26. The functionality in this menu is only limited to the axis where the menu is located.

- 
Edit Values: Edit all values in the axis. It does not change any of its internal numerical data. It only edits the label values in the axis. Clicking the menu will bring up a popup window (Figure 4.27). The popup window displays all current values and their associated text boxes where the new values can be entered. Enter new values where you want to change them. Leave the same value where you do not want them to change. You have to make sure that text boxes for new values are all filled even though you do not want to alter some values. The *Autofill* box is designed to fill values with the fixed step size. The value starts from the “Start” value and increments (or decrements) with the specified “Step” size. Prefix and suffix are concatenated at the beginning and the end of

each value. In the example in Figure 4.27, the prefix and the suffix contains a space character at the end and at the beginning, respectively, to insert spaces at the beginning and the end of the numeric value.

- **✂ Remove Values:** This menu is used to remove values from the axis. It will bring up a popup window (Figure 4.28). Values to be removed can be selected using check boxes. The removal of values actually deletes the numerical values in the data accordingly. The operation is irreversible and currently an *undo* operation is not supported in the MiddaughSuite. Always make copies of the original data by using the  *Duplicate* menu before you actually perform irreversible actions. It is sometimes necessary to extract a part of data. The  *Duplicate* and  *Remove Values* menus can be used in this case, although the MiddaughSuite does not have an extract menu. The data should be duplicated first, then unnecessary parts are removed to obtain the desired remaining portion.

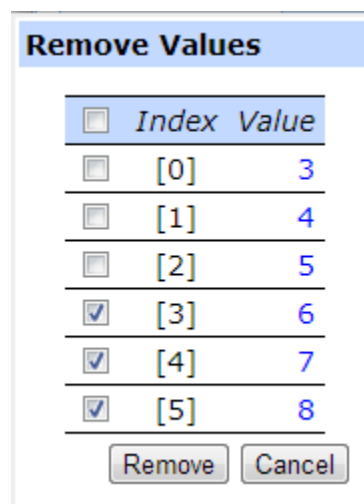







Figure 4.28 A popup window for the *Remove Values* menu.

-  **Split Axis:** This menu is used to reshape an axis into multiple axes. Clicking this menu will bring up a popup window as described in Figure 4.29. In this example, a *Sample* axis that contains 48 values (A1-P1, A2-P2, A3-P3) will be reorganized into three axes (*Sample*, *pH*, and *Run*) based on their properties. The *Sample* axis contains three values of “Protein A”, “Protein B”, “Protein C”. The *pH* axis contains 6 values from 3 to 8. The *Run* axis contains three values from 1 to 3. The previous sample value such as A1 can be decomposed into more detailed combinations such as “Sample: Protein A, pH:3, Run:1.” The *Split Axis* operation replaces the selected axis with the newly assigned axes. It should be noted that the number of the latter combination can be different than the number of values in the original axis. In the previous example, the number of the combination of three axes is 3 samples \times 6 pH values \times 3 runs = 54, while the original sample axis contains only 48 values. The resulting data contains more space than can be covered by the original data; thus, those unassigned region remains zero. The  icon in the blue menu bar is used to add a new column of an axis. The name of the new axis can be edited using the  icon. The  icon will remove the column. After a new axis is added, new values for the axis can be added using the  icon that is located next to the combo box. If a value is added once, the value immediately becomes available for all rows in the axis.

Split an Axis


























































































































































Index	Current	Sample  	pH  	Run  
[0]	A1	Protein A  	3  	1  
[1]	B1	Protein A  	3  	2  
[2]	C1	Protein A  	3  	3  
[3]	D1	Protein A  	4  	1  
[4]	E1	Protein A  	4  	2  
[5]	F1	Protein A  	4  	3  
[6]	G1	Protein A  	5  	1  
[7]	H1	Protein A  	5  	2  
[8]	I1	Protein A  	5  	3  
[9]	J1	Protein A  	6  	1  
[10]	K1	Protein A  	6  	2  
[11]	L1	Protein A  	6  	3  
[12]	M1	Protein A  	7  	1  
[13]	N1	Protein A  	7  	2  
[14]	O1	Protein A  	7  	3  
[15]	P1	Protein A  	8  	1  
[16]	A2	Protein A  	8  	2  
[17]	B2	Protein A  	8  	3  
[18]	C2	Protein B  	3  	1  
[19]	D2	Protein B  	3  	2  
[20]	E2	Protein B  	3  	3  
[21]	F2	Protein B  	4  	1  
[22]	G2	Protein B  	4  	2  
[23]	H2	Protein B  	4  	3  

Figure 4.29 A popup window for the *Split Axis* menu

-  **Sort in Ascending Order:** Sort the values in the axis in ascending order. The internal data is also sorted accordingly.
-  **Sort in Descending Order:** Sort the values in the axis in descending order. The internal data is also sorted accordingly.
-  **Remove One-Value Axis:** Remove this one-value axis. The menu is only enabled for the one-value axis.

4.6 Figure Menu

The *Figure* menu provides the ability to visualize multidimensional data in the form of a number of different graphs and diagrams. The type of graphs includes the line graph, bar graph, contour graph, color plot, scatter plot matrix, empirical phase diagram, radar chart and Chernoff face diagram. The *Figure* menu screen displayed in Figure 4.30 consists of three different regions - *Menu*, *Options*, and *Graph Viewing Area*. The general steps to create a figure are as follows:

1. Select a figure type from the *Menu*.
2. Select a range of data and related options from the *Options*.
3. Click the "View graph/diagram" button that is located between *Menu* and *Options*, or at the end of *Options*.
4. The resulting graph will show up in the *Graph Viewing Area*.

The created figure is simply an image file of Portable Network Graphic (PNG) format.

The figure can be saved or copied for the use in various documents. Clicking the right button of the mouse on the figure will bring up a popup menu.

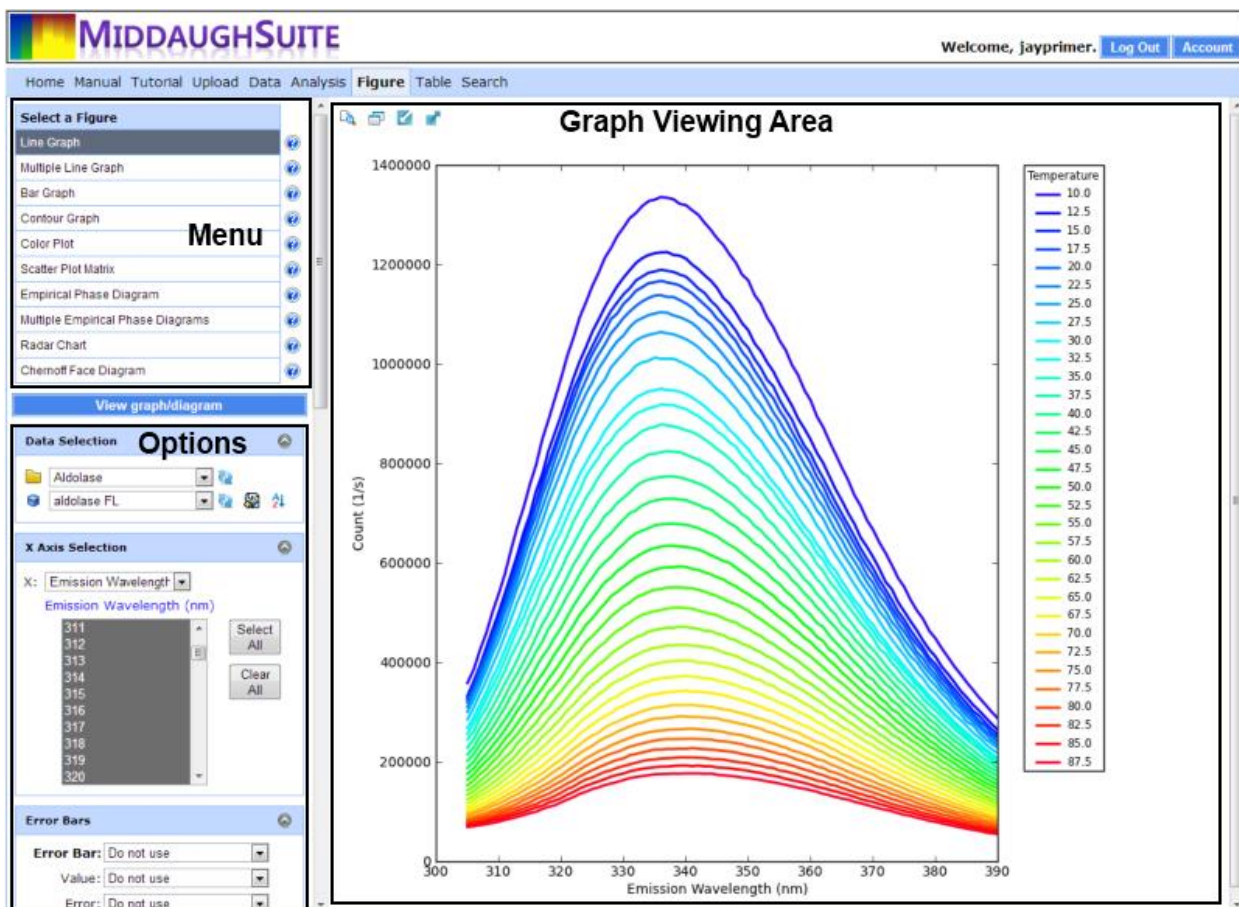


Figure 4.30 Figure menu screen. The screen consists of Menu, Options, and Graph Viewing Area.

4.6.1 Line Graph

The Line Graph menu allows you to create a single line graph. MiddaughSuite provides pre-defined colors, markers, and styles for multiple data lines. Each axis in multidimensional data can be assigned to determine colors, markers, or styles of lines. For example, fluorescence spectra as a function of temperature and pH values are being visualized as line graphs in Figure 4.31B.

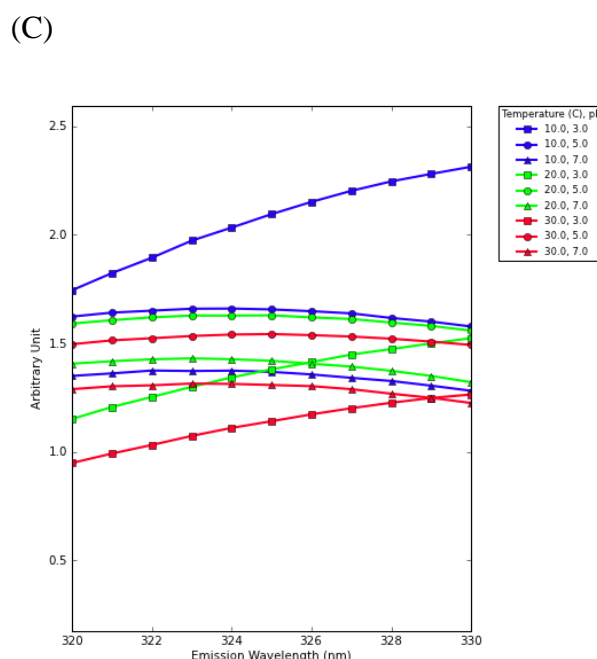
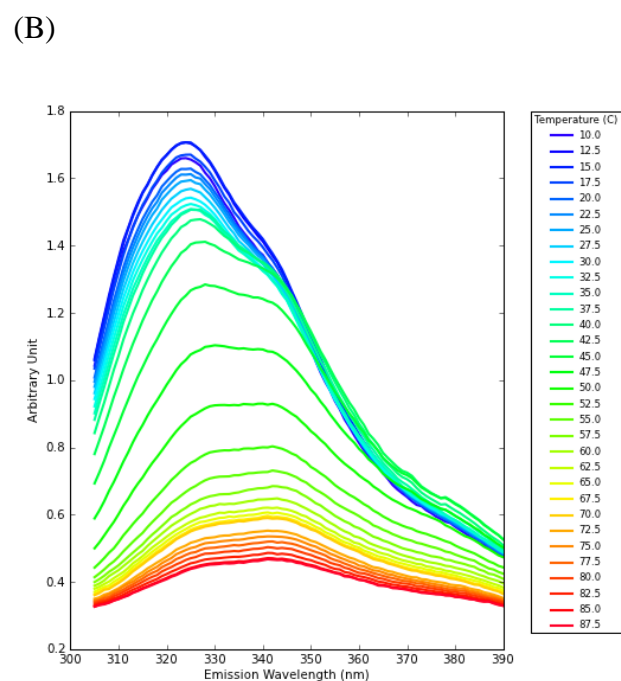


Figure 4.31 Fluorescence spectra of emission wavelengths from 305 to 390nm as a function of temperature and pH: (A) Axes and values and (B) An example of a line graph at a specific pH value. The emission wavelength is assigned to the x axis and the temperature axis to line colors. (C) An example of a line graph. The emission wavelength is assigned to the x axis, the temperature axis to line colors, and the pH axis to line markers.

Data Selection

The first step is to select data for the line graph in the *Options* panel. The data selection panel allows you to select a project and its data.

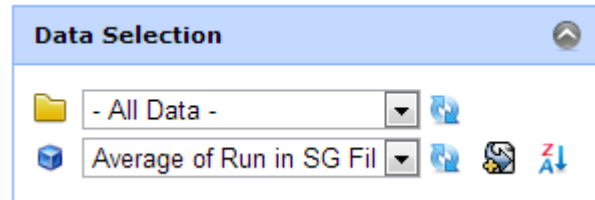


Figure 4.32 Data Selection in the Options panel

- **Project:** The combo box lists all projects.
- **Data:** The combo box lists all data in the project selected above.
- **Reload:** The project or data can be reloaded using this icon. It is necessary to refresh the list after data is added or modified in other menus.
- **Sort Option:** There are options that determine the order of data listed in the combo box.

The options are toggled when the button is clicked.

- **Sort by Creation Time:** Data are sorted by the time they are created. The creation time cannot be changed once the data is created.
- **Sort by Name:** Data are sorted by their name.
- **Sort by Updated Time:** Data are sorted by their last modified time. The time is automatically updated whenever any properties of the data are modified. The last modified time is located at the title section of the data detail panel.
- For each sort type option, you can select ascending or descending order.
 - **Ascending Order**
 - **Descending Order**

X Axis Selection

After the data to be visualized as a line graph is selected, the x axis should be selected in the X Axis Selection box in the Options panel as displayed in Figure 4.33. The range of values for the x axis can be selected as well. The excluded values will not be displayed in the graph. The name of the selected axis will be displayed as the x axis name in the line graph. It should be noted that there is no need to select the y axis because the line graph plots the selected data. The name of the y axis is “Data Type” in the data detail panel.

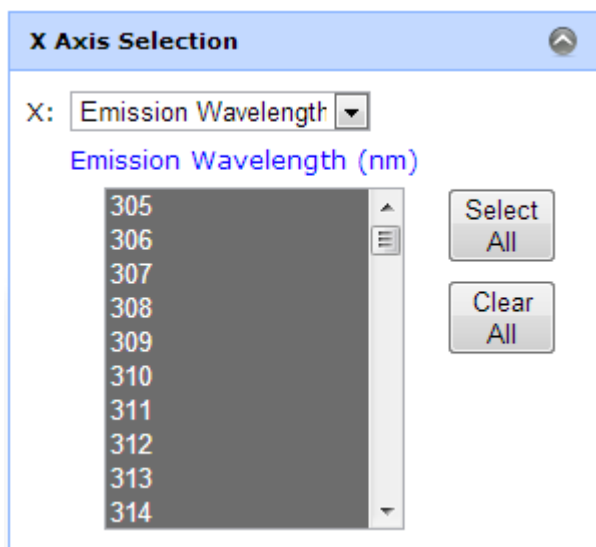


Figure 4.33 X Axis Selection in the Options panel

By changing the x axis of your data, different properties of the data can be revealed. The example in Figure 4.34 demonstrates the emission spectra and the temperature melt from the same fluorescence measurement.

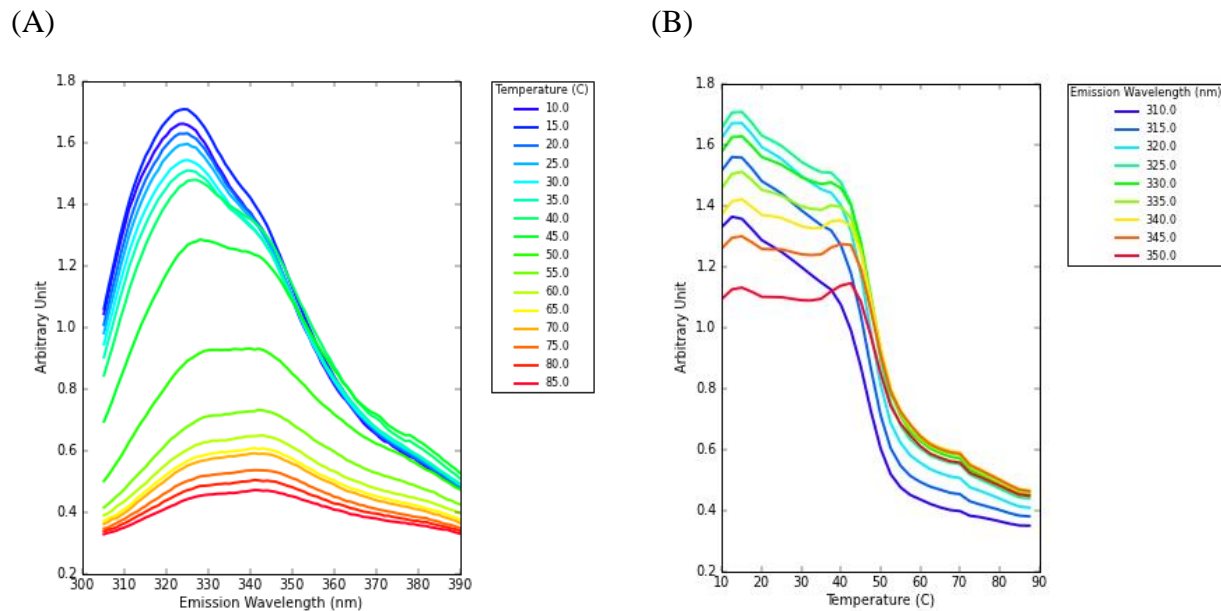


Figure 4.34 (A) The emission spectra and (B) the temperature melt of the same fluorescence measurement.

Error Bars

If you want to display an error bar for each data point, you should select an axis that contains both the averaged value and the standard deviation. MiddaughSuite provides a function called “Average” to calculate the average and the standard deviation. If you do not want to display error bars, select “Do not use” in the Error Bar option.

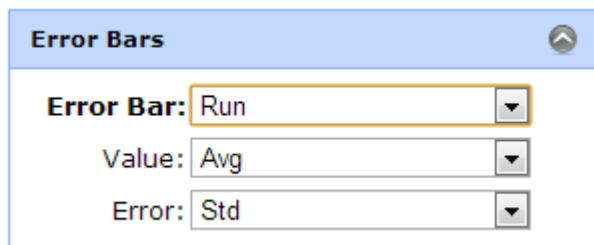


Figure 4.35 Error Bars in the Options panel

Line Color

MiddaughSuite provides a pre-defined set of colors for data. The color set starts with blue and ends with red. The color set is divided by the number of values in the selected axis in the Line Color panel (Figure 4.36). Then, each color is assigned to values sequentially in the selected axis. If the line color is not assigned (i.e. “Do not use” value is selected), all lines will be colored with the same default color, blue.

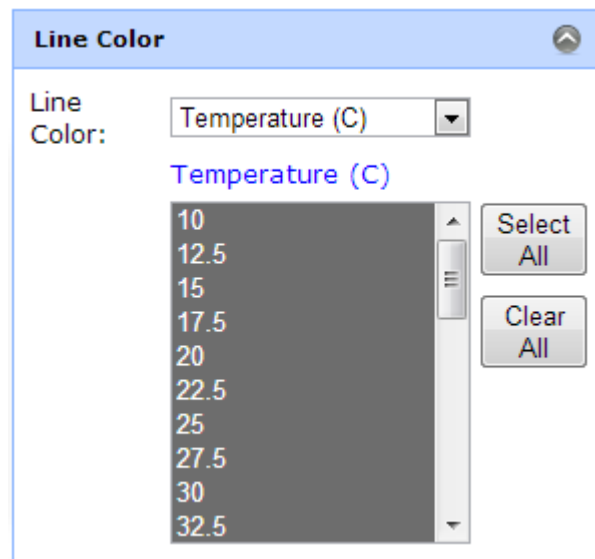
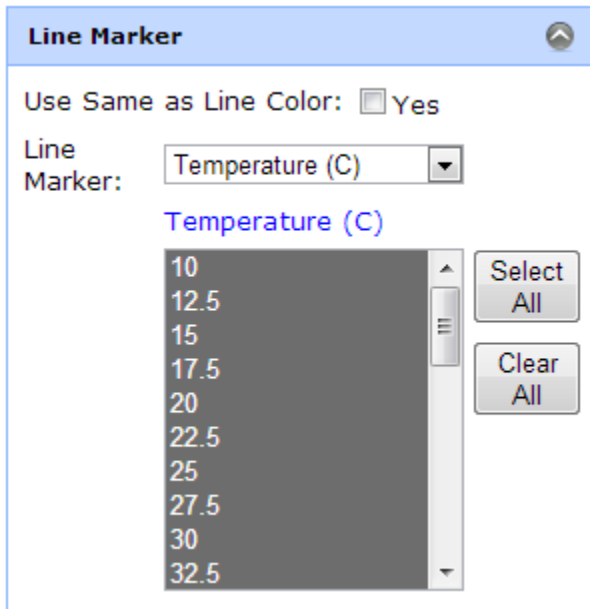


Figure 4.36 Line Color in the Options panel

Line Marker

MiddaughSuite provides a pre-defined set of line markers. The values in the selected axis will be assigned with pre-defined symbols as displayed in the Figure 4.37. If the number of the values is greater than the number of symbols, then the assignment will start over.

(A)



(B)

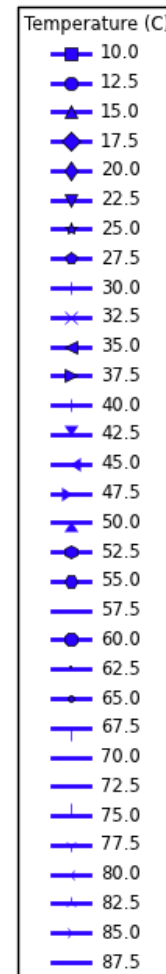


Figure 4.37 (A) Line Marker in the Options Panel and (B) Pre-defined Markers in the MiddaughSuite

- **Use Same as Line Color:** This option is only available when it is combined with the Line Color option. This option enables the assignment of both the colors and markers to the same axis. The example in the Figure 4.38 demonstrates that both the color and marker of the lines change based on the temperature values.

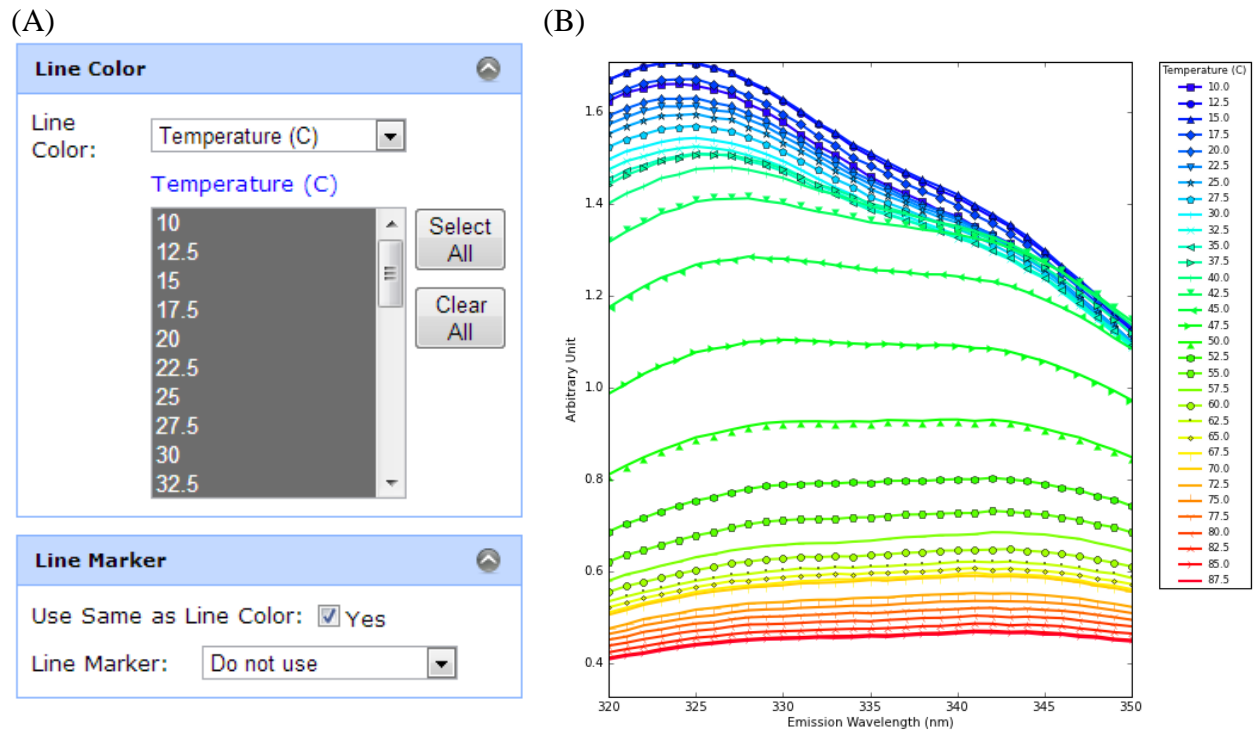


Figure 4.38 (A) “Use Same as Line Color” Option (B) Different markers are assigned to temperature values as well as line colors.

Line Style

MiddaughSuite provides a pre-defined set of line styles. The values in the selected axis will be assigned with pre-defined styles. If the number of the values is greater than the number of styles, then the assignment will start over. The example in the Figure 4.39 shows pre-defined styles. It should be noted that there are currently only four styles available: solid, dashed, dotted, and dashed-and-dotted lines.

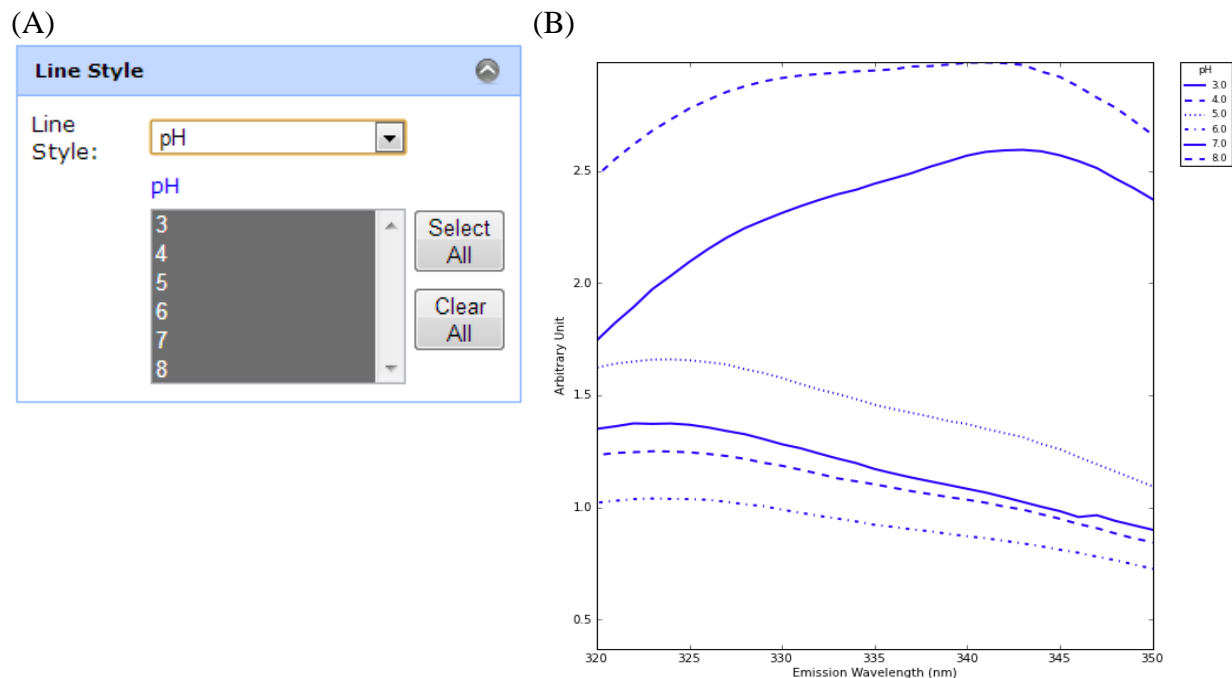


Figure 4.39 (A) Line Style in the Options panel (B) Pre-defined line styles in the MidaughSuite

Remaining Axes

If the data has more axes than the options above, all of the remaining axes and their values should be specified in this panel. It is recommended that one selects only one value for each axis because multiple selections will share the same line color, marker and style setting.

Manual Zoom

The viewing range of any graph in the MidaughSuite is automatically calculated if not specified by users. To manually set up the viewing range, the “Check to set a viewing range” option in the *Manual Zoom* in the *Options* panel should be checked as displayed in the Figure 4.40. Minimum

and maximum values for x and y axes can be specified. If any value is left blank, an automatically calculated value will be used instead.

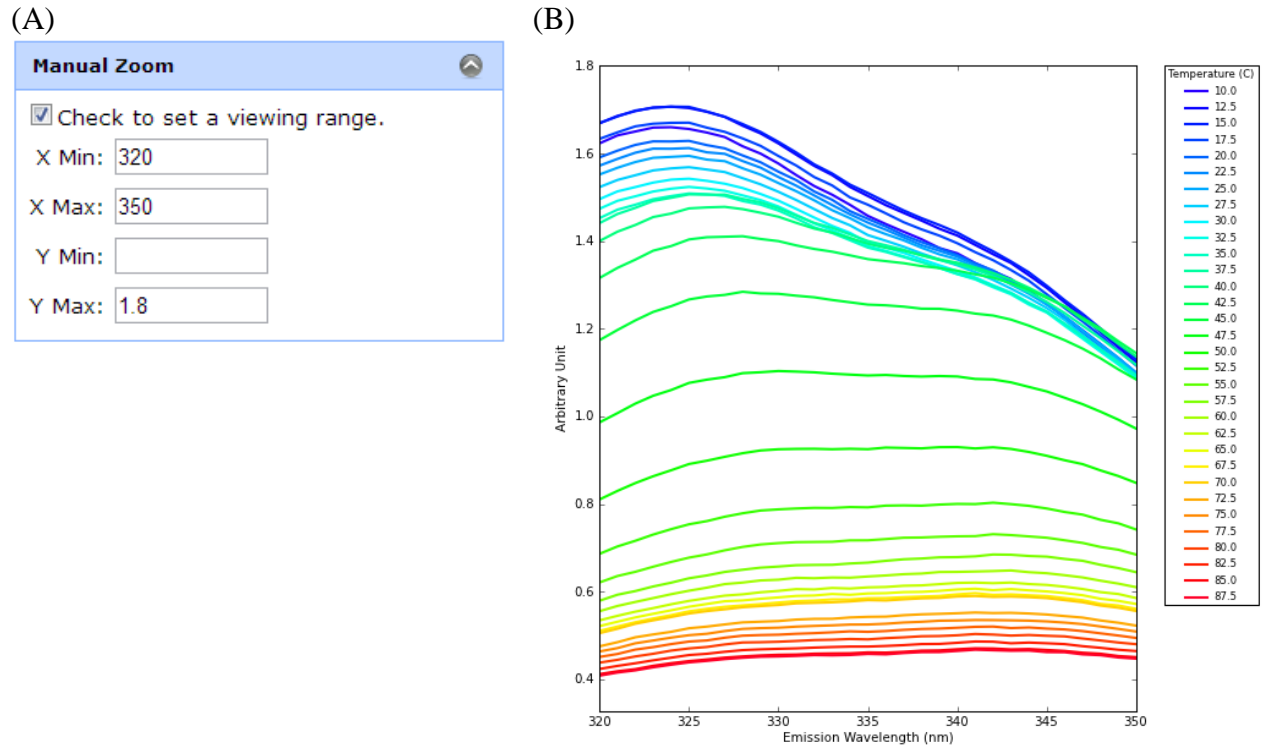


Figure 4.40 (A) Manual Zoom in the Options panel (B) The line graph with specified range. The system calculates minimum of the y axis because it is not specified.

Size Option

The size of a figure can be specified in *Size Option* in the *Options* panel. The size is defined as follows:

- Figure Width (pixels) = Width (in inches) × DPI (dots per inches)
- Figure Height (pixels) = Height (in inches) × DPI (dots per inches)

(A)

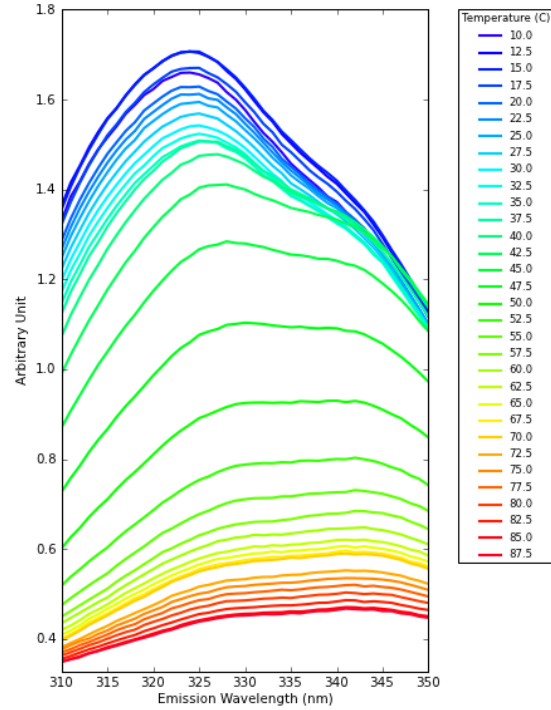
Size Option

Width: inches

Height: inches

DPI:

Width (pixels) = $6 \times 75 = 450$
Height (pixels) = $8 \times 75 = 600$



(B)

Size Option

Width: inches

Height: inches

DPI:

Width (pixels) = $4 \times 112.5 = 450$
Height (pixels) = $5.33 \times 112.5 = 600$

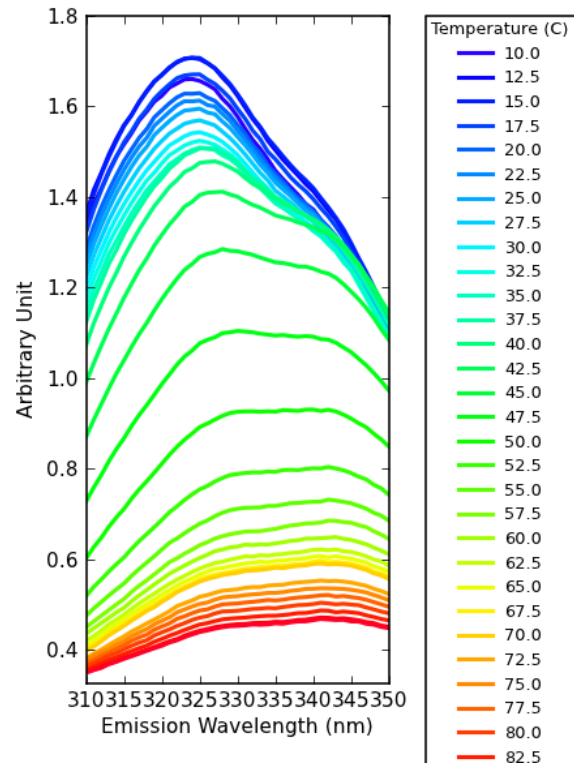


Figure 4.41 Same size (450×600) of figures with different size options

In the example in Figure 4.41, both figures (A) and (B) have the same size (460 × 600) in pixels with different size options. Because the font size and the line width are fixed relative to the width and height of the figure given in inches, figure (B) has larger fonts and thicker lines than figure (A). In addition, the legend in figure (B) is truncated because of a lack of space.

4.6.2 Multiple Line Graph

The *Multiple Line Graph* consists of multiple line graphs placed in the grid as displayed in the Figure 4.43. It requires two additional axes to define the grid in which the line graphs are placed. These axes are called *Group X* (column) and *Y* (row) axes and can be selected in the *Group Selection* in the *Options* panel (Figure 4.42). The names of the axis and values are placed as the titles of each row and column. The *Multiple Line Graph* is created as a single image file.

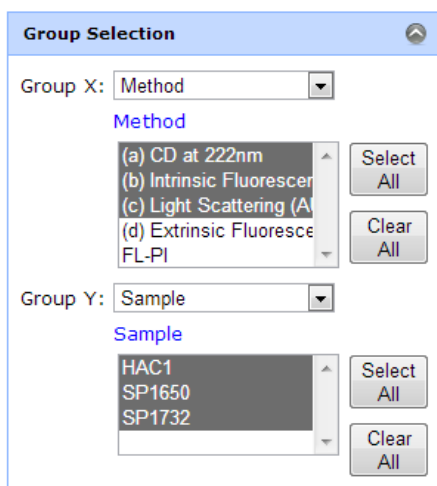


Figure 4.42 Group Selection in the Options panel

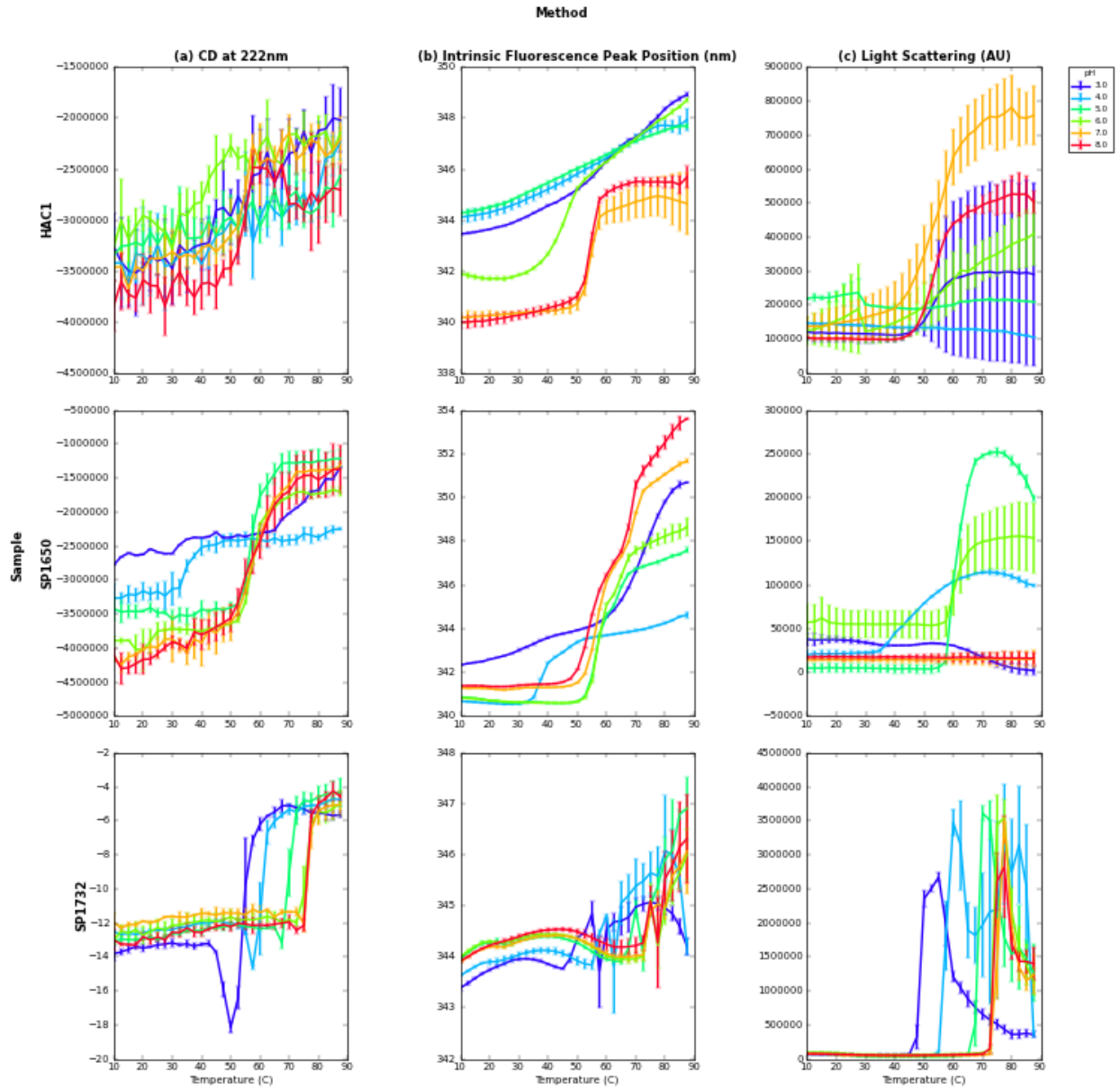


Figure 4.43 An example of Multiple Line Graph. Group axes selected in the Figure 4.42 are used.

4.6.3 Bar Graph

The Bar Graph menu is used to create a single bar graph such as shown in Figure 4.44.

MiddaughSuite provides pre-defined colors for multiple bars. All general options are the same as those in the line graphs.

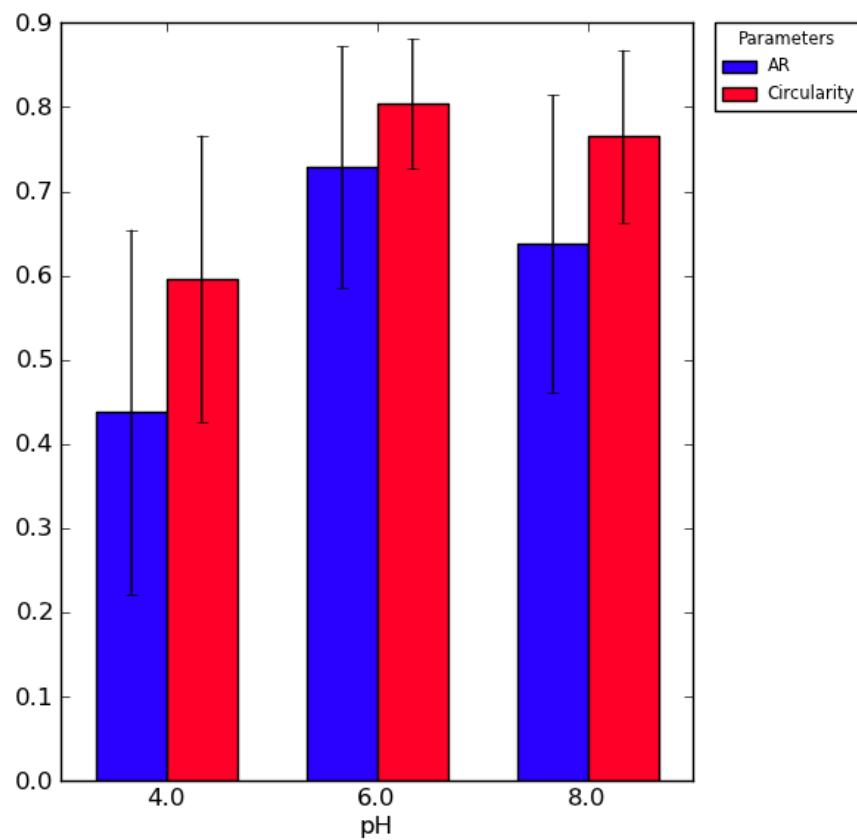
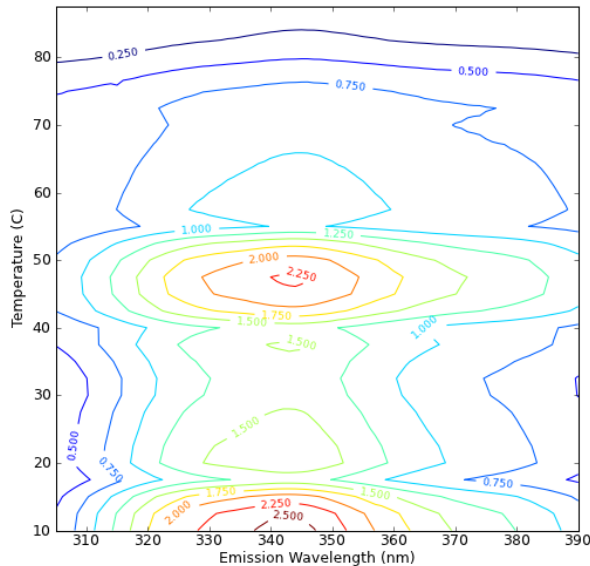


Figure 4.44 An example bar graph.

4.6.4 Contour Graph

The *Contour Graph* menu is used to create a single contour graph. A contour graph consists of multiple contour lines. Each contour line connects points in x and y coordinates whose value is the same. The x and y axes should be specified in the *Axes Selection* in the Options panel. The value for each contour line is determined by the number of levels option. The number of levels determines the number of contour lines between the minimum and maximum values of the data. Figure 4.45 demonstrates examples of contour plots with different numbers of levels. All other general options are the same as those in the line graphs.

(A) Number of Levels = 10



(B) Number of Levels = 30

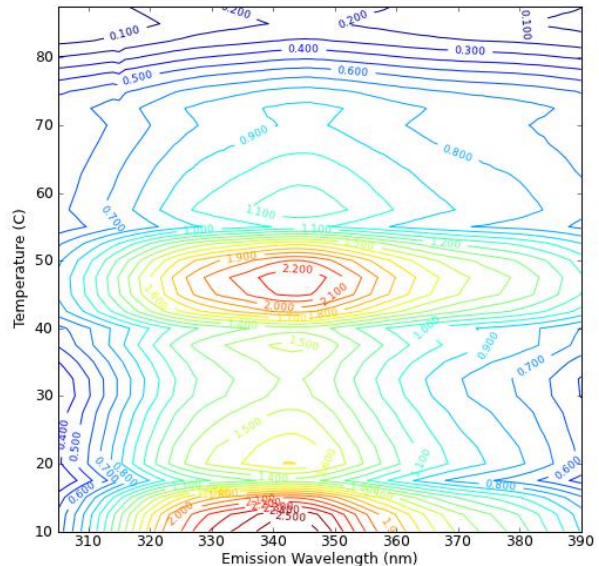


Figure 4.45 Example contour plots with (A) number of levels = 10 and (B) number of levels = 30

4.6.5 Color Plots

The *Color Plot* menu is used to create a single color plot such as illustrated in Figure 4.46. A color plot is a two-dimensional plot where data points with x and y coordinates are represented as individual colors. MiddaughSuite provides a pre-defined set of colors for the color plot. All other general options are the same as those used in the line graphs.

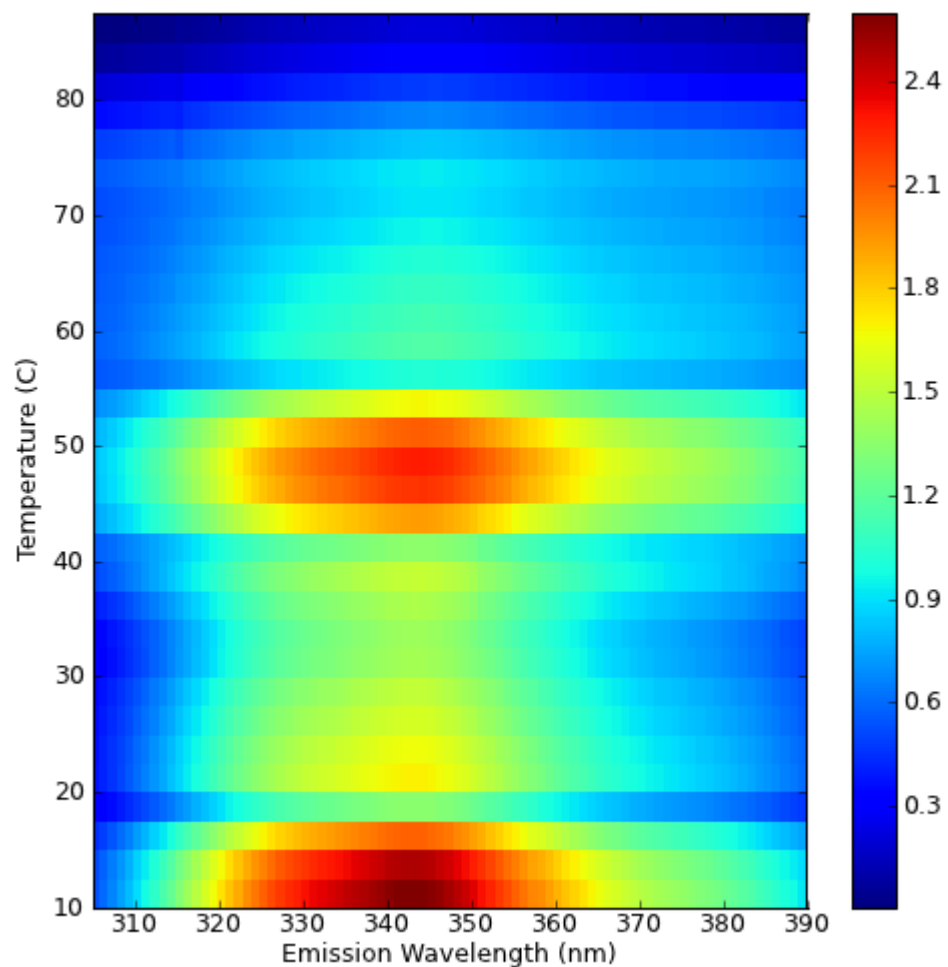


Figure 4.46 An example of a color plot.

4.6.6 Scatter Plot Matrix

The scatter plot matrix is a two-dimensional matrix that consists of scatter plots and histograms. Suppose there are N samples that have M attributes. This data can be given as an $N \times M$ matrix. The objective of the scatter plot matrix is to investigate the correlation between each pair of attributes. Each scatter plot displays the correlation between a pair of attributes by plotting N samples according to their attribute values. Since there can be M -choose-2 pairs of attributes, the same number of scatter plots consists of the half of the scatter plot matrix. The other half would be filled with the same pairs but their x and y axes are flipped. The diagonal of the scatter plot matrix is filled with histograms that show the distribution of the attribute values of the samples.

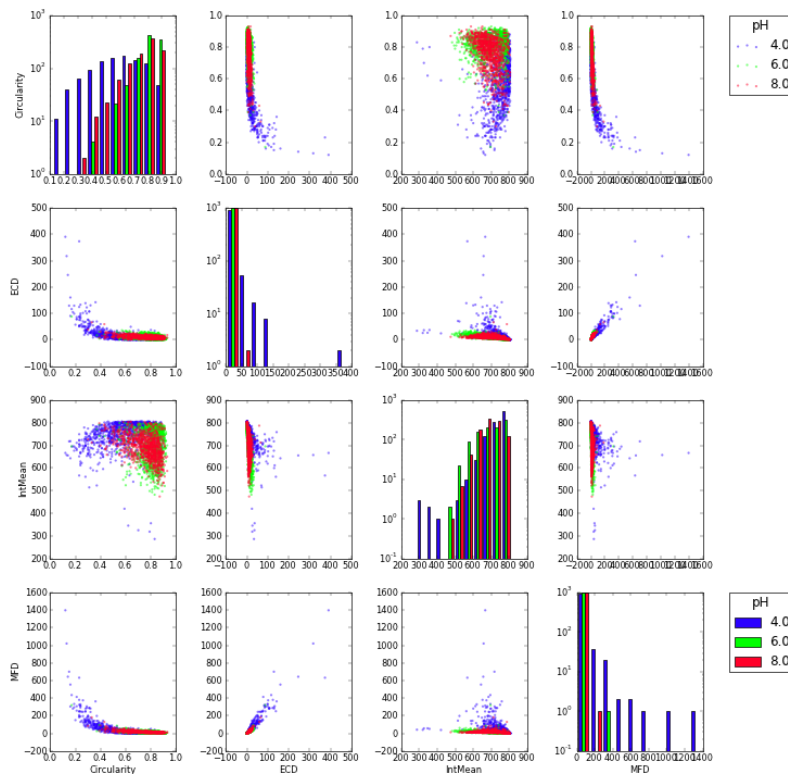
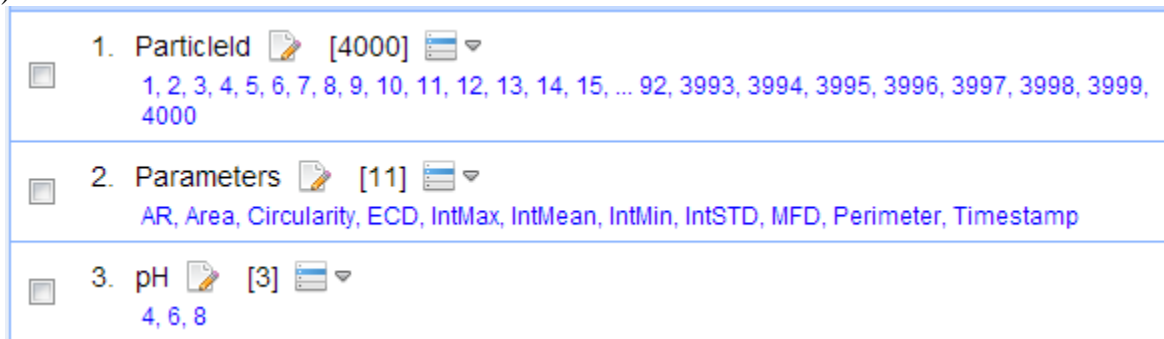


Figure 4.47 An example of a scatter plot matrix.

(A)

	A	B	C	D	E	F	G	H	I	J	K	L
1												
2		Parameters										
3	ParticleId	ECD	AR	Circularity	IntMean	IntSTD	IntMin	IntMax	Area	Perimeter	MFD	Timestamp
4	1	1.125	0.6	0.68	799.63	6.12	786	805	8	14.66	1.125	0
5	2	1.125	0.58	0.79	794.25	11.92	772	811	12	15.49	1.375	0
6	3	1.125	0.59	0.73	797.22	5.07	789	802	9	14.49	1.125	0
7	4	3.375	0.54	0.57	785.72	16.81	745	809	43	41.04	7.875	0
8	5	1.125	0.47	0.6	804.46	6.09	792	810	13	21.31	2.125	0

(B)



1. ParticleId [4000]
1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, ... 92, 3993, 3994, 3995, 3996, 3997, 3998, 3999, 4000

2. Parameters [11]
AR, Area, Circularity, ECD, IntMax, IntMean, IntMin, IntSTD, MFD, Perimeter, Timestamp

3. pH [3]
4, 6, 8

Figure 4.48 (A) Particle information in a sample organized in Excel file from MFI data (B) MidaughSuite data of 4000 particles with 11 parameters at pH 4, 6, and 8.

Figure 4.47 is an example of a scatter plot matrix that shows 1000 particles present in IgG samples at pH 4, 6, and 8 observed using a Micro Flow Imaging (MFI) instrument. MFI instruments produce particle information with 11 different attributes as described in Figure 4.48. The resultant data should be organized as an Excel file (Figure 4.48A) to be uploaded to MidaughSuite because the MFI file format is not directly supported. The uploaded data can be merged into data such as Figure 4.48B if samples are present under various conditions (e.g. pH 4, 6 and 8). Colors can be assigned to distinguish sample conditions.

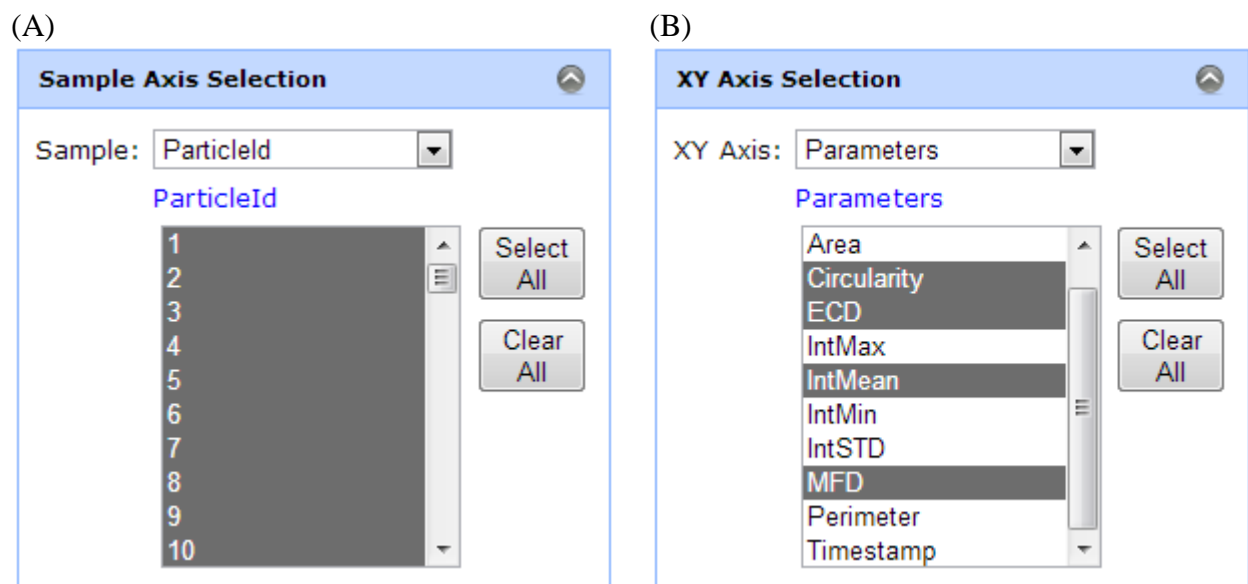


Figure 4.49 (A) Sample Axis Selection and (B) XY Axis Selection in the Options panel

Sample Axis Selection

A sample axis should be selected to distinguish each sample. In the previous example, the axis ‘ParticleId’ contains particle numbers and should be selected in the *Sample Axis Selection* in the *Options* panel (Figure 4.49A).

XY Axis Selection

After a sample axis is selected, its attributes axis should be selected. In the previous example, the axis ‘Parameters’ contains all attributes for the samples with 4 out of 11 attributes selected (Figure 4.49B). Each pair of selected attribute values becomes x and y axes of scatter plots in the scatter plot matrix.

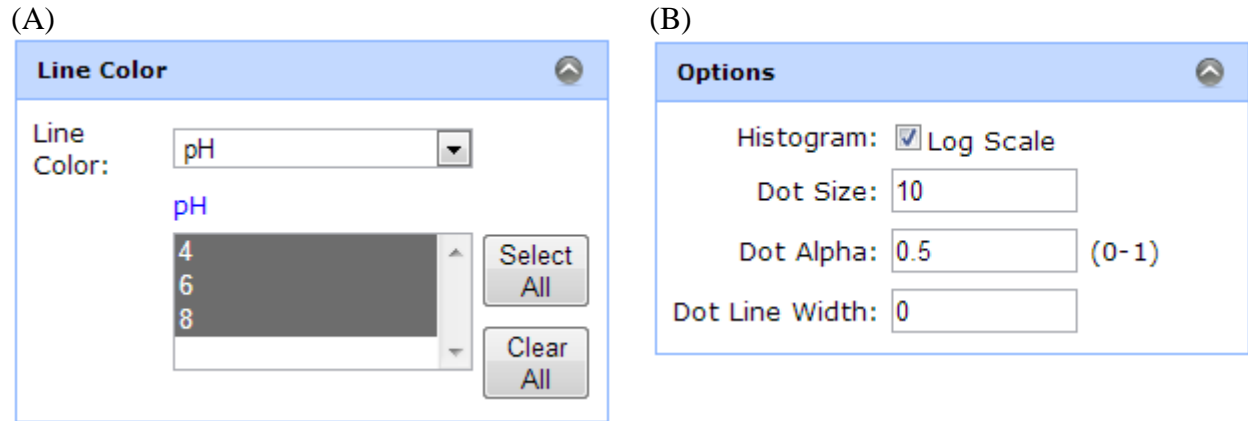


Figure 4.50 (A) Line Color and (B) Options in the Options panel

Line Color

Colors can be assigned to help visualize samples in different conditions. The axis that represents an individual condition (e.g. pH values) can be selected in this panel (Figure 4.50A). If the “Do not use” option is selected, all samples are colored blue.

Options

There are several options available in the options panel as displayed in the Figure 4.50B. The effects of these options are shown in the Figure 4.51.

- **Histogram - Log Scale:** If selected, a log scale is applied to all histograms.
- **Dot Size:** The dot size determines the size of the circle that represents each sample.
- **Dot Alpha (0-1):** The dot alpha determines the opacity of the circle. (0 for transparent, 1 for opaque)
- **Dot Line Width:** The dot line width determines the width of the circumference of a circle.

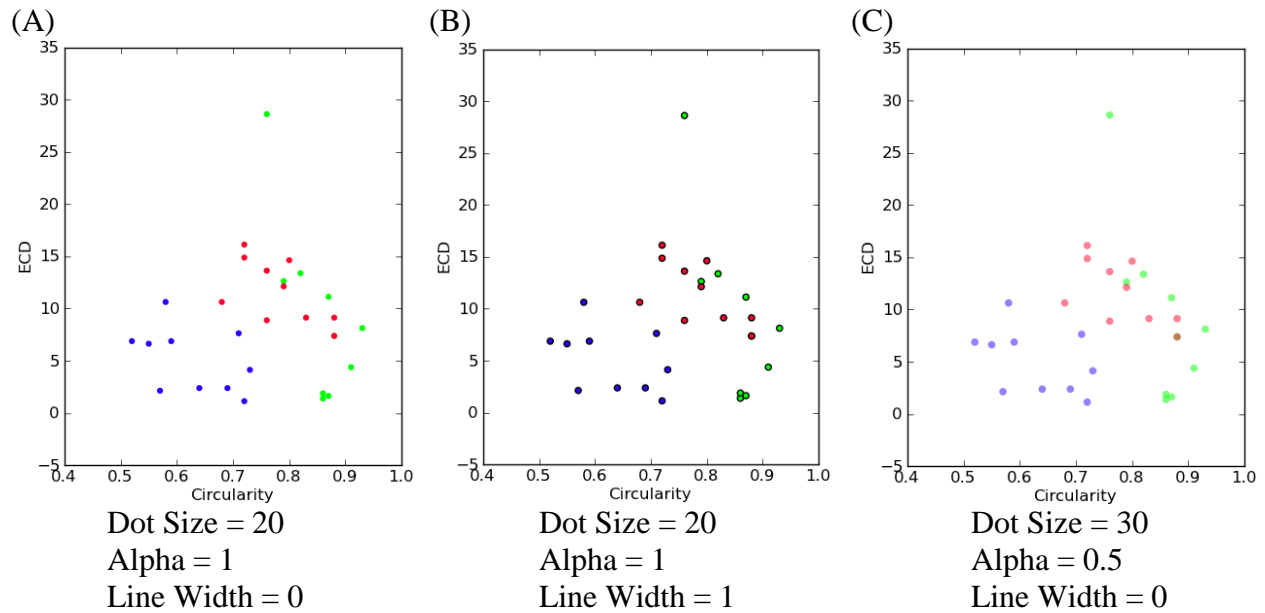


Figure 4.51 Examples of dot options.

4.6.7 Empirical Phase Diagram

The Empirical Phase Diagram menu is used to create a two-dimensional colored diagram as displayed in the Figure 4.52. Each point in the two-dimensional space is represented as a block of color. It is necessary to assign three values to each point because a color consists of RGB components. To construct an EPD, data should have at least three axes, – x, y and color axes. The color axis should contain three values for RGB assignment. Both traditional and three-index EPDs can be constructed using this menu. Traditional EPDs can be constructed by assigning the SVD results to colors, while three-index EPDs are created by assigning structural indices to colors.

Figure 4.52 shows an example of an EPD of BSA. It demonstrates additional features such as overlaying clustering results on the EPD and reference RGB components on the right. Figure 4.53 shows the Options panel for the EPD menu used to assign x, y and color axes, clustering results, and an option to display RGB components.

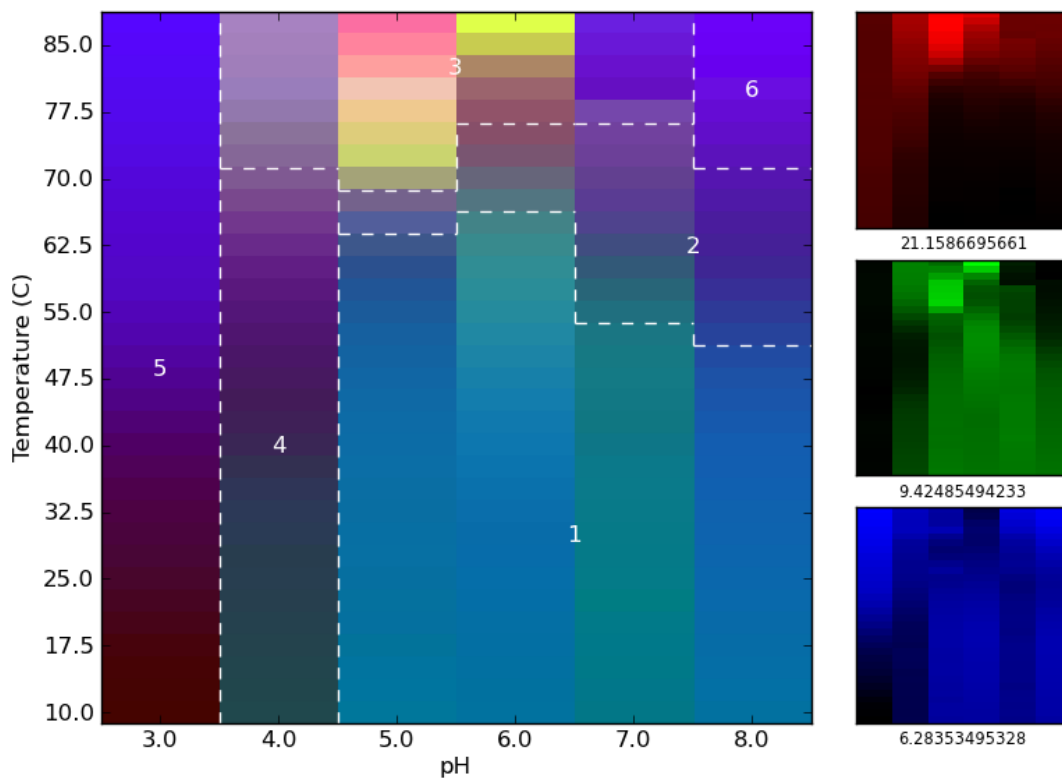


Figure 4.52 Empirical Phase Diagram of BSA with 6 clusters and RGB components shown.

Color Selection

After the x and y axes are selected in the Axes Selection (Figure 4.53A), a color axis should be selected to assign color values to each point in the x and y coordinates (Figure 4.53B). To map a color, 2D values in one color components are normalized between zero and one. This value represents the intensity of the selected color component (i.e. zero for black, one for the maximum intensity of the color). This mapping can be changed using the options available for each RGB component:

- **Min:** If specified, the minimum value for this component is fixed to this value before normalization. Any smaller value is ignored and replaced with this min value.

(A)

Axes Selection

X: pH

pH

3
4
5
6
7
8

Select All

Clear All

Y: Temperature (C)

Temperature (C)

10
12.5
15
17.5
20
22.5
25
27.5
30
32.5

Select All

Clear All

(B)

Color Selection

Color: Singular Value

R: 21.1586695660897

Min:

Max:

Log Scale:

Invert:

G: 9.42485494233044

Min:

Max:

Log Scale:

Invert:

B: 6.28353495327729

Min:

Max:

Log Scale:

Invert:

(C)

Clustering Display

Check if you want to display a clustering result.

- All Data -

Clustering.xlsx

Clustering: Sample

Sample

Clustering

Method

BSA

Color: White

(D)

RGB Components

RGB Components: Show

Figure 4.53 Options panel for EPD menu: (A) Axes Selection, (B) Color Selection, (C) Clustering Display and (D) RGB Components

- **Max:** If specified, the maximum value for this component is fixed to this value before normalization. Any larger value is ignored and replaced with this max value.
- **Log Scale:** If checked, the component values are rescaled with a log function before normalization.
- **Invert:** If checked, the component values are mapped to inverted color intensity after normalization. The value zero will be assigned to the max intensity of the color, while the value of one to black.

Clustering Display

Clustering information can be overlaid on an EPD. The clustering information is clustering numbers with the same x and y axes and their values as EPD data (Figure 4.54). The same numbers indicate the same clusters. This clustering information should be provided as separate MiddaughSuite data and can be selected in the *Clustering Display* as displayed in Figure 4.53C. The clustering information can be obtained as a result from clustering analyses (e.g. k-Means clustering) in the Analysis menu. Color options determine the color of borders and cluster numbers.

RGB Components

Figure 4.53D shows the RGB components option that determines whether an EPD includes RGB components or not. RGB components are three small diagrams of the same x and y dimensions that display only a single color component corresponding to a particular measurement type. This helps understand how each colored component contributes to the EPD by using methods that are specially sensitive to a type of structure (e.g. CD, secondary structure; intrinsic fluorescence,

tertiary structure; light scattering, aggregation). The label of each component diagram is the selected color value in the color axis.

	pH							
Temperature (C)	3	4	5	6	7	8		
10	5	4	1	1	1	1		
12.5	5	4	1	1	1	1		
15	5	4	1	1	1	1		
17.5	5	4	1	1	1	1		
20	5	4	1	1	1	1		
22.5	5	4	1	1	1	1		
25	5	4	1	1	1	1		
27.5	5	4	1	1	1	1		
30	5	4	1	1	1	1		
32.5	5	4	1	1	1	1		
35	5	4	1	1	1	1		
37.5	5	4	1	1	1	1		
40	5	4	1	1	1	1		
42.5	5	4	1	1	1	1		
45	5	4	1	1	1	1		
47.5	5	4	1	1	1	1		
50	5	4	1	1	1	1		
52.5	5	4	1	1	1	2		
55	5	4	1	1	2	2		
57.5	5	4	1	1	2	2		
60	5	4	1	1	2	2		
62.5	5	4	1	1	2	2		
65	5	4	2	1	2	2		
67.5	5	4	2	2	2	2		
70	5	4	3	2	2	2		
72.5	5	3	3	2	2	6		
75	5	3	3	2	2	6		
77.5	5	3	3	3	3	6		
80	5	3	3	3	3	6		
82.5	5	3	3	3	3	6		
85	5	3	3	3	3	6		
87.5	5	3	3	3	3	6		

Figure 4.54 Clustering information used in Figure 4.52

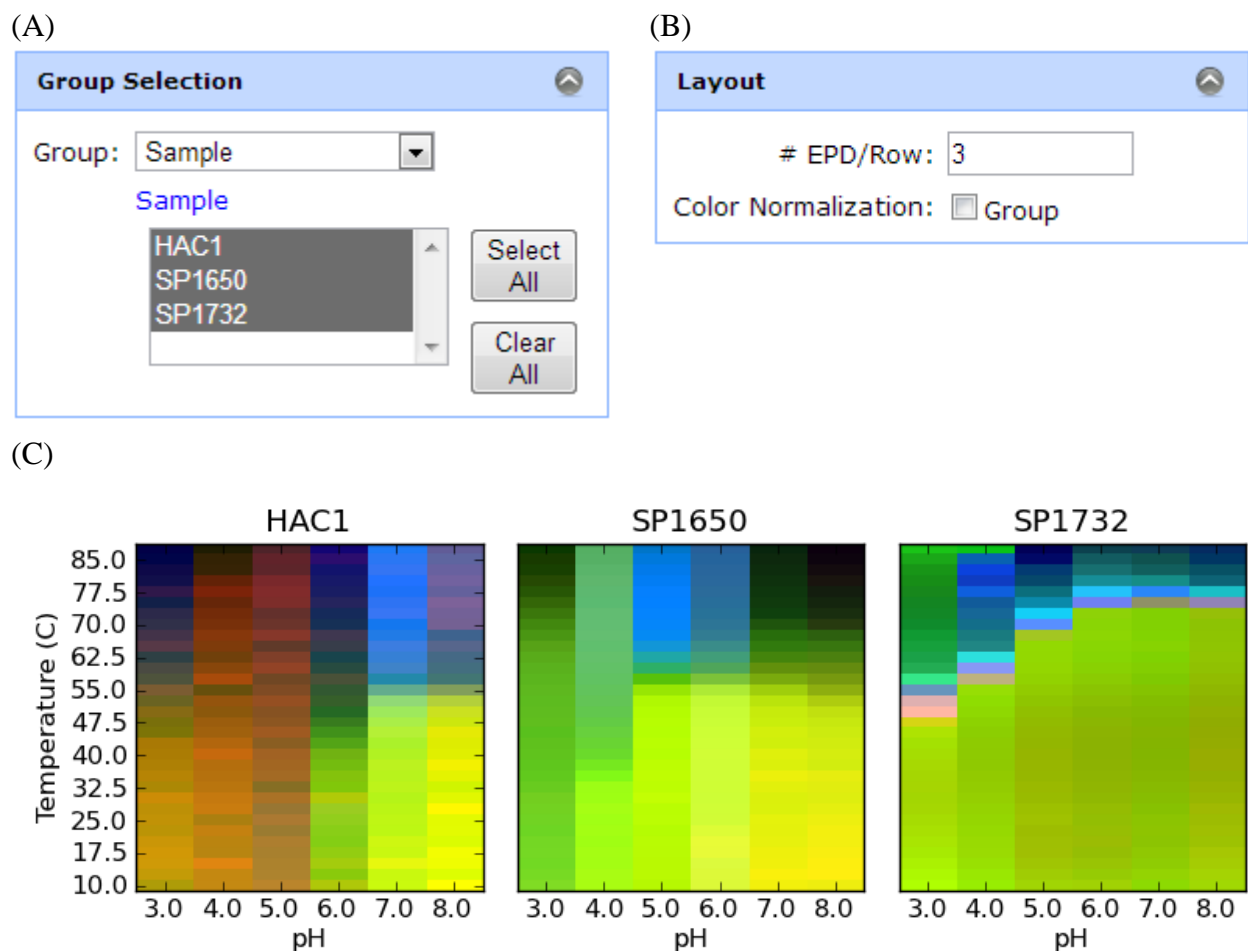


Figure 4.55 Multiple Empirical Phase Diagram menu: (A) Group Selection, (B) Layout in the Options panel and (C) An example of three EPDs created using this menu.

4.6.8 Multiple Empirical Phase Diagrams

The Multiple Empirical Phase Diagram menu is used to create multiple EPDs in a single image. Additional axis should be selected in *Group Selection* (Figure 4.55A) to specify the EPDs. The selected EPDs are placed in a grid whose column has the number of EPDs specified in the *Layout* panel (Figure 4.55B). The Color Normalization option in the *Layout* panel determines whether EPDs share the same color space or not. If selected, each color component value of all EPDs are normalized together and mapped to the same color space. In this case, the same color

regions in any EPD indicate similar structural behaviors. This option is useful when visualizing EPDs from one SVD result from all sample data. If not checked by default, individual EPDs display their own colors.

4.6.9 Radar Chart

The Radar Chart menu is used to create radar charts placed in a two-dimensional grid. Each block in the two-dimensional grid contains a radar chart that visualizes multiple values in the method axis. MidaughSuite supports ten values in the method axis as displayed in Figure 4.56A.

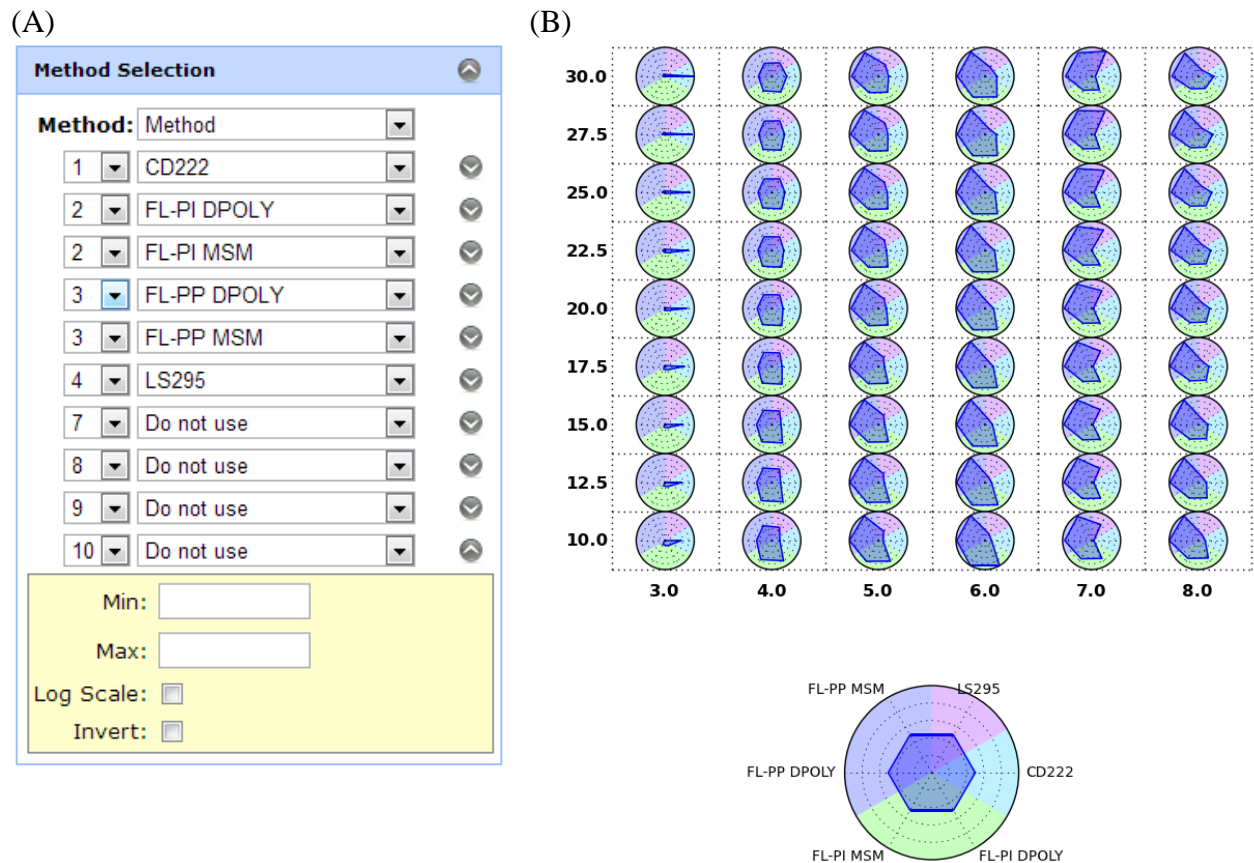


Figure 4.56 (A) Method Selection in the Options panel and (B) an example Radar chart

Each method value has the same options of *Min*, *Max*, *Log Scale*, and *Invert* as the EPD menu. Each option also works similarly to those in the EPD menu. Each method value is normalized between zero and one and mapped to the equiangular axis in a polar coordinate system. *Min* and *Max* values and the *Log Scale* option are applied before normalization. The *Invert* option reversely maps values to the axis (i.e. value one to center, value zero to radius) if selected.

MiddaughSuite supports a group coloration feature. Multiple methods can form groups and each group share the same background color. Grouping can be done by assigning the same number to multiple method values. Figure 4.56B demonstrates an example of group coloration. Two method values, FL-PI DPOLY and FL-PI MSM are mapped to group 2 and FL-PP DPOLY and FL-PP MSM to group 3. The other single values such as CD222 and LS295 are assigned to group 1 and 4, respectively. If every group contains only a single member, the group coloration feature is deactivated and every group has a white background.

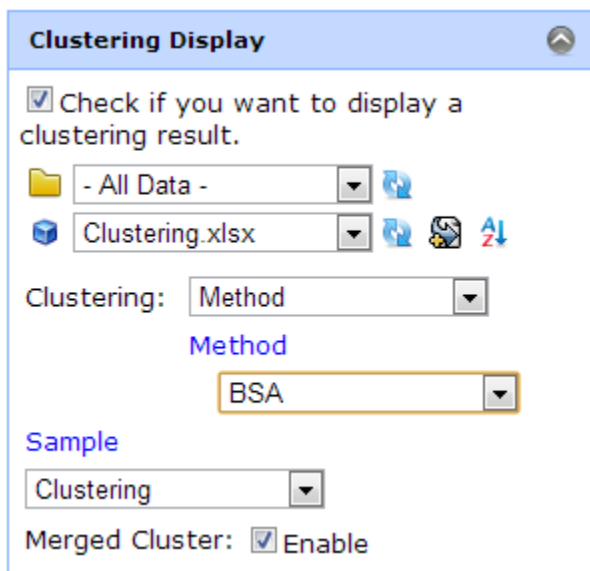


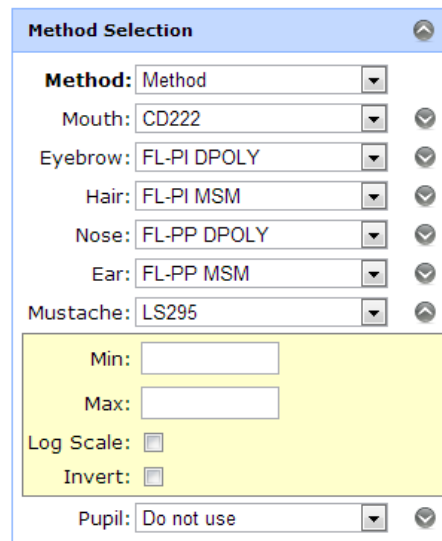
Figure 4.57 Clustering Display in the Options menu for Radar Charts and Chernoff Face Diagram

MiddaughSuite supports the *Merged Cluster* option in the Clustering Display for both radar charts and Chernoff face diagrams as displayed in Figure 4.57. If selected, every point inside a cluster is averaged and the averaged values in each method are used to draw a representative iconic feature for the cluster.

4.6.10 Chernoff Face Diagram

The Chernoff Face Diagram menu is used to create Chernoff faces placed in a two-dimensional grid. Each block in the two-dimensional grid contains a Chernoff face that visualizes multiple values in the method axis as a facial feature. MiddaughSuite support seven facial features in the method axis as displayed in the Figure 4.58.

Each method value has the same options of *Min*, *Max*, *Log Scale*, and *Invert* as the EPD and Radar Charts menu. Each method value is normalized between zero and one and mapped to facial features. *Min* and *Max* values and the *Log Scale* option are applied before normalization. The *Invert* option reversely maps values to the facial features if selected.



The image shows a dialog box titled "Method Selection" with a blue header and a light blue border. It contains several dropdown menus and checkboxes. The "Method" dropdown is set to "Method". Below it are seven facial features: Mouth (CD222), Eyebrow (FL-PI DPOLY), Hair (FL-PI MSM), Nose (FL-PP DPOLY), Ear (FL-PP MSM), and Mustache (LS295). Each feature has a dropdown menu and a checkmark icon to its right. Below these is a yellow-shaded area containing "Min:" and "Max:" text labels with empty input boxes, "Log Scale:" with an unchecked checkbox, and "Invert:" with an unchecked checkbox. At the bottom is a "Pupil:" dropdown menu set to "Do not use" with a checkmark icon to its right.

Figure 4.58 Method Selection in the Options panel for Chernoff Face Diagram

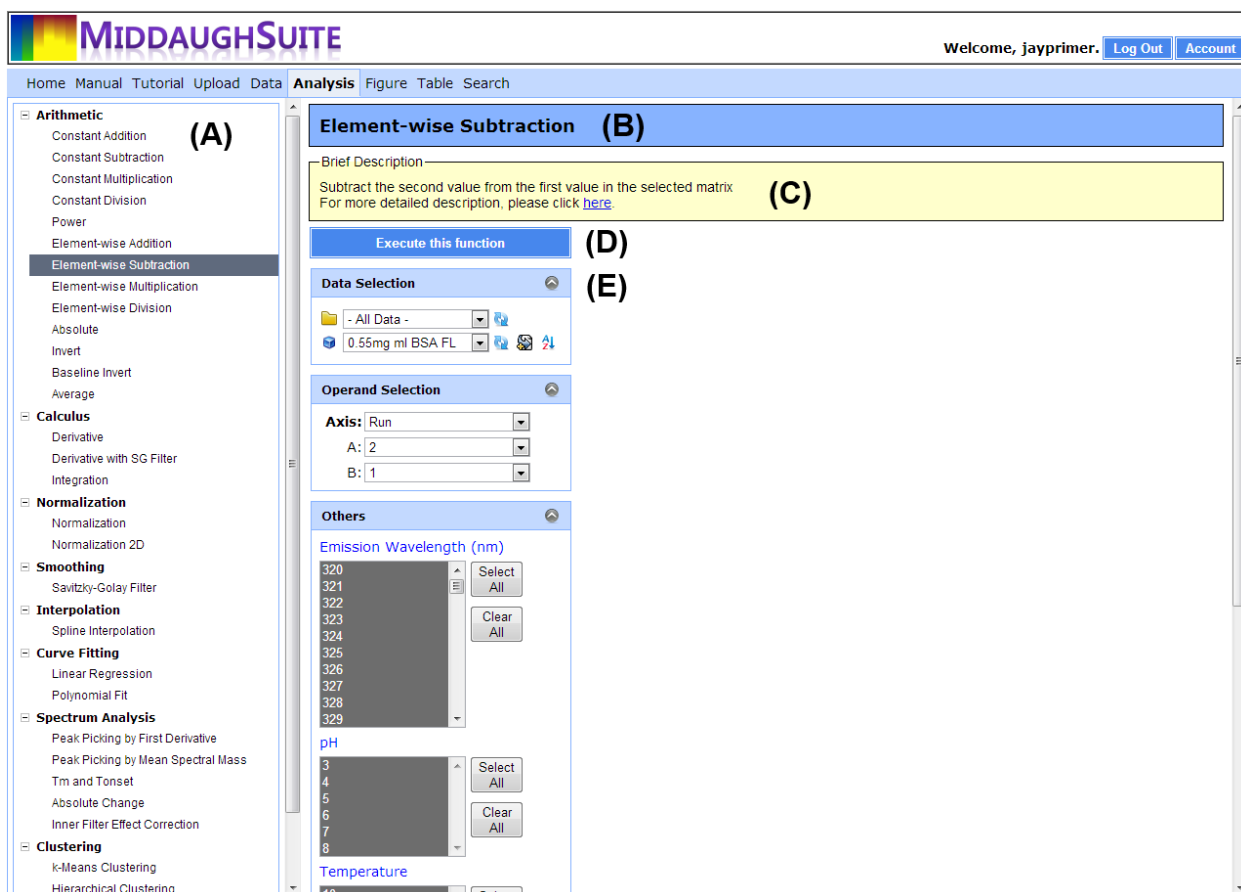


Figure 4.59 Analysis menu screen: (A) Functions list, (B) selected function name, (C) brief description of selected function, (D) “Execute this function” button, and (E) data, value range, and options panel for the selected function.

4.7 Analysis Menu

The *Analysis* menu provides various analysis functions that are applicable to MiddaughSuite data. Every function is designed to efficiently handle multidimensional data. MiddaughSuite currently supports 31 functions in 9 categories. The *Analysis* menu screen as displayed in Figure 4.59 consists of five components – functions list, function name panel, description panel, “Execute this function” button, and data and option selection panels. The functions list displays available

functions. One function in the list can be selected and the information and input panels for the selected function are displayed in the right screen. Below the name and description of the selected function, arguments for the function should be selected. Each function has its own data and option selection panels. After data and options are selected, clicking the “Execute this function” button will initiate execution of this function. The result will be always in the form of a multidimensional matrix and will be automatically added to the data list.

4.7.1 Element-wise Subtraction

The *Element-wise Subtraction* function is useful for subtraction of buffer data from sample data. It works similarly to matrix subtraction. The difference is that this function requires arguments of two values in one axis in one matrix instead of two matrices of the same size. This requirement automatically fulfills the restriction of matrix subtraction that two arguments should be the same size. Users should prepare both sample and buffer data under identical conditions and merge them into one matrix.

Figure 4.60A shows an example of fluorescence data that contains triplicate data from both protein and buffer under a wide range of pH and temperature conditions. After this data is selected for the *Element-wise Subtraction* function, both protein and buffer in the sample axis should be selected in the *Operand Selection* panel as displayed in Figure 4.60B. The range of values in the remaining axes can be specified in the *Others* panel. In this manner, it is possible to specify the two arguments required for matrix subtraction.

In many cases, it would be difficult to get the same amount of buffer data as the protein data. For such cases, buffer data can be duplicated to fill the other conditions before they are merged with sample data.

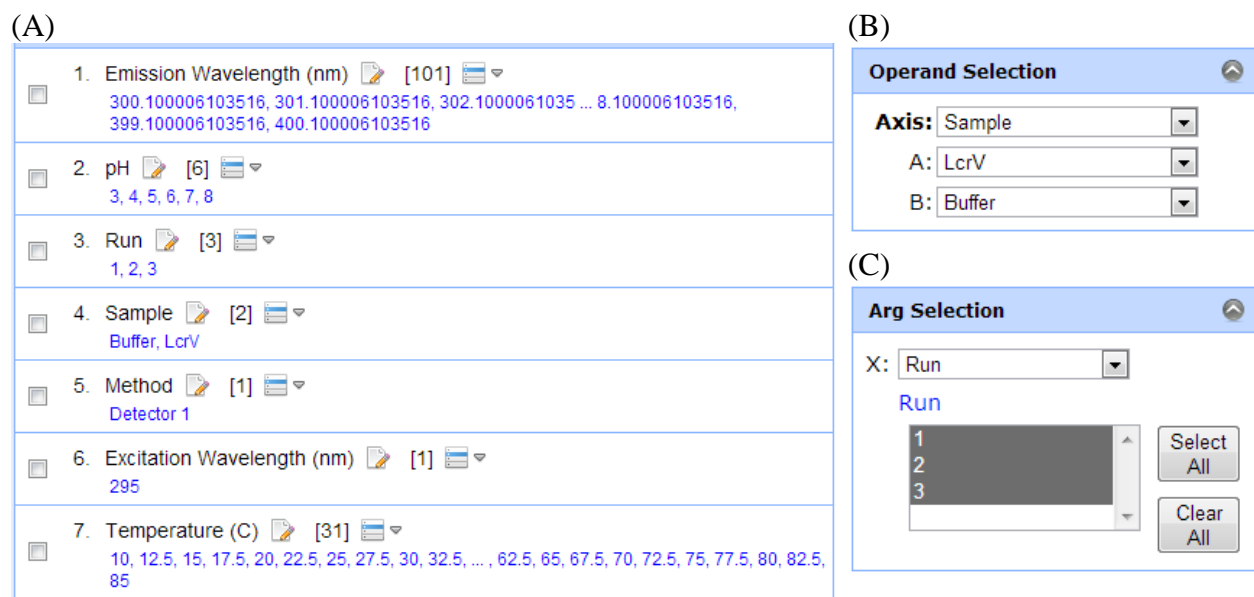


Figure 4.60 (A) Example of Fluorescence Data for Buffer Subtraction and Average functions, (B) Operand Selection for Element-wise Subtraction. This function calculates $A-B$. (C) Arg Selection for Average. This function calculates the average and standard deviation of all selected values.

4.7.2 Average

The *Average* function is used to calculate an average and standard deviation from multiple runs of experimental data. Similar to the *Element-wise Subtraction* function, the *Average* function requires only one matrix. Multiple values in one axis in the input matrix should be selected in the *Arg Selection* panel (Figure 4.60C) and their average and standard deviation will be calculated.

The result matrix contains the same dimensions and values as the input matrix except the averaged axis. The averaged axis will have ‘Avg’ and ‘Std’ values instead of the original values, which represent the average and standard deviation, respectively. This axis and values can be used to draw error bars in the figures.

4.7.3 Normalization and Normalization2D

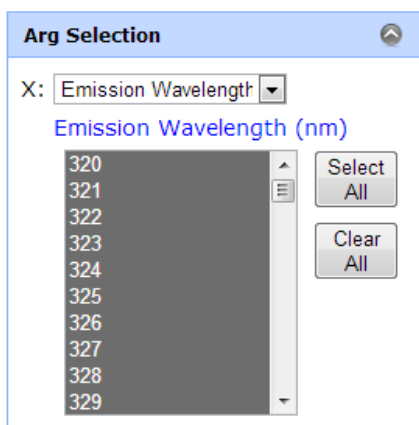
The *Normalization* and *Normalization2D* functions are used to normalize data. The *Normalization* function takes one dimensional data for normalization. In the fluorescence data example in Figure 4.60A, one must select “Emission Wavelength (nm)” axis as a reference axis to normalize spectra under all conditions (Figure 4.61A). Figure 4.61C shows normalized spectra from 10°C to 87.5°C.

The *Normalization2D* function takes two axes as arguments (Figure 4.61B). These two dimensional data in all other conditions are normalized together. As shown in the Figure 4.61D, the “Temperature (C)” as well as the “Emission Wavelength (nm)” axis is selected as arguments and the relative intensity of temperature-dependent spectra is preserved.

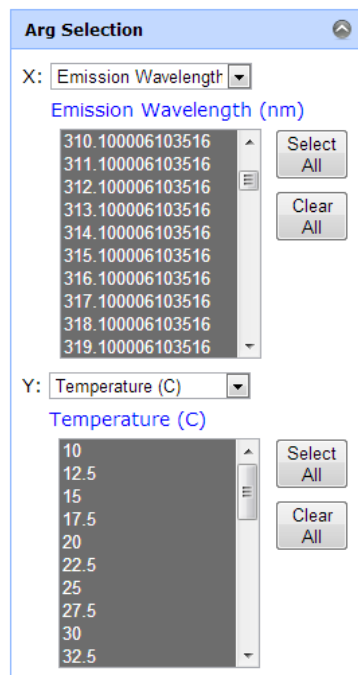
4.7.4 Savitzky-Golay Filter

The Savitzky-Golay Filter is a widely used smoothing function. It requires two parameters – window size and order. The filter performs a local polynomial regression. The parameter order is used to determine the order of the polynomial function used in the regression. The window size determines the number of points used for the regression. The regression is used to find a smoothed value of the center point of the window. Therefore, the window size should be an odd number. The regression is applied to all sample points as the window moves.

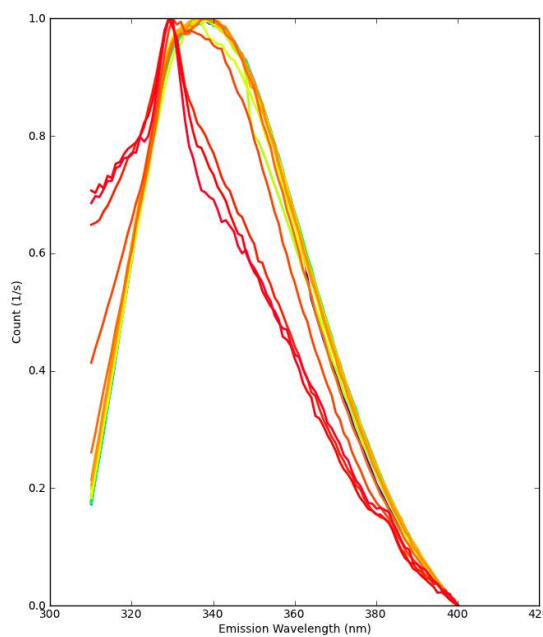
(A)



(B)



(C)



(D)

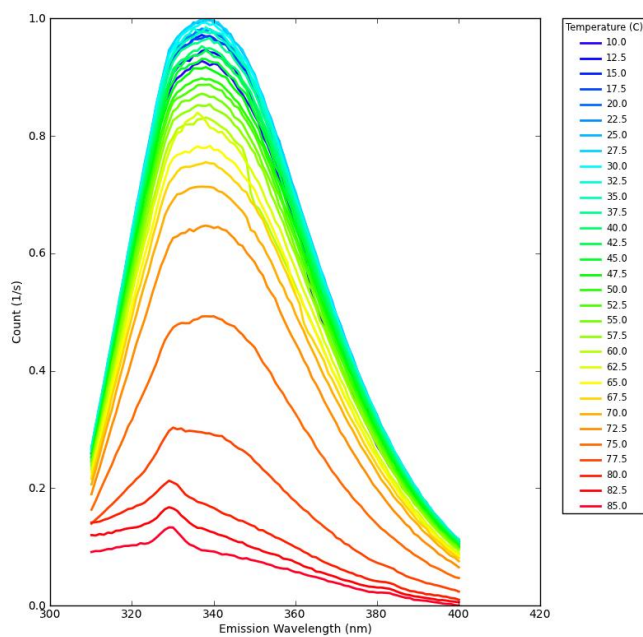


Figure 4.61 (A) Arg Selection for Normalization, (B) Arg Selection for Normalization2D, (C) Example of normalized fluorescence spectra, and (D) Example of normalized fluorescence spectra in which the relative intensity of the spectra as a function of temperature is preserved using Normalization2D.

4.7.5 Spline Interpolation

The *Spline Interpolation* function is used to replace values in one axis with others while data is preserved. For example, one set of data is collected with temperature from 10°C to 90°C with 1°C increments and the other sets with temperatures from 10°C to 87.5°C with 2.5°C increment. For operations such as merging data, it is required that two different temperature axes should be matched while the shape of the data is preserved. The *Spline Interpolation* function takes two axes as displayed in the Figure 4.62. The data on the left is recalculated with the axis values given on the right. The result will have the same axis values as the target axis.

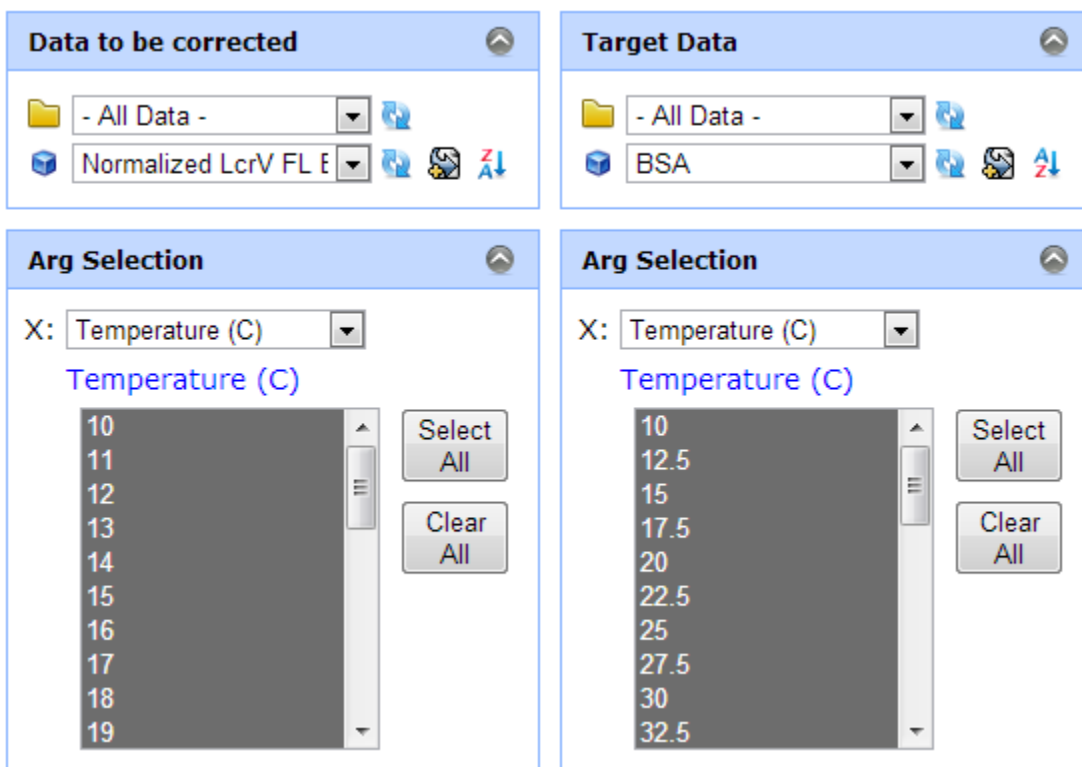


Figure 4.62 Data Selection for Spline Interpolation function

4.7.6 Peak Picking by Mean Spectral Center of Mass

The *Peak Picking by Mean Spectral Center of Mass (MSM)* function is used to calculate the spectral centroid as an estimate of the intensity (area) and wavelength position of a spectral peak. This analysis is widely used for intrinsic fluorescence spectra from tryptophan residues in proteins. The emission spectrum from tryptophan displays different spectral properties due to the polarity of local environment surrounding the indole side chain. If tryptophan is located in an apolar environment, the emission peak is found below 340nm. When the peak is shifted up to 340-355nm, it is generally considered to be more exposed to polar environments such as the aqueous solvent. The sensitivity of the Trp emission spectrum originates from the selective activation of two isoenergetic transitions in the indole side chain. In contrast, tyrosine is selectively insensitive to its local environment polarity because it has a single electronic state.

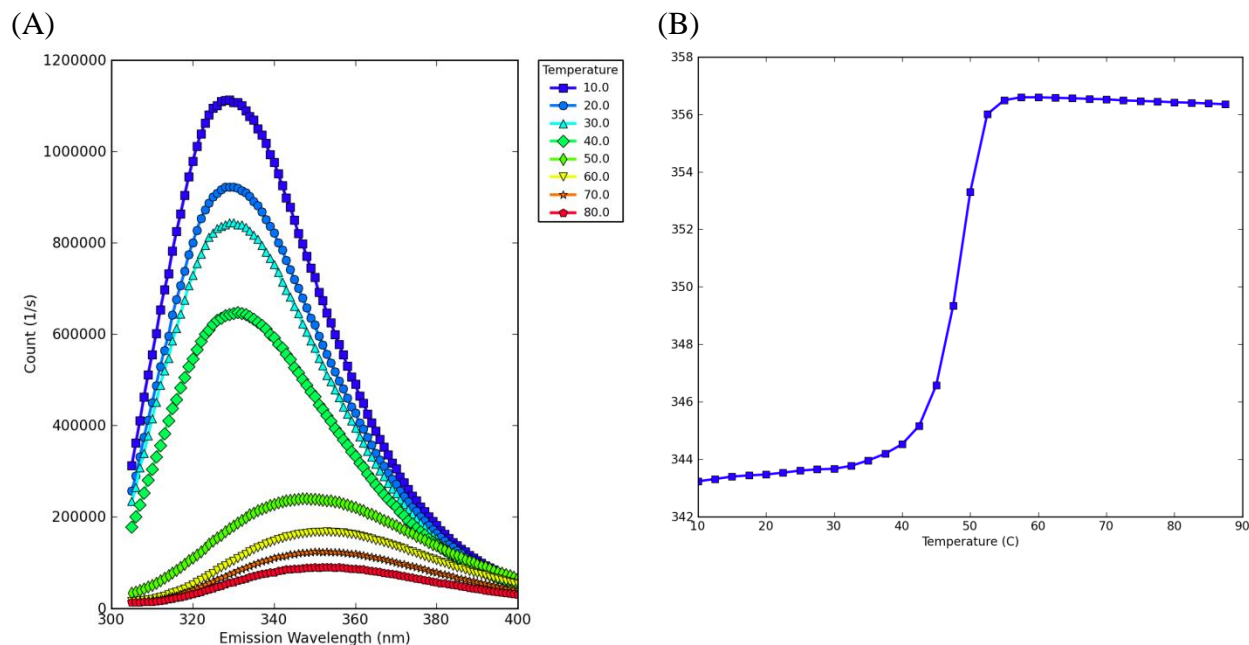


Figure 4.63 (A) Trp fluorescence spectra of Chymotrypsin at pH 7 (B) Peak Shift by Mean Spectral Center of Mass

Figure 4.63A shows an example of the tryptophan fluorescence emission spectra of Chymotrypsin at pH 7 from 10°C to 80°C. The peak position of each spectrum is calculated using the MSM function. It should be noted that the calculated peak is not an actual peak maximum. Rather, it reflects the mean of spectral energy. In many cases, the MSM method works much better than identifying the actual maximum because it is less sensitive to noise. It is common for macromolecules to contain multiple tryptophan residues and the observed fluorescence emission is the sum of the individual emission spectra.

4.7.7 k-Means Clustering

The *k-Means Clustering* function is used to find clusters (physically-associated) data. In the *k*-means clustering algorithm, the number of clusters is selected by a user. All sample points are partitioned into the given number of clusters in which each point belongs to the nearest cluster based on the distance from the center of the cluster. The algorithm tries to minimize sum of distances from all points and iteratively updates the location of the cluster centroid.

Figure 4.64 shows input panels for the *k-Means Clustering* function. The input axes are used to transform the data into a two-dimensional matrix as an input for the *k*-Means Clustering algorithm. The axes X, Y and Sample are combined together to form rows while the method axes form the columns of the input matrix. The number of clusters is not an input parameter since this function varies with the number of clusters from 2 to 10. The axes X, Y and Sample are reconstructed in the result, while the method axes are replaced with the clustering results.

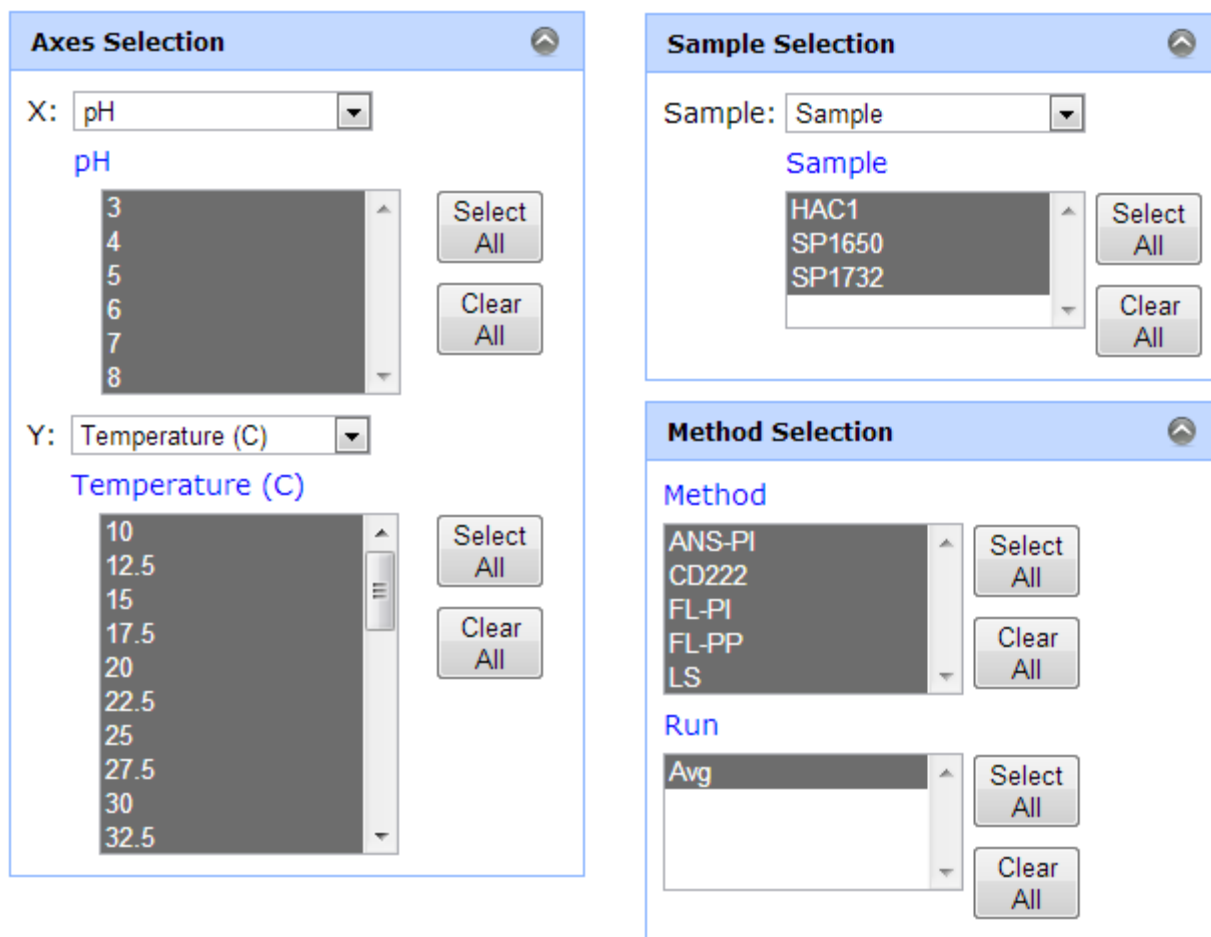


Figure 4.64 Input panels for k-Means Clustering and Singular Value Decomposition

4.7.8 Singular Value Decomposition

The *Singular Value Decomposition* function is used to calculate the singular value decomposition of the input data. The singular value decomposition of the input matrix M is a factorization of the following form:

$$M = U \times \Sigma \times V^T$$

where U , V are unitary matrices and Σ is a diagonal matrix with singular values on the diagonal. This function returns the matrix U with singular values as the axis values. The matrix V is ignored.

This function uses the same input panels and a method to construct the input matrix as the *k-Means Clustering* function. The axes X , Y and $Sample$ are combined together to form rows while the method axes form columns of the input matrix M . The axes X , Y and $Sample$ are reconstructed in the result matrix U .

4.8 Table Menu

The *Table* menu provides functionality to view numerical data in the form of two dimensional tables and to download the generated tables as an Excel file. The *Table* menu screen as displayed in the Figure 4.65 consists of two panels – the input to the left and the output to the right. The input panel consists of data, sheet, axes selection and the “View” button. The Table menu can show three dimensional data as a series of two dimensional tables. The axes selection determines the row and column of the two dimensional table, while the sheet selection decides in which sheet the two dimensional table resides. The generated tables can be organized into an Excel file that contains multiple sheets in each of which a two-dimensional table is located. Clicking the “Download as Excel file” button in the output panel starts downloading the Excel file.

Figure 4.65 Table menu screen: (A) Sheet Selection, (B) Axes Selection, (C) “View” button, and (D) Table view and “Download as Excel file” button

4.9 Search Menu

The Search menu provides functionality to search importable projects from other users in the MiddaughSuite. The importable projects are the projects whose “Share With” option in the project detail panel is set to “Staff Only” or “Public”. If the project is set to “Staff Only”, only users that are marked as a staff or a super user in the MiddaughSuite system can search the project. If it is set to “Public”, any users in the MiddaughSuite can search the project. The searched projects can be browsed and imported to your account.

The screenshot displays the MIDAUGH SUITE Search menu. At the top, the user is logged in as 'jayprimer'. The search bar (A) contains the text 'aldolase FL'. The search results table (B) shows one entry: 'Aldolase' by 'jayprimer' on '13-02-05 10:10:55'. The project detail panel (B) shows 'Aldolase' with a description, attached file, and share settings. The data list panel (C) lists 'aldolase PTL.zip', 'aldolase CD, ABS', 'aldolase FL' (selected), and 'aldolase LS'. The data detail panel (D) for 'aldolase FL' shows a grid of fluorescence spectra, a list of dimensions (Emission Wavelength, Method, Excitation Wavelength, Temperature, Sample, pH, Run), and a history log of changes.

Figure 4.66 Search menu screen: (A) Search panel, (B) Project Detail Panel, (C) Data List Panel, (D) Data Detail Panel, and (E) “Import This Project To Your Account” button

Figure 4.66 shows the Search menu screen. The screen consists of six components – search panel, project detail panel, data list panel, data detail panel and the “Import This Project To Your Account” button. You can search projects using the search panel. If a searched project is selected, the information about the project is displayed in the project detail, data list, and data detail panels. These panels are identical to panels in the *Data* menu. Clicking the *Import* button initiates copying the project data into your account.

Chapter 5. Conclusion

The pharmaceutical market is expected to continuously grow due to the recent dramatic success of biopharmaceutical products. This growing market is driving more rapid development of candidate biologics. The size and complexity in structure of biologics, however, makes the molecules more susceptible to chemical or structural changes leading to lower potency or altered immunogenicity. Sustaining the stability of the macromolecule becomes one of the major challenges in the future development of biopharmaceutical products.

One of the most common approaches to improve the stability of a macromolecule is to find a proper composition of inactive additives (so called “excipients”) that help stabilize the macromolecule. A better understanding of macromolecular behavior should proceed to the initiation of the screening process for stabilizing compounds to help narrow the large number of available excipients and experimental conditions that need to be explored.

Techniques such as X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy and hydrogen deuterium exchange mass spectroscopy (HDX-MS) provide detailed structural information about a macromolecule but are generally not available or too time-consuming or complex for routine and frequent use. Classical biophysical techniques such as UV absorbance, circular dichroism (CD), fluorescence, light scattering, and Fourier transform infrared (FTIR) spectroscopy provides less detailed information about protein structure but a combination of all of their information may be sufficient to study structural changes under many circumstances at adequate resolution.

The empirical phase diagram (EPD) method was developed as a convenient method to summarize experimental data from multiple techniques and to visualize macromolecular behavior into an intuitive colored diagram. It can provide a global picture of structural changes in a macromolecule over a wide range of experimental conditions. The method employs singular value decomposition (SVD) to extract major patterns from a large collection of data. The extracted patterns are visualized as color changes in a two dimensional space of environmental stresses (e.g. temperature and pH).

The current EPD method suffers from a number of deficiencies that primarily reflect the use of color to represent the state of the protein's structural integrity. This includes a lack of meaningful relationship between color itself and actual molecular features as well as limitations resulting from color deficiencies in vision and blindness which are possessed by a substantial portion of the human population.

In chapter 2, three data visualization approaches were introduced – the three-index EPD, radar charts, and Chernoff face diagrams. These methods are designed to overcome the difficulties discussed above.

The three-index EPD method displays secondary, tertiary, and quaternary structural changes with a pre-defined color scheme. The overall structural changes of a macromolecule can be displayed with fixed colors. A color yellow represents the native state of a macromolecule and the color blue an aggregated state. A darker color, close to black, represents a maximally altered conformational state without any notable aggregation. These colors are considered the major indicators, while other colors such as brown and green represent partially altered states depending on the differentially reduced color levels of tertiary and secondary structure.

The radar chart and the Chernoff face diagram are similar to each other in that iconic features are designed to reflect structural characteristics of macromolecules and placed in a two-dimensional grid of environmental stresses. The changes in iconic features as a function of environmental stress represent the overall macromolecular response to the given stresses. The advantage of these diagrams over the original EPDs is that these diagrams do not rely on colors and can display more than three kinds of data and thus more complex elements of structure. It is, however, difficult to read an exact value or to detect subtle changes in values with these two approaches, especially with Chernoff face diagrams.

The biophysical characterization of macromolecules using the various kinds of visualization techniques discussed above requires a large volume of experimental data. Recent development of high-throughput and multi-functional instruments enables such large-scale data collection while it greatly reduces the overall cost, time and labor. However, current data analysis procedures supplied with instruments cannot match the data generation speed due to the lack of dedicated software.

The need for analysis software that can easily combine data from various instruments, organize a large volume of data, quickly apply mathematical functions, visualize the result with a number of different graphs and diagrams including EPDs has been greatly increased. Such software is also desired that provides a way to share accumulated data among researchers. These requirements are listed and discussed in the Chapter 3.

MiddaughSuite is web-based analysis software for biophysical characterization. It is developed in this work to meet the user requirements described above. MiddaughSuite is developed as a website rather than stand-alone software. The website has many advantages over

stand-alone software. It is designed to support a group of researchers in individual laboratories and beyond. Users do not need to manage individual software and can gain easy access to MiddaughSuite from any remote computers using various kinds of browsers. The centralized data server can store all data from a laboratory. The stored data can easily be shared among researchers. Recent development in dynamic HTML technology makes it possible to incorporate interactive components in web pages. The interactive components provide a way to make an intuitive and easy-to-use interface for users. Because the graphical user interface of software determines the quality of a user's experience with the software, it is important to design a high quality user interface.

The web user interface of MiddaughSuite is described in Chapter 4. MiddaughSuite has six main menus including *Upload*, *Data*, *Figure*, *Analysis*, *Table*, and *Search*. The *Upload* menu is used to upload data files from various instruments. It also supports a user-supplied Excel file so that users can upload their own data or data from currently not supported instruments. The *Data* menu is used to manage uploaded data. It supports projects and tags for users to easily categorize data. The *Figure* menu is for visualization of data using various kinds of graphs and diagrams. It supports presentations such as line and bar graphs as well as the original EPDs, three-index EPD, radar charts, and Chernoff face diagrams. The *Analysis* menu is used to apply mathematical functions to data. Currently 31 functions in 9 categories are supported in the *Analysis* menu. These functions are designed for application to multidimensional matrices. If more data is merged to form a larger multidimensional data, the analysis functions can be applied more efficiently. Using the *Table* menu, actual numerical values in the data can be extracted and

downloaded in the format of Excel files. Finally, users can share their projects. In the *Search* menu, other users can search and import projects that are marked to be shared.

In conclusion, MidaughSuite combined with high-throughput multimodal spectrophotometers makes it possible to obtain various types of EPDs rapidly, possibly within a single day, with a minimal amount of protein and effort. The MidaughSuite system will be continuously upgraded to support more instruments, functions and diagrams and promises significant aid in the development of protein pharmaceuticals and vaccines.