Trinity University
Digital Commons @ Trinity

Computer Science Honors Theses | Computer Science Department

4-20-2011

Applying the Diversity Map, a Visualization Technique, to the Protein Data Bank

Onyekwere Maduka Ogba
Trinity University, oogba@trinity.edu

Follow this and additional works at: http://digitalcommons.trinity.edu/compsci_honors

 Part of the Computer Sciences Commons

Recommended Citation
Ogba, Onyekwere Maduka, "Applying the Diversity Map, a Visualization Technique, to the Protein Data Bank" (2011). *Computer Science Honors Theses*. 27.
http://digitalcommons.trinity.edu/compsci_honors/27

This Thesis open access is brought to you for free and open access by the Computer Science Department at Digital Commons @ Trinity. It has been accepted for inclusion in Computer Science Honors Theses by an authorized administrator of Digital Commons @ Trinity. For more information, please contact jcostanz@trinity.edu.

Applying the Diversity Map, a Visualization Technique, to the Protein Data Bank

Onyekwere Maduka Ogba

Abstract

In this research, the Diversity Map, a visualization technique created in the Metoyer research lab in Oregon State, is used to visualize the diversity of all the molecules deposited in the Protein Data Bank. Data was split into a period of three years from 1990 to 2010 and analyzed individually by Dr. Laura Hunsicker-Wang in the department of chemistry with a research focus in protein, and ten student participants with varying level of knowledge in chemistry. Results show that the Diversity Map is a powerful tool in understanding data from the Protein Data Bank and furthermore, has strong potential in being useful to the scientific community.

Acknowledgements

I would like to thank Dr. Ronald Metoyer for inviting me to work with him as a summer research assistant in his lab at Oregon State University. This thesis would not have happened if I was not there. I want to thank Tuan Pham, a member of Metoyer's lab, who helped me in understanding the background code of the Diversity Map program. I'd like to thank Dr. Hunsicker-Wang for being the expert advisor and participant of my informal case study. I thank the 10 student participants who sacrificed time to be an integral part of my study. Finally, I want to thank Dr. Mark Lewis for being my thesis advisor, as well as my thesis committee members, Dr. John Howland and Dr. Hunsicker-Wang.

Applying the Diversity Map, a Visualization Technique, to the Protein Data Bank

Onyekwere Maduka Ogba

A departmental thesis submitted to the

Department of Computer Science at Trinity University

in partial fulfillment of the requirement for Graduation

April 20, 2010

_____ _____

Thesis Advisor Departmental Chair

_____

Associate Vice President

for

Academic Affairs

This thesis is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs License, which allows some noncommercial copying and distribution of the thesis, given proper attribution. To view a copy of this license, visit http://creativecommons.org/licenses/ or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.

Applying the Diversity Map, a Visualization Technique, to the Protein Data Bank

Onyekwere Maduka Ogba

TABLE OF CONTENTS

1 Introduction

1.1 Protein Data Bank

The Protein Data Bank was established in 1971 as a joint effort of the Brookhaven National Laboratory and the Cambridge Crystallographic Data Centre. The goal of the creators was to bring together all proteins for easy and efficient access to everyone. The database began with 13 structures by the end of 1974 with a slow and steady increase in size until a boost in the numbers in the late 1980s (Kirchmair et. al 2008). The boost is attributed to the increase in computational power, as well as the establishment of structural genomic initiatives (Kirchmair et. al 2008). The success of this software, like any other application, is heavily dependent of the power of the hardware (i.e. - how much memory is present to accommodate the growing amount of data). As of January 4th, 2011, the protein data bank website states that there are currently 70,303 structures deposited in the database.

A valid entry into the database requires the scientist to input structural and functional information about the protein molecule as well as the experimental technique used to discover the protein strand. This information must be in the format deemed appropriate by the developers of the online bank. In early stages of the Protein Data Bank, less emphasis was placed on standardizing these submitted entries (more flexible options were left to the submitting parties), hence causing the data set to not conform to each other. For example, one scientist might enter *E*

Coli as the source of a protein, while another may enter *Escherichia Coli* as the source of another protein. These two entries, scientifically, refer to the same thing. However, computationally (by strict string comparison), the two entries will be classified as two different sources. In recent years, standardization became a concern to the developers of the Bank (Kirchmair et al 2008). One of the challenges faced would be to cure the already non-standardized data in the database since these initiatives began after the institution of the Protein Data Bank and then to implement new technologies to ensure that current and future additions to the database are standardized. One standardization technique is to change some of the input fields from plain text fields to a more rigid entry form like a check box with the list of possible entries, hence making that input field uniform and ultimately making analysis of the Protein Data Bank more feasible.

1.2 Diversity Map

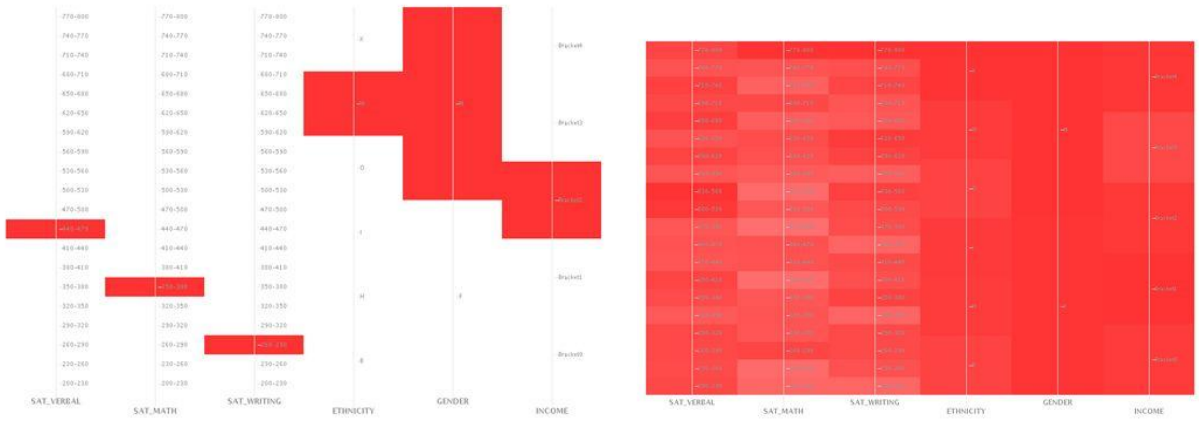

Fig. 1. Data set with (a) Low Diversity and (b) High Diversity using the Diversity Map Representation

The Diversity Map is a visualization tool created and designed by Metoyer and his research group in Oregon State University for the purpose of visualizing the diversity of large multivariate data sets (Pham et. al 2010)**.** Each attribute is represented as a column on the map and each value on the column is represented as a semi transparent rectangular block, which will be referred to as a bucket. The number of objects in a bucket is denoted by its degree of color intensity. Therefore, a darker shaded bucket signifies that a larger number of objects are in that particular bucket and vice-versa for a lighter shaded bucket. Furthermore, the shade evenness of a column denotes the level of diversity of that column. This evenness shows that the values on the column are distributed across the buckets. Using this definition, figure 1a can be classified as a set of objects with lower diversity than that of figure 1b.

2 Previous Research

Although, many tools have been developed to analyze the Protein Data Bank, most of them are focused on analyzing one protein structure in the database at a time. Programs like iMolTalk and Protein Structure Analysis Package analyze PDB complexes, including their 3D molecular structure, a visualization of their Ramachandran plots, and information of their intra- and inter interactions (Kirchmair et. al 2008). In addition, there are other kinds of programs that primarily screen the protein database in search of particular answers. For example, a computational method was developed to screen the data bank for "putative pockets possessing a specific binding site chemistry and geometry" (Campagna-Slater et. al 2010). SuMo (Surfing The Molecules) screens protein structures and substructures to detect similar binding sites of another protein (Kirchmair

et. al 2008). None of these applications deal with the level of abstraction that the Diversity Map is able to offer.

An interesting new visualization tool that seems to have as high a level of abstraction as the Diversity Map is BioGenet. This tool creates, visualizes and analyzes biological networks. This functional connection between proteins is made by the BioGenet server which detects protein-protein interactions between molecules (Martin et. al 2010). Although we get similar levels of abstraction, BioGenet does not answer the question of diversity, in terms of richness and evenness that the Map does. However, there is a growing desire to understand biological and molecular diversity. One of Dr. Metoyer's research with the Diversity Map was to analyze a Moth civilization in a defined ecological field in Oregon State University. Moreover, there is an entire journal dedicated to molecular diversity.

Proteins are adept to being unique. They are made up of amino acids (life's building blocks) that are coded by DNA sequences, which are made up of nucleotides. A single change in nucleotide can cause a change in a DNA sequence, which causes a change in amino acids. A slight variation in the DNA sequence can drastically cause a difference in protein structure and function (Nedelkov 2008). These changes in nucleotides can and do occur and hence, lead to different kinds of proteins in the human body. Therefore, it is necessary that we understand the nature of human protein diversity.

3 Experimental Method

3.1 Preliminary Meeting

To determine an appropriate use of the Diversity Map in chemistry, it was important to the research to find an expert in that field to guide the entire process. I contacted Dr. Hunsicker-Wang, a professor of chemistry at Trinity University in which one of her research areas focuses on Protein analysis, to be an advisor as well as the participant of the case study in analyzing the visualization technique. Preliminary meetings with Dr. Hunsicker-Wang were for the purpose of figuring out an appropriate application for the Diversity Map. The goal was to ensure that this study would be useful to the scientific community as well as easily streamline with the Map. For the application to be easily streamlined, it would require a biological system that contains a large number of objects with set attributes, and set values for each attributes. Using these guidelines, we decided that the Protein Data Bank was an appropriate database to apply the Diversity Map.

Once the decision was made to use the Protein Data Bank, Dr. Ronald Metoyer and his research group at Oregon State University were contacted in order to decide on the attributes for each protein that we wanted to study. Each of these attributes needed to have specific and recorded values. By definition, the Diversity Map "maps" each object with its attributes denoted by a column on the visualization screen using its specific attribute value.

3.2 PDB Attributes Used

Dr. Hunsicker-Wang and I concluded that these attributes most efficiently explained the structural information of each of the proteins in the Data band and were suitable in studying the diversity of the Protein Data Bank.

- Resolution
- Deposition Date
- Structure Title
- Macromolecule type
- Structure molecular weight
- Space group

3.2.1 Resolution

The obtained resolution of a particular protein molecule is dependent on the level of advancement of the experimental technique used in discovering the molecule, purity of samples, quality of crystals, strength of X-Rays, quality of detector, and method to process data. Intuitively, one would suspect that the resolutions obtained in earlier years are generally lower to those of more recent years. Although Dr. Hunsicker-Wang suspected this claim, it was not so clear how diverse the range of resolutions would be for each of the data sets from 1990 to 2010. In this experiment, we plan to see how effective the Diversity Map is in answering the question regarding the diversity of the resolution in each of the year's groupings.

3.2.2 Deposition Date

The date in which a molecule is submitted to the online Protein database is always recorded. This would be an integral part of my experiment given that I will be separating the data sets using this attribute. Dr. Hunsicker's prediction is that the amount of proteins deposited will increase as the years pass. Furthermore, we decided it will be worthwhile to analyze other attributes and how they evolve over the years.

3.2.3 Macromolecule Type

Given that we are analyzing the Protein Data Bank, it is not a surprise that the majority of the molecular type will be Proteins. However, there are other macromolecules present in the Bank, such as DNA, RNA, and the different combinations of the three. The distribution of these, we thought, will be interesting to analyze.

3.2.4 Structure Molecular Weight

This is the weight of the entire protein structure (or DNA, RNA) deposited in the Protein Data Bank. Just like the others, the structure molecular weight is self reported and subject to human induced errors. Dr. Hunsicker-Wang proposed that proteins deposited in the earlier years may have been smaller than those deposited in the later years due to the fact that they probably did not have the technological advancements to study larger molecules until later dates.

3.2.5 Experimental Technique

Dr. Hunsicker-Wang states that there are two main techniques for determining the structures of proteins (DNA, or RNA), X-Ray diffraction and Nuclear Magnetic Resonance (NMR). However, she claimed that it will be interesting to determine the degree to which the ratio between the two differs over the years.

3.2.6 Space Group

This attribute explains the symmetry of the repeating molecules in the Protein crystal. According to Dr. Hunsicker-Wang, one can think of this attribute as the spatial position of all the atoms in the protein molecule relative to the other copies of the proteins. A common notation in the protein field has been used to describe the different spatial arrangements (space groups). Dr. Hunsicker-Wang, through her previous research, hypothesizes that the $P_{2_1\,2_1\,2_1}$ space group would be prevalent.

3.3 PERL and MySQL Utilization

The PDB is efficient in producing desired data in an excel spreadsheet format. A sample of the excel spreadsheet one would see is shown in Figure 2

PDB ID	Space Group	Exp. Method	Dep. Date	Resolution	Structure MW	Macromol. Type
101M	P 6	X-RAY DIFFRACTION	1997-12-13	2.07	18112.99	* Protein
102M	P 6	X-RAY DIFFRACTION	1997-12-15	1.84	18010.86	* Protein
103M	P 6	X-RAY DIFFRACTION	1997-12-16	2.07	18093.99	* Protein
104M	P 1 21 1	X-RAY DIFFRACTION	1997-12-18	1.71	18030.79	* Protein
105M	P 1 21 1	X-RAY DIFFRACTION	1997-12-18	2.02	18030.79	* Protein
106M	P 6	X-RAY DIFFRACTION	1997-12-21	1.99	18182.04	* Protein
107M	P 6	X-RAY DIFFRACTION	1997-12-22	2.09	18209.09	* Protein
108M	P 6	X-RAY DIFFRACTION	1997-12-23	2.67	18209.09	* Protein
109M	P 6	X-RAY DIFFRACTION	1997-12-22	1.83	18133.94	* Protein
10GS	C 1 2 1	X-RAY DIFFRACTION	1997-08-14	2.20	47830.95	* Protein
10MH	H 3 2	X-RAY DIFFRACTION	1998-08-10	2.55	44768.76	* Protein * DNA
110M	P 6	X-RAY DIFFRACTION	1997-12-23	1.77	18118.91	* Protein
111M	P 6	X-RAY DIFFRACTION	1997-12-24	1.88	18160.99	* Protein
112M	P 6	X-RAY DIFFRACTION	1997-12-24	2.34	18146.96	* Protein
117E	P 21 21 21	X-RAY DIFFRACTION	1998-09-15	2.15	65203.82	* Protein
11AS	P 1 21 1	X-RAY DIFFRACTION	1997-12-02	2.50	73531.64	* Protein
11GS	C 1 2 1	X-RAY DIFFRACTION	1997-11-03	2.30	48369.21	* Protein
12AS	P 1 21 1	X-RAY DIFFRACTION	1997-12-02	2.20	74226.08	* Protein
12E8	P 1 21 1	X-RAY DIFFRACTION	1998-03-14	1.90	94839.80	* Protein
12GS	C 1 2 1	X-RAY DIFFRACTION	1997-11-19	2.10	48013.39	* Protein
13GS	C 1 2 1	X-RAY DIFFRACTION	1997-11-20	1.90	48752.93	* Protein
13PK	P 21 21 21	X-RAY DIFFRACTION	1996-11-23	2.50	181549.56	* Protein
14GS	C 1 2 1	X-RAY DIFFRACTION	1997-11-29	2.80	47146.27	* Protein
15C8	C 1 2 1	X-RAY DIFFRACTION	1998-03-18	2.50	46360.60	* Protein
16GS	C 1 2 1	X-RAY DIFFRACTION	1997-11-30	1.90	47242.32	* Protein
16PK	P 21 21 21	X-RAY DIFFRACTION	1998-05-18	1.60	45571.98	* Protein
17GS	C 1 2 1	X-RAY DIFFRACTION	1997-12-07	1.90	48010.28	* Protein
17RA	P 1	SOLUTION NMR	1998-08-04		6729.09	* RNA
18GS	C 1 2 1	X-RAY DIFFRACTION	1997-12-07	1.90	48093.09	* Protein
1914	P 43 2 2	X-RAY DIFFRACTION	1997-11-13	2.53	26562.90	* Protein
19GS	C 1 2 1	X-RAY DIFFRACTION	1997-12-14	1.90	49084.55	* Protein
19HC	P 1 21 1	X-RAY DIFFRACTION	1998-11-27	1.80	73973.79	* Protein
1A00	P 21 21 21	X-RAY DIFFRACTION	1997-12-08	2.00	64565.79	* Protein
1A01	P 1 21 1	X-RAY DIFFRACTION	1997-12-08	1.80	64381.59	* Protein

Figure 2 - Portion of PDB Data Set downloaded from

The PDB allows one to select the desired attributes for each protein and displays the results in a spreadsheet format prior to downloading the data. Data extraction becomes more efficient because the results obtained from the download contain only pertinent information for analysis to the user. In this case, only the six attributes in the columns in Figure 2 were selected. Once this spreadsheet is downloaded, a series of steps is taken to prepare the data.

3.3.1 Data Preparation

Preparation of data for the Diversity Map program is a very important part of this research. Like any other software application, you can only program an application to do exactly what you want it to do. This degree of flexibility hence depends on the programmer. In this case, Metoyer's research group purposely set guidelines for the input data that are required of the data before ensuring a successful visualization. Explained below, are the series of preparations that I did to correctly apply the Diversity Map to the data sets gotten from the Protein Data Bank.

- **The input data must be in the tab delimited text (.txt) format**

This is relatively easy to accomplish. Data could be received from the Protein Data Bank either as an excel file or a comma-separated file (csv) format. This could then be saved as a tab delimited text using a simple "save as" option on Microsoft Excel 2007.

- **First round of data formatting (Perl)**

I realized that once the Excel formatted sheet to a text format was saved, blocks with multiple entries split the line into two. The most prominent of this was in the *Macromolecular Type* column. Although most values are Protein, DNA, or RNA, there are some values which are actually a hybrid of Protein and either DNA or RNA denoted as either * *Protein* * *DNA* or * *Protein* * *RNA*. To fix this, a Perl script was written, as seen in Figure 3.

```
while($next=<FILE>){

        if ($next =~ /^ \*/){
        chomp $text;
                $text = $text.$next;
                $next = <FILE>;
        }

#       modify(\$text);

        print OUTPUT $text;
        $text = $next;
}
```

Figure 3 - Perl Script row concatenation

Figure 3 shows that the Perl script loops through the entire data set taking each row as a string of text. The first step is more of an error detector which checks for blocks with multiple entries. After the first download from the Protein Data Bank, multiple entries in a block are divided by '*' in the text file. However, I noticed through trial and error that after further manipulation of

the data through MySQL, the text file reads the second part of the data and moves whatever is remaining in that row into a new row. To rectify this problem, an initial file was run through the Perl script which essentially detects the occurrence of '*' and concatenates the string before and after the character.

```
$$text =~ s/SOLID-STATE NMR/NMR/gi;
$$text =~ s/FIBER DIFFRACTION/DIFF./gi;
$$text =~ s/ELECTRON MICROSCOPY/ELEC MICR/gi;
$$text =~ s/ELECTRON CRYSTALLOGRAPHY/ELEC CRYS/gi;
$$text =~ s/SOLUTION NMR/NMR/gi;
$$text =~ s/SOLUTION NMR\, XRAY DIFF/MIX/gi;
```

Figure 4a - Perl code for shortening value names

After checking for unwanted split lines, the data looped through a code, like Figure 4a**,** which shows a list of code lines that serve the purpose of shortening phrases. The first line is instructing the script to locate any occurrence of "Solid-State NMR" and converting that to "NMR", ignoring the letter-case.

```
$$text =~ s/2008\-[0-9]+\-[0-9]+/2008/gi;
$$text =~ s/2009\-[0-9]+\-[0-9]+/2009/gi;
$$text =~ s/2010\-[0-9]+\-[0-9]+/2010/gi;
$$text =~ s/ \*[ ]DNA/DNA/gi;
$$text =~ s/ \*[ ]RNA/RNA/gi;
$$text =~ s/ \*[ ]Protein/Protein/gi;
$$text =~ s/\"ProteinDNA\"/Protein\/DNA/gi;
$$text =~ s/\"ProteinRNA\"/Protein\/RNA/gi;
$$text =~ s/DNARNA/DNA\/RNA/gi;
$$text =~ s/DNA\/RNA Hybrid/DNA\/RNA/gi;
$$text =~ s/ProteinDNA\/RNA Hybrid/Protein\/DNA\/RNA/gi;
$$text =~ s/X-RAY DIFFRACTION/XRAY DIFF/gi;
```

Figure 4b - Perl Code for editing value names

Then, the data was run through a series of commands to modify various strings in the database. Figure 4b shows the list of various commands that each script must run through. Some which include simplifying the *Deposition Date* values from the 'year-month-day' format to just 'year' (shown in the first three lines of Figure 4b) and standardizing all entries of a Protein-DNA or Protein-RNA hybrid to 'Protein/DNA' and 'Protein/RNA' respectively. This is a core part of the research simply because without this rigorous form of data preparation, the Diversity Map would be useless in visualizing this data set. There would be unnecessary clutter in the visualization and would hinder any form of analysis on the data.

- **Second round of data formatting (MySQL)**

After going through the series of string manipulation using the Perl script, the data is uploaded into Navicat Lite, an open source MySQL database manipulation software. MySQL is effective in my research for two reasons. First, it is easier to manipulate data in a column in MySQL than in Perl, while Perl seems more powerful manipulating data in a row. Secondly, it is easier to manipulate strings as numerical values in the MySQL Database. These two features will be useful in formatting my data.

The first thing that is done is to fill up all the empty blocks, blocks in which the user did not enter a value for, with the string, "null" as shown in Figure 5.

```
UPDATE `data set format 1990-1992`
SET
`Space Group` = (CASE WHEN (`Space Group` IS NULL or `Space Group` = "") THEN "null" else `Space Group` END),
`Exp. Method` = (CASE WHEN (`Exp. Method` IS NULL or `Exp. Method` = "") THEN "null" else `Exp. Method` END),
`Dep. Date` = (CASE WHEN (`Dep. Date` IS NULL or `Dep. Date` = "") THEN "null" else `Dep. Date` END),
`Structure MW` = (CASE WHEN (`Structure MW` IS NULL or `Structure MW` = "") THEN "null" else `Structure MW` END),
`Macromol. Type` = (CASE WHEN (`Macromol. Type` IS NULL or `Macromol. Type` = "") THEN "null" else `Macromol. Type` END),
`Resolution` = (CASE WHEN (`Resolution` IS NULL or `Resolution` = "") THEN 20 else `Resolution` END);
```

Figure 5 - MySQL query for filling up all blank blocks with the string “null”

Figure 5 shows an example of a MySQL query. In this query, the data set is *updated* by *setting* the values of each of the attributes to the string “null” *when* the values of each of the attributes are empty. This query to fill visualization with "null" would play an important role in enhancing understanding of the final visualization. The only exception to this rule was for the *Resolution* column. Instead of a “null” string, the empty blocks were replaced with ‘20’. This was an arbitrary number picked to signify that the user did not enter a value for the resolution used. I picked the number ‘20’ instead of the string ‘null’ because I would later have to treat the entire column as one with only numerical values.

The next set of updates was focused on the resolution and molecular weights column. The Diversity Map works best when numerical values are grouped into ranges, with each range mapped to a bucket in the visualization (the definition of discretization). The visualization will be meaningless if these groupings are not made because one would not be able to make any claim regarding the diversity of the attribute. However, the range at which one sets will determine the level of specificity that one is willing to have. A bigger range will mean lower specificity and vice versa since we essentially hide more information about the data set by making these groupings. Therefore, a further consultation with Dr. Hunsicker-Wang, an expert in

the field of proteins, was arranged in order to determine how wide the range should be for the resolution and molecular weight. She advised that the resolution should have a range of 0.2 (Example, 2.0-2.2) and the molecular weight should have a range of 5000 (Example: 50,000-55,000). Figure 6a and 6b, show the code used to update the entire file column.

```
UPDATE `data set format 1990-1992`
SET Resolution =
(CASE
when Resolution = 'null' then Resolution = "null"
WHEN Resolution >= 0.00 AND Resolution < 0.20 THEN '0.0-0.2'
when Resolution >= 0.20 AND Resolution < 0.40 then '0.2-0.4'
when Resolution >= 0.40 AND Resolution < 0.60 then '0.4-0.6'
when Resolution >= 0.60 and Resolution < 0.80 then '0.6-0.8'
when Resolution >= 0.80 AND Resolution < 1.00 then '0.8-1.0'
when Resolution >= 1.00 AND Resolution < 1.20 then '1.0-1.2'
when Resolution >= 1.20 and Resolution < 1.40 then '1.2-1.4'
when Resolution >= 1.40 AND Resolution < 1.60 then '1.4-1.6'
when Resolution >= 1.60 AND Resolution < 1.80 then '1.6-1.8'
when Resolution >= 1.80 and Resolution < 2.00 then '1.8-2.0'
when Resolution >= 2.00 AND Resolution < 2.20 then '2.0-2.2'
when Resolution >= 2.20 AND Resolution < 2.40 then '2.2-2.4'
when Resolution >= 2.40 and Resolution < 2.60 then '2.4-2.6'
when Resolution >= 2.60 AND Resolution < 2.80 then '2.6-2.8'
when Resolution >= 2.80 and Resolution < 3.00 then '2.8-3.0'
when Resolution >= 3.00 AND Resolution < 3.20 then '3.0-3.2'
when Resolution >= 3.20 AND Resolution < 3.40 then '3.2-3.4'
when Resolution >= 3.40 and Resolution < 3.60 then '3.4-3.6'
when Resolution >= 3.60 AND Resolution < 3.80 then '3.6-3.8'
when Resolution >= 3.80 and Resolution < 4.00 then '3.8-4.0'
when Resolution >= 4.00 and Resolution < 20.0 then '4.0-20.0'
when Resolution >= 20.0 then '20.0'
END)
```

Figure 6a - MySQL code to update resolution

```
(CASE
WHEN `Structure MW` >= 0 AND `Structure MW` < 5000 THEN '0-5000'
WHEN `Structure MW` >= 5000 and `Structure MW` < 10000 THEN '5000-10000'
when `Structure MW` >= 10000 and `Structure MW` < 15000 THEN '10000-15000'
when `Structure MW` >= 15000 and `Structure MW` < 20000 THEN '15000-20000'
when `Structure MW` >= 20000 and `Structure MW` < 25000 then '20000-25000'
when `Structure MW` >= 25000 and `Structure MW` < 30000 then '25000-30000'
WHEN `Structure MW` >= 30000 and `Structure MW` < 35000 then '30000-35000'
WHEN `Structure MW` >= 35000 and `Structure MW` < 40000 then '35000-40000'
when `Structure MW` >= 40000 and `Structure MW` < 45000 then '40000-45000'
when `Structure MW` >= 45000 and `Structure MW` < 50000 then '45000-50000'
when `Structure MW` >= 50000 and `Structure MW` < 55000 then '50000-55000'
when `Structure MW` >= 55000 and `Structure MW` < 60000 then '55000-60000'
when `Structure MW` >= 60000 and `Structure MW` < 65000 then '60000-65000'
when `Structure MW` >= 65000 and `Structure MW` < 70000 then '65000-70000'
when `Structure MW` >= 70000 and `Structure MW` < 75000 then '70000-75000'
when `Structure MW` >= 75000 and `Structure MW` < 80000 then '75000-80000'
when `Structure MW` >= 80000 and `Structure MW` < 85000 then '80000-85000'
when `Structure MW` >= 85000 and `Structure MW` < 90000 then '85000-90000'
when `Structure MW` >= 90000 and `Structure MW` < 95000 then '90000-95000'
when `Structure MW` >= 95000 and `Structure MW` < 100000 then '95000-100000'
when `Structure MW` >= 100000 and `Structure MW` < 105000 then '100000-105000'
when `Structure MW` >= 105000 and `Structure MW` < 110000 then '105000-110000'
when `Structure MW` >= 110000 and `Structure MW` < 115000 then '110000-115000'
when `Structure MW` >= 115000 and `Structure MW` < 120000 then '115000-120000'
when `Structure MW` >= 120000 and `Structure MW` < 125000 then '120000-125000'
when `Structure MW` >= 125000 and `Structure MW` < 130000 then '125000-130000'
when `Structure MW` >= 130000 and `Structure MW` < 135000 then '130000-135000'
when `Structure MW` >= 135000 and `Structure MW` < 140000 then '135000-140000'
when `Structure MW` >= 140000 and `Structure MW` < 145000 then '140000-145000'
when `Structure MW` >= 145000 and `Structure MW` < 150000 then '145000-150000'
when `Structure MW` >= 150000 and `Structure MW` < 155000 then '150000-155000'
when `Structure MW` >= 155000 and `Structure MW` < 160000 then '155000-160000'
when `Structure MW` >= 160000 and `Structure MW` < 165000 then '160000-165000'
when `Structure MW` >= 165000 and `Structure MW` < 170000 then '165000-170000'
when `Structure MW` >= 170000 and `Structure MW` < 175000 then '170000-175000'
when `Structure MW` >= 175000 and `Structure MW` < 180000 then '175000-180000'
when `Structure MW` >= 180000 and `Structure MW` < 185000 then '180000-185000'
when `Structure MW` >= 185000 and `Structure MW` < 190000 then '185000-190000'
when `Structure MW` >= 190000 and `Structure MW` < 195000 then '190000-195000'
when `Structure MW` >= 195000 and `Structure MW` < 200000 then '195000-200000'
when `Structure MW` >= 200000 then 'Over 200000'
END)
```

Figure 6b - MySQL code to update Structure MW

Another point to emphasize here is my reasoning for inserting a '20', instead of a null, for resolution blocks with no values. As seen in Figure 6a**,** I treat the entries in resolution as numerical values and manipulate them as such (i.e. using the '>=' and '<' to compare numerical values) and then the resulting value is a string.

3.3.2 Data Integration

META File Generation

In order for the data set to be properly entered into the Java program, a META file needs to be created. This is an index of all the possible values (buckets) for each of the attributes that the

data set contains. To obtain this list, I used a MySQL query (Figure 7) to get the list of possible values for each of the attributes, order it alpha-numerically, and then paste it into a text file.

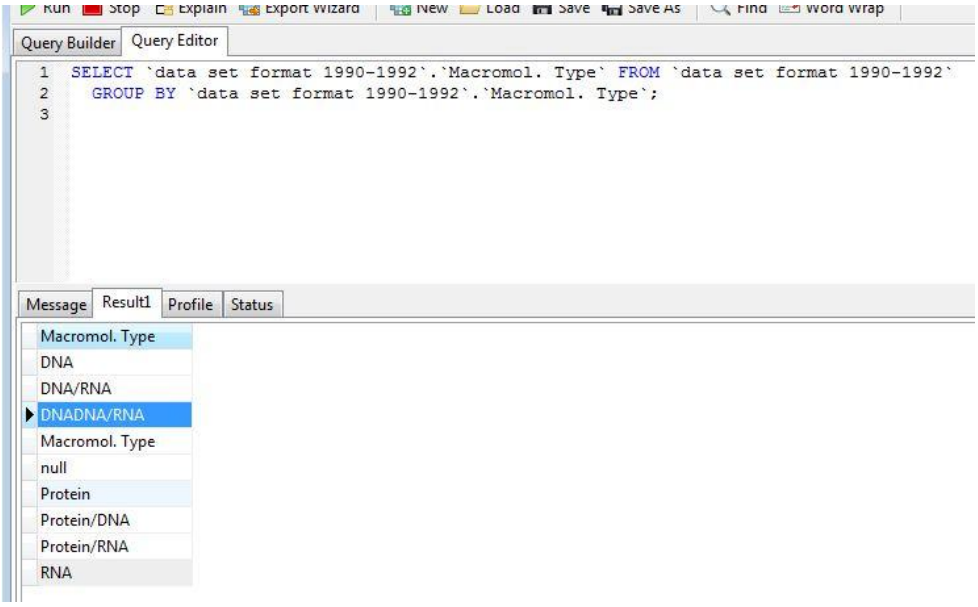

Figure 7 - generating the possible values for Macromol. Type in the 1990-1992 data set

This ordered list needs to be appended into the META file in a strict format of "*Attribute Name Attribute Value"*. This can be easily done using a Perl script (shown in Figure 8**)** that inserts the Attribute name in front of the value.

```
#usr/local/bin/perl

$inputfile = "meta convert file.txt";
$outputfile = "data_org_meta.txt";

print "$inputfile\n";
print $outputfile;
open(FILE, $inputfile);
open(OUTPUTFILE, ">> $outputfile");

my $line;
while($text=<FILE>){
        chomp $text;
        print OUTPUTFILE "Macromol. Type        $text\n";
};

close(FILE);
~
~
~
~
```

Figure 8 - appends attribute values to META file in the correct form

The program opens the input file containing the alpha numerically ordered attribute values and the output META file. The *while* loop goes through every string in the input and inserts the name of the attribute, in this case, Macromol. Type, in front of the string and then appends that string to the output file (Figure 9**).**

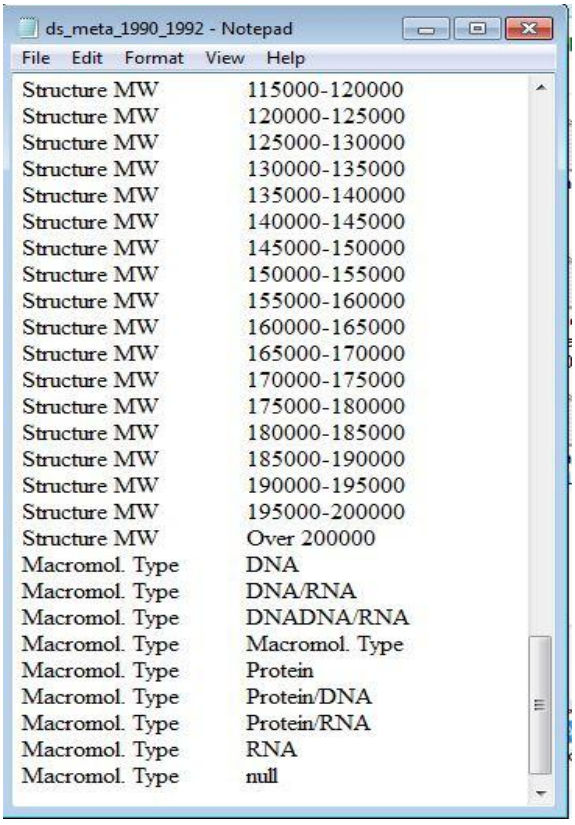

Figure 9 - part of a META file

Java Code Integration

Once the Meta file and data set file are both prepared, data integration to the system becomes very straightforward. Since the program had already been made by the Dr. Metoyer's research team in Oregon State University (Pham et. al 2010)**,** the only changes in the code that need to be made is the string name for the data set and meta file, as well as the string array declaring the attributes for the data set. Once this is done, the program is able to generate the visualization map. Figure 10 shows the data set Diversity Map for the proteins deposited in the Protein Data Bank between January 1, 1990 and December 31, 1992.

Figure 10 - Diversity Map visualization for Data Set 1990-1992

As seen in Figure 10, each column represents a unique attribute and each rectangular block in that column represents a value of the attribute. Also, a darker shade of blue in a particular bucket indicates a larger amount of objects in that category. Consequently, the diversity of a column is defined as the evenness in distribution of color intensity across that column. The *Resolution* and *Structure MW* attributes have been discretized (making a continuous value set discrete by placing them into a broader bucket). The range of 0.2 and 5000 for the *Resolution* and *Structure MW* respectively was determined by Dr. Huncicker-Wang as one that would be broad enough to synchronize properly with this Map but still retain a certain level of specificity to be informative to the user.

Using the mouse to hover on a particular bucket displays the exact amount of objects (individuals) that have been placed in that bucket hence, increasing the amount of detail of the

Diversity Map. For example, in Figure 10, the mouse was placed on the *1992* bucket on the *Dep. Date* (Deposition Date) column and on the top right corner, it states that "There are 524 individuals in the '1992' bucket". Also, by clicking and holding on to any of the buckets, the visualization filters its distribution to only the objects in the selected bucket. Thereby, adding an extra level of detail to the program. In addition to this, I have placed emphasis on the deposition years and added a radio button filter on the lower right corner of Figure 10. By clicking on any of the radio buttons, the Diversity Map filters the visualization to only show the distribution for only the objects that fall into the specific year specified. The advantage of this feature is that the user does not have to hold the click of the mouse on the desired bucket.

One can also opt to temporarily remove the labels of each bucket by clicking on the check box for *Draw Labels.* Figure 11 shows the same visualization without the labels.

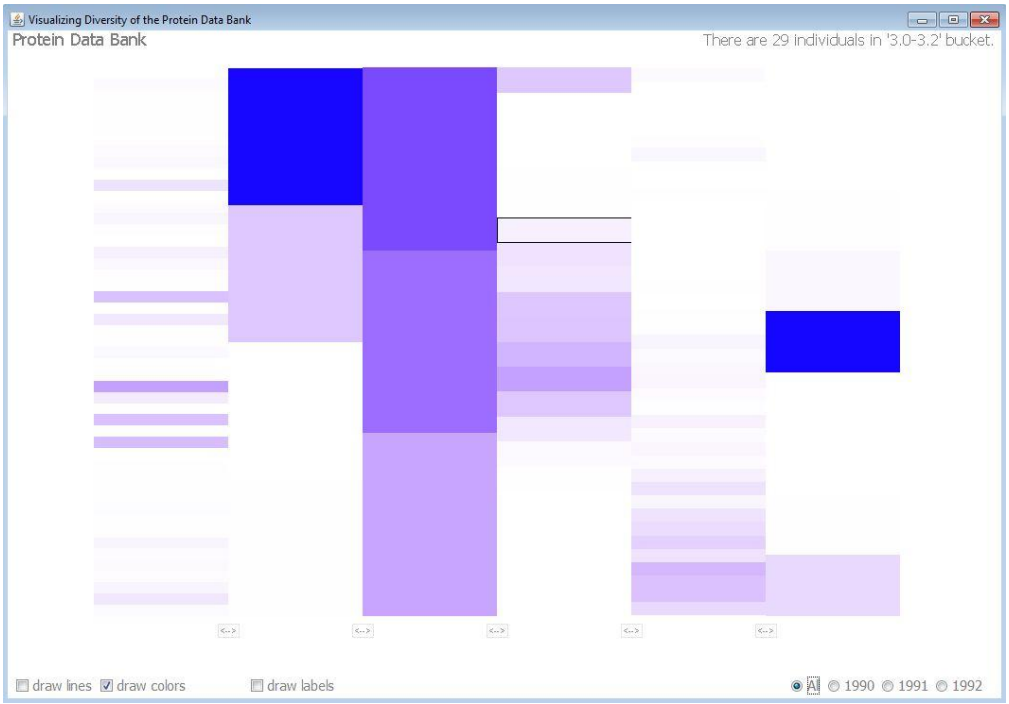

Figure 11 - Diversity Map for Data Set 1990-1992 without label

By doing this, one can focus simply on the distribution (diversity) of the each of the column, as well as the entire data set without being distracted with the details of each of the attribute names. Unlike the others, this feature adds another level of abstraction to the Diversity Map.

Another feature of the Diversity Map is the *draw lines* checkbox. Clicking on this check box and "un-clicking" the *draw colors* check box shows the correlation between attributes (columns). Figure 12 shows the same data set from 1990-1992 with the *draw lines* as well as the *1991 radio button filter* feature applied.

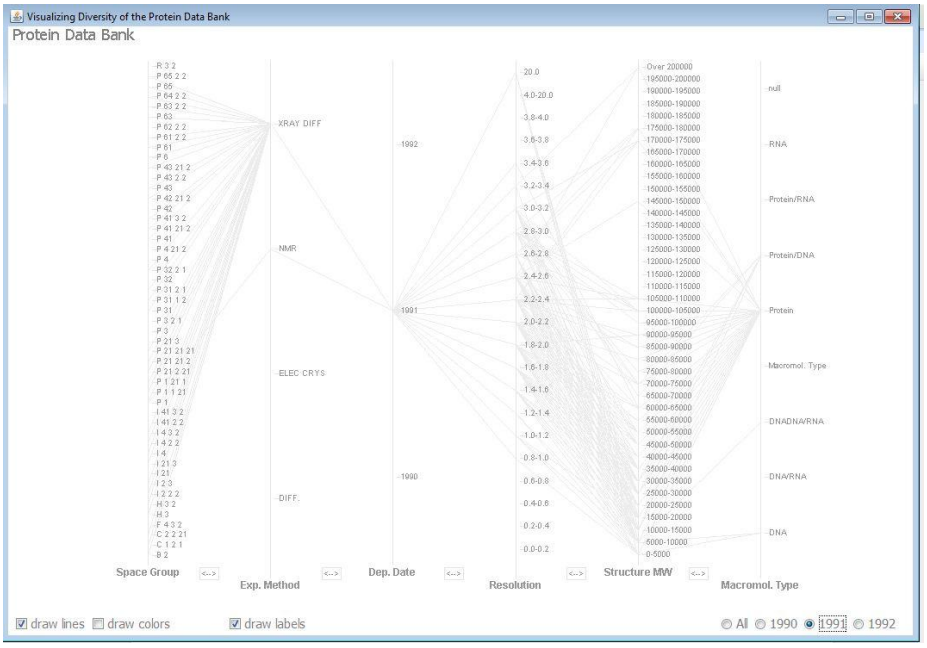

Figure 12 - Data Set 1990-1992 with "draw lines" box clicked and 1991 filter

When there is a correlation between two buckets of different attributes, a connecting line is drawn. A line drawn from one bucket to another signifies that there is at least one object with the two attribute values. For example, a conclusion could be made that all DNA deposited into the Protein Data Bank in 1991 were of molecular weight between the range of 0-10,000. However, since this connection line is only drawn for adjacent attributes (columns), it would be difficult to draw conclusions on correlation between non-adjacent attributes on the Map. Consequently, another level of detail was implemented into the Diversity Map in order to resolve the problem of finding correlation beyond adjacent attributes. As seen in Figure 13 a red line is highlighted when the mouse is place above the desired line to observe.

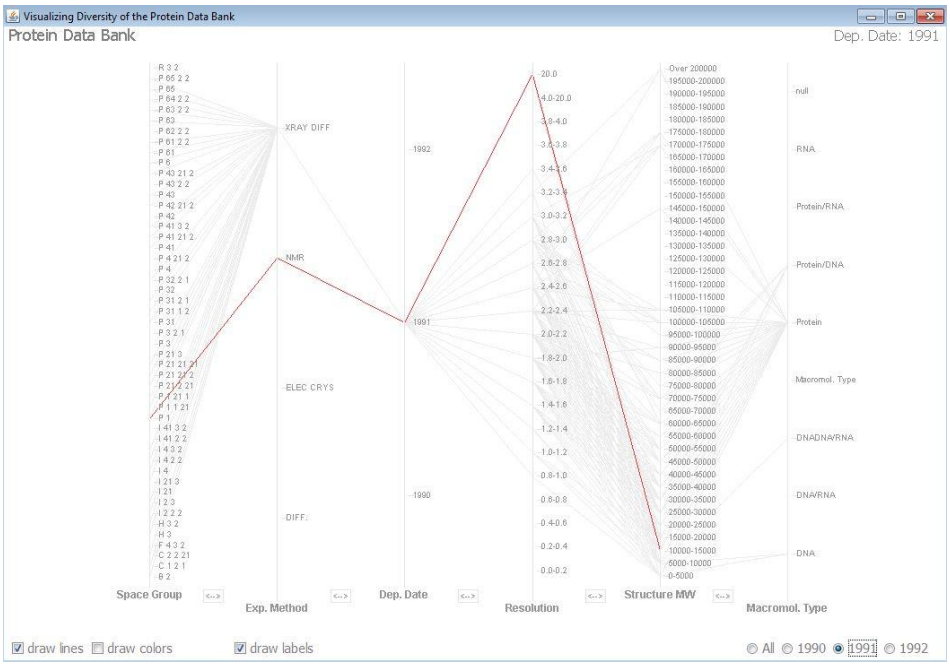

Figure 13 - Data Set 1990-1992 highlighted red line correlation among attributes

Utilizing the extra level of detail that his feature brings, one can deduce, in this example, that at least one NMR-derived object in 1991 have the Space Group of P 1, resolution of 20.0 (which is

the arbitrary value for non-specified resolution values), and molecular weight between 10000 and 15000.

3.4 Case Study

The main goal of this thesis is to determine if this Diversity Map is applicable to the Protein Data Bank. In order to do this, Dr. Hunsicker-Wang, a chemistry professor at Trinity University, who is an expert in proteins, volunteered to be part of my informal case study. The focus of this study was on the analysis of the data set visualization map from 1990 to 2010.

3.4.1 First Phase

The first part of the informal study was to determine which attributes are the most efficient to visualize while still being informative to the user of the Map. Multiple sessions were organized in order to accomplish this. After doing this, we examined the list of possible values for each attribute to determine which one needed edits. Ranging from shortening *X-Ray Diffraction* to *X-Ray Diff.* to determining that 0.2 is an appropriate range for the *Resolution* bucket, a rigorous process was undertaken to clean up the data. The result of this part of the case study is the result of the data preparation stage.

3.4.2 Second Phase

Second phase of the informal study was primarily to test the visualization obtained from the result of the first phase of the study. Dr. Hunsicker-Wang was given two out of the seven visualizations of the Protein Data Bank, displayed individually, and was asked to analyze the Map using all of its features. She was asked to state the strengths and weaknesses of the visualization in relation to the excel spreadsheet of the data sets that she had previously analyzed.

In addition, she looked at two other sets of visualizations (comparison experiment) one was far apart in years (1990-1992 and 2008-2010) and one was close to each other (1999-2001 and 2002-2004). She commented on the correlations that she noticed between the visualizations. I used her analysis of the test to formulate questions for the final phase of my informal study: the student participant study section.

3.4.3 Third Phase

The final phase of the study involved preparing and executing formal case studies with student participants with basic experience in chemistry. It was a more formalized study than that of the second phase with a controlled set of questions stemmed from Dr. Hunsicker-Wang's analysis of the Map. The ten student participants were asked to fill an introductory questionnaire stating their class year, their level of experience with chemistry, filling in chemistry classes already taken, and state if they usually wear contact lenses or prescription glasses.

Before the study began, students were given an introduction to the entire thesis describing origin and development of the visualization idea since 2009. They were told about the previous study with Dr. Hunsicker-Wang and the purpose of them being present and participating in the study. They were given a tutorial of the Diversity Map using Data Set from 1996-1998 of the Protein Data Bank as tutorial Map (Figure 14).

Figure 14 - Data Set 1996-1998

The participants were aware of all the features of the Map and advised to utilize as many of the features to answer the questions that they were about to be asked. In addition, the concept of color intensity and diversity as it relates to the Diversity Map was formally explained to the participants so as to ensure they know what kind of concepts to think of when the questions are asked. Also, none of the questions required the students to have a deep understanding of chemistry in order to answer them. Rather, the questions were phrased to focus on the student's understanding of solely the visualization technique. Before the study began, students were asked if they had any questions about all that had being said and then asked to respond vocally to every question asked of them so as to ensure their understanding of the question. They were given visualizations that looked like Figure 14 and questions were broken down into three forms in order to determine certain aspects of the visualization.

Basic Questions

These questions were posed to determine the participant's general understanding of the visualization itself. Focus was to determine the student's ability to locate attributes on the Map as well as understand the idea of color intensity.

- Could you point out the *Experimental Method* column?
- Could you point out the *Macromol. Type* column?
- What is the most prominent bucket in the *Experimental Method* column?
- What is the most prominent bucket in the *Macromol. Type* column?

Applied Questions

These questions were posed to determine if the participants understand the features of the Diversity Map. The focus was on allowing the student participant to explore the features, use them to answer the questions, as well as possibly make inferences, correlations, and conclusions.

- What is the most common resolution of this data set?
- What is the most prominent bucket in the Space Group column
- Click and hold on to "1.8-2.0" resolution range. Do you notice differences between the filtered Map and the general Map? If so, explain them?
- Click and hold on to "20.0" resolution range, Do you notice differences between the filtered Map and the general Map? If so, explain them?
- Could you explain the distribution of the *Resolution* column?
- Could you explain the distribution of the *Structure MW* column?
- In your opinion, is the Space Group column diverse or not?

Comparison Questions

These questions were posed as the student participants were allowed to look between two specified data set visualizations. The goal of this was to observe how the student participants made a judgment call on which of the Maps were more diverse.

- What are the similarities/differences between the *Dep. Date* columns in both visualizations?
- What are the similarities/differences between the *Structure MW* columns in both visualizations?
- What are the similarities/differences between the *Resolution* columns in both visualizations?
- Which Map is more diverse?

All correspondence was recorded with the consent of the participant and was promised to only be used to report responses and observations in this thesis paper.

4 Results

Three levels of analysis on the Protein Data Bank were derived from the informal case study. The first results were obtained from Dr. Hunsicker-Wang who gave insights into the strengths and weaknesses of the application of this Diversity Map to the Protein Data Bank, and elaborated on the interesting deductions from each of the visualizations as well as the efficiency of the Diversity Map at its fundamental level of abstraction. The second set of results was from the student participant case studies. The third finding was a result of Hunsicker-Wang's analysis of the same data set with the additional line feature which should give the Map a greater level of detail.

4.1 Expert Analysis

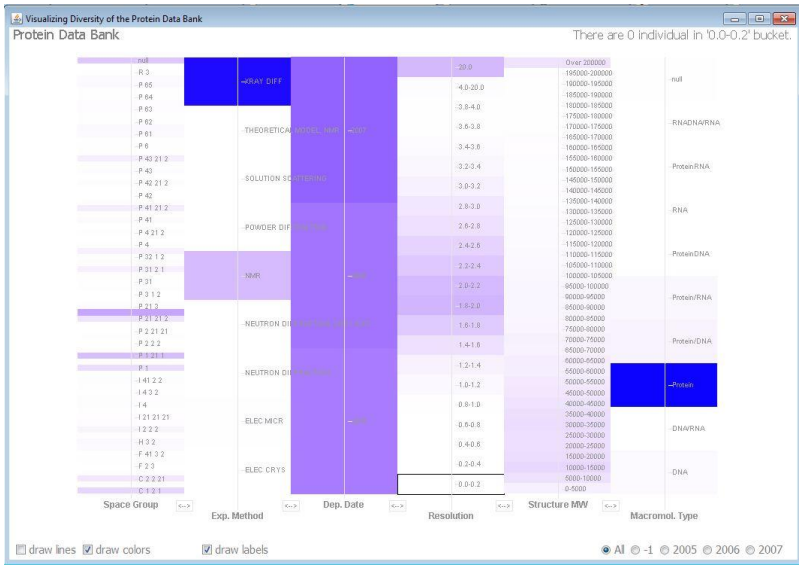

Figure 15 - Data Set 2005-2007

Dr. Hunsicker-Wang was asked to analyze the Data Set 2005-2007 (Figure 15) and four other data sets in my research**,** first explaining what was noticed by observing the entire visualization and then looking at trends through each of the columns. Her first impressions of the data are the objects in the darkest buckets. In the experimental methods column, it was noticed that X-Ray diffraction bucket carried the most amount of objects. This observation was reflected in the Macromolecular type column as well. It was extremely easy to notice the objects of highest color intensity. It was immediately noticed that the increase in intensity over the years in the deposition date column. It is interesting to note that the relatively slight change in color intensity across a column was very noticeable to Dr. Hunsicker-Wang at first glance. The next phase was to analyze each column.

Space Group

Dr. Hunsicker-Wang was interested in seeing if there was a certain space group that is prevalent in the Protein Data Bank deposits. Through prior research experience, it was predicted that the

space group $P_{21\,21\,21}$ should be one of the most prominent in the Data Bank. This was based on the nature of proteins that she had worked with in her laboratory. While analyzing the Map, it was immediately noticed that the most prominent bucket (most intense in color) in Data Set 2005-2007, as well as all the data sets visualizations she looked at, was indeed the *P 21 21 21* bucket.

Experimental Method

The distribution of this column also fell within Dr. Hunsicker-Wang's initial prediction that *X-Ray Diff.* would be the bucket with the most amount of values. Again, this prediction stemmed from her research in the field. This column, as correctly stated, expresses very low diversity.

Deposition Date

Dr. Hunsicker-Wang noted that the increase in color intensity in ascending order aligned with the perceived trend of the Protein Data Bank. With an increased awareness of the Bank, and an increased access to experimental methods, a large amount of data is deposited yearly into the database. However, there was an anomaly that was easily noticed using the Diversity Map. When Dr. Hunsicker-Wang was asked to compare the data sets of 1990-1992 and 2008-2010, it was noticed that 2010 had a lower amount of deposited protein objects than 2009.

Resolution

Dr. Hunsicker-Wang was interested in determining the most common resolution for a protein structure. She determined using the Diversity Map that 1.8-2.2 is the most common range across the years. However, in the earlier years, she was able to notice that the resolution leaned towards 2.0-2.2, while towards 1.8-2.0 in the later years, which was a reasonable trend because of technological advances to obtain proteins at a higher resolution in the later years. In addition, when comparing resolution trends

between earlier and later years, she noticed that the distribution of the resolution objects were wider at later years, hence, more diverse than that of the earlier years.

Structure Molecular Weight

Dr. Hunsicker-Wang mentioned that the higher amount of color intensity to the lower ranges in the molecular weight columns is evidence of the fact that smaller molecules are easier to crystallize and hence, are more prominent on the data bank. When comparing earlier and later data sets, she noticed that, similar to the resolution, the *Structure MW* columns of later years are higher in diversity than those of earlier years.

Overall Trends

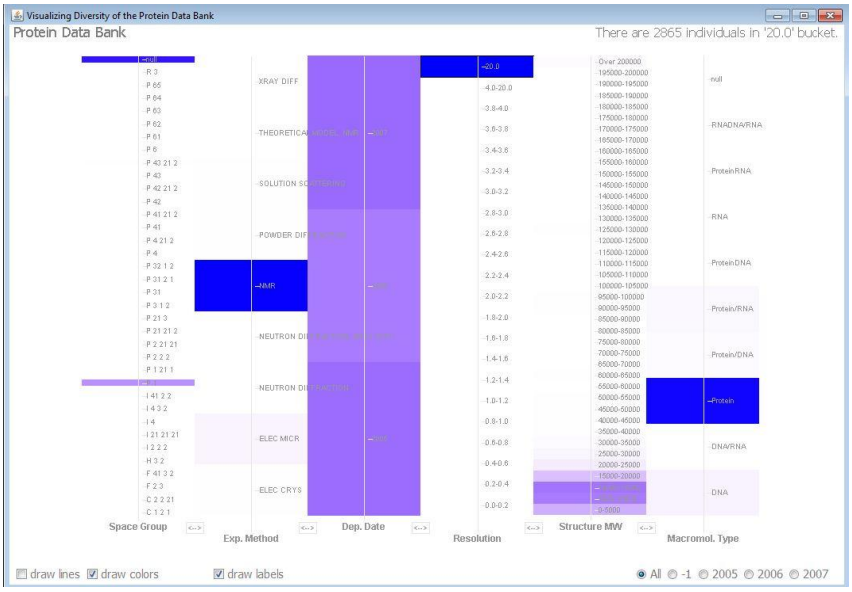

Figure 16 - Data Set 2005-2007 with 20.0 highlighted

In addition to analyzing individual columns, Dr. Hunsicker-Wang was asked to find correlations between columns on the Map. She used the radio buttons to filter the objects for 2005, 2006, 2007 and she did see the correlation between the *Deposition Date* and the *Resolution.* She noticed that the more recent the date of deposition, the higher the resolution. Another major correlation was between the arbitrary '20' value

for the resolution and the rest of the data sets. Unspecified resolution values were from objects obtained from NMR and have the lower structural molecular weight (Figure 16). Overall, she concluded that there was a high amount of diversity in the space group column, a low amount of diversity in the experimental method, more diversity in resolution and molecular weight but with clear trends, and finally a predominance of proteins in the macromolecular type column.

4.2 Participant Survey

Ten students volunteered to partake in this study. Of the ten students, three were three seniors, six juniors and one first-year student. All had taken the general chemistry 1&2 (combined at Trinity University) course. Nine of them had taken up to organic chemistry 1, six of them with organic chemistry II background, and three of them up to had taken the biochemistry 1 course. None had consequential visual impairments and if without 20/20 vision, stated that they were using contact lenses or prescription glasses at the time of the experiment.

All participants were able to answer the basic questions listed above with ease after they had understood the Map and gone through the 5 minutes tutorial. Each of them pointed out the appropriate columns that represented the attribute in question. Furthermore, all participants correctly determined the buckets with the highest amount of objects by comparing the intensity of the color of a bucket in a column. This showed that the basic concept of the Map could be easily understood and one could answer basic questions of the Map without prior knowledge of the data set.

Regarding the applied questions, surprisingly, all participants were able to determine the most common resolution range in the data set, which was 1.8-2.0. This could be classified as a difficult question, considering the closeness in color shade of the adjacent buckets in the resolution column. However,

participants were able to distinguish the shades without hovering on the bucket to view the exact amount of objects in it.

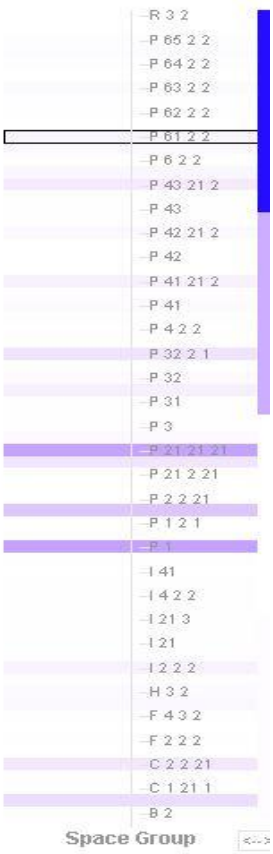

Figure 17

However, different answers were obtained when the participants were asked to determine if the Space Group attribute is diverse (Figure 17). This question was posed for two reasons; to verify that they completely understood the definition of diversity as it is related to the Map, and to analyze how the participants were thinking about the data in the visualization map. There were two main views to answer the questions. Some students claim that the *Space Group* is diverse because the color is distributed, albeit coarsely, across the column. The other group claims that *Space group* is not diverse because of the *holes* in the column (buckets with relatively little objects in them) that produces significant breaks in the data. One student noted that if the buckets were rearranged to have all the objects with color on one end and the

relatively white buckets on the other end, about half of the column would be white. The participant stated that that was how the conclusion that "either answer could be argued" was made.

When asked to explain the distribution of the Resolution and Structure Molecular Weight, all the students were able to describe the trend in color shades as they saw on the columns. For the *resolution,* all students were able to notice that the distribution peaked between 1.8 and 2.2 and declined with an almost equal gradient on both sides of the column. Three of the students who had additional experience in mathematics were able to describe that the resolution distribution in all visualizations fit a bell curve. For the *Structure Molecular Weight*, once again, all student participants were able to determine that most of the deposited objects were situated in the lower molecular weight region.

Participants were able to notice major differences between the visualization they observed holding on to the 20.0 resolution bucket (see figure 16) and that of the un-filtered map. All of the participants noted that the *NMR* bucket was exclusively shaded under the experimental method. They noticed the lack of diversity in the *Space Group* column and the shift of the structural molecular weight to the lowest buckets. However, most participants did not notice changes within the distributions when those changes were relatively minor. For example, Dr. Hunsicker-Wang noticed a slight shift of the resolution distribution to 1.8-2.0 range in the later years than in the earlier years. Only 1 out of 10 student participants was able to point this out.

The final question on comparing diversity between two data sets gave the most interesting observations. I asked each participant to compare Data Set 1990-1992 and Data Set 2008-2010 and determine which one was more diverse. Although their answer to this question was important, the method they used to arrive at a conclusion was equally as intriguing. Out of the 10 students that participated, 3 of them said that the Data Set from 1990-1992 is more diverse than that from 2008-2010. The reason behind this conclusion was in the concept of randomness. They reasoned that since some columns in the 1990-1992 data set have

less of a pattern than those in 2008-2010, the former had to be more diverse than the latter. The definition of diversity as it relates to this Map was explained in the beginning of each survey as the evenness, or a wide spread/distribution, of color intensity across a column. This definition never alluded to the "randomness" of the color distribution and so should not have been a factor in this reasoning. However, this randomness is a preconceived notion of diversity. Diversity equals variety, in which to some may infer randomness. However, a diverse set of objects do not necessarily have to be random, at least according to our definition in this research. Nevertheless, the remaining seven participants chose the 2008-2010 data set as the more diverse one because they claimed it was more evenly distributed, overall, across the column. One student mentioned the less amount of extreme color intensity (dark blue, and white) in the 2008-2010, leaving the visualization to be filled with "light blue" spread across the visualization. One other student reasoned that due to the higher amount of technology in the later years, more proteins have been discovered and entered into the database hence, giving the Map a wider variety than that of the earlier years.

4.3 Diversity Map with Lines Drawn

Dr. Hunsicker-Wang also analyzed the Diversity Map visualization with the lines across all the columns and noticing any consequential difference, if any, in interpreting or analyzing the data set. It was determined that this form of visualization did not aid in a deeper understanding of the Map. The majority of the information that was displayed in the lines version of the Map was already obtained from the block version. The only extra piece of information that she stated was slightly more interesting in this version was that for the first time, one could begin to visualize correlation between adjacent columns. In figure 12, she stated that the visualization showed that DNA molecules on average had lower molecular weights than those of proteins in that data set, which as you can see through the connecting lines, fall into a wider range of molecular weights than the rest of the macromolecular types. However, this kind of correlation does not span across columns because the lines drawn do not relay any extra quantitative information

regarding the amount of objects that fit the correlation visualized. Without this quantitative measurement, Dr. Hunsicker-Wang states that no additional information can be deduced from this data.

5 Conclusions

The Diversity Map has proven to be an overall effective way of looking at the data from the Protein Data Bank. The informal case study by a resident protein analyst, Dr. Hunsicker-Wang, has shown that the Map is a more powerful tool than an excel spreadsheet in looking at the entire Protein Data Bank at a higher level of abstraction. One is able to deduce trends if any in attributes across the Map and also between two data sets. Filtration of the data set gives a greater level of detail to analyze correlations between buckets of one data set to trends in the entire visualization. Compared to the excel spreadsheet, it is easier to spot out extreme cases of bucket color intensity, making it immediately evident upon looking at the Map which values have the highest and lowest amount of objects in them. Upon a more careful analysis of the Map, one can notice trends that would have been very difficult to point out in excel spreadsheets. For example the shift in the peak of the normal distribution of resolutions from the 2.0-2.2 range to the 1.8-2.0 range can be identified using the Map. This was found without any complex form of data manipulation that would have been required of the user if using an excel spreadsheet.

The Diversity Map, however, lacks a level of detail that can be found in an Excel spreadsheet. In an excel spreadsheet, the user can observe all information about one specific object in the data set. This is impossible to do using the Diversity Map. In fact, the Diversity Map was not designed to look at specific objects; it was made to describe the diverse nature of entire data sets. This trade-off between detail and abstraction is something that needs to be discussed in terms of

level of importance to the scientific user. Dr. Hunsicker-Wang mentioned in one of our meetings that this visualization technique could be seen as a significant compliment to an excel spreadsheet (or any kind of visualization - graph that could be derived from the data).

Nevertheless, in the birthing stage of the Diversity Map, this research shows that the Diversity Map is applicable to the Protein Data Bank and has promise for being useful to scientists interested in research related to the Bank. The next step would be to formalize the participant study, add more features to the Map, as well as have a more comprehensive list of attributes displayed on the Diversity Map. Inferring from this research, one can predict that further studies will solidify the idea that Diversity Map, the visualization technique from Oregon State, may be a tool for making our constantly growing amount of biological and chemical data considerably more feasible to understand and manipulate.

Bibliography

Campagna-Slater, V., ARROWSMITH, a., Zhao, Y., Schapira, M. "Phamacophore Screening of the Protein Data Bank for Specific Binding Site Chemistry." *Journal of Chemical Information and Modeling* 50 (2010): 358-367.

Kirchmair, J., Markt, P., Distinto, S., Schuster, D., Liedl, K., Langer, T., Wolber, G. "The Protein Data Bank (PDB), Its Related Services and Software Tools as Key Components for In Silico Guided Drug Discovery." *Journal of Medical Chemistry* 51 no. 22 (2008): 7021-7040.

Martin, A., Ochagavia, M., Rabasa, L., Miranda, J., Fernandez-de-Cossio, J., Bringas, R. "BioGenet: a new tool for gene network building, visualization and analysis." *BMC Bioinformatics* (2010).

Nedelkov, Dobrin. "Discovering Human Protein diversity."*Macedonian Journal of Chemistry and Chemical Engineering* 27 no. 2 (2008): 99-106.

Pham, T., Hess, R., Ju, C., Zhang, E., Metoyer, R. "Visualization of Diversity in Large Multivariate Data Sets." *Visualization and Computer Graphics, IEEE Transactions on*16 no. 6 (2010): 1053 - 1062

Appendix

Perl Script for Data Preparation

```
open(FILE, $input) or die("could not open file");
open(OUTPUT, "> $output") or die("could not open output file");

my $line;
my $text=<FILE>;
my $next;

sub modify {
	my ($text)=shift;
	$$text =~ s/SOLID-STATE NMR/NMR/gi;
	$$text =~ s/FIBER DIFFRACTION/DIFF./gi;
	$$text =~ s/ELECTRON MICROSCOPY/ELEC MICR/gi;
	$$text =~ s/ELECTRON CRYSTALLOGRAPHY/ELEC CRYS/gi;
	$$text =~ s/SOLUTION NMR/NMR/gi;
	$$text =~ s/SOLUTION NMR\, XRAY DIFF/MIX/gi;
	$$text =~ s/\([A-Za-z0-9]\)//gi;
	$$text =~ s/1990\-[0-9]+\-[0-9]+/1990/gi;
	$$text =~ s/1991\-[0-9]+\-[0-9]+/1991/gi;
	$$text =~ s/1992\-[0-9]+\-[0-9]+/1992/gi;
	$$text =~ s/1993\-[0-9]+\-[0-9]+/1993/gi;
	$$text =~ s/1994\-[0-9]+\-[0-9]+/1994/gi;
	$$text =~ s/1995\-[0-9]+\-[0-9]+/1995/gi;
	$$text =~ s/1996\-[0-9]+\-[0-9]+/1996/gi;
	$$text =~ s/1997\-[0-9]+\-[0-9]+/1997/gi;
	$$text =~ s/1998\-[0-9]+\-[0-9]+/1998/gi;
	$$text =~ s/1999\-[0-9]+\-[0-9]+/1999/gi;
	$$text =~ s/2000\-[0-9]+\-[0-9]+/2000/gi;
	$$text =~ s/2001\-[0-9]+\-[0-9]+/2001/gi;
	$$text =~ s/2002\-[0-9]+\-[0-9]+/2002/gi;
	$$text =~ s/2003\-[0-9]+\-[0-9]+/2003/gi;
	$$text =~ s/2004\-[0-9]+\-[0-9]+/2004/gi;
	$$text =~ s/2005\-[0-9]+\-[0-9]+/2005/gi;
	$$text =~ s/2006\-[0-9]+\-[0-9]+/2006/gi;
	$$text =~ s/2007\-[0-9]+\-[0-9]+/2007/gi;
	$$text =~ s/2008\-[0-9]+\-[0-9]+/2008/gi;
	$$text =~ s/2009\-[0-9]+\-[0-9]+/2009/gi;
	$$text =~ s/2010\-[0-9]+\-[0-9]+/2010/gi;
	$$text =~ s/ \*[ ]DNA/DNA/gi;
	$$text =~ s/ \*[ ]RNA/RNA/gi;
	$$text =~ s/ \*[ ]Protein/Protein/gi;
```

```
        $$text =~ s/\"ProteinDNA\"/Protein\/DNA/gi;
        $$text =~ s/\"ProteinRNA\"/Protein\/RNA/gi;
        $$text =~ s/DNARNA/DNA\/RNA/gi;
        $$text =~ s/DNA\/RNA Hybrid/DNA\/RNA/gi;
        $$text =~ s/ProteinDNA\/RNA Hybrid/Protein\/DNA\/RNA/gi;
        $$text =~ s/X-RAY DIFFRACTION/XRAY DIFF/gi;
}
while($next=<FILE>){

        if ($next =~ /^ \*/){
        chomp $text;
                $text = $text.$next;
                $next = <FILE>;
        }

        modify(\$text);

        print OUTPUT $text;
        $text = $next;
}

modify(\$next);
print OUTPUT $next;

close(OUPUT);
close(FILE);
```

MySQL Query for Data Preparation

```
UPDATE `data set format 1990-1992`
SET
`Space Group` = (CASE WHEN (`Space Group` IS NULL or `Space Group` = "") THEN "null"
else `Space Group` END),
`Exp. Method` = (CASE WHEN (`Exp. Method` IS NULL or `Exp. Method` = "") THEN "null"
else `Exp. Method` END),
`Dep. Date` = (CASE WHEN (`Dep. Date` IS NULL or `Dep. Date` = "") THEN "null" else
`Dep. Date` END),
`Structure MW` = (CASE WHEN (`Structure MW` IS NULL or `Structure MW` = "") THEN
"null" else `Structure MW` END),
`Macromol. Type` = (CASE WHEN (`Macromol. Type` IS NULL or `Macromol. Type` = "")
THEN "null" else `Macromol. Type` END),
```

```
`Resolution` = (CASE WHEN (`Resolution` IS NULL or `Resolution` = "") THEN 20 else
`Resolution` END);

UPDATE `data set format 1990-1992`
SET Resolution =
(CASE
when Resolution = 'null' then Resolution = "null"
WHEN Resolution >= 0.00 AND Resolution < 0.20 THEN '0.0-0.2'
when Resolution >= 0.20 AND Resolution < 0.40 then '0.2-0.4'
when Resolution >= 0.40 AND Resolution < 0.60 then '0.4-0.6'
when Resolution >= 0.60 and Resolution < 0.80       then '0.6-0.8'
when Resolution >= 0.80 AND Resolution < 1.00 then '0.8-1.0'
when Resolution >= 1.00 AND Resolution < 1.20 then '1.0-1.2'
when Resolution >= 1.20 and Resolution < 1.40       then '1.2-1.4'
when Resolution >= 1.40 AND Resolution < 1.60 then '1.4-1.6'
when Resolution >= 1.60 AND Resolution < 1.80 then '1.6-1.8'
when Resolution >= 1.80 and Resolution < 2.00       then '1.8-2.0'
when Resolution >= 2.00 AND Resolution < 2.20 then '2.0-2.2'
when Resolution >= 2.20 AND Resolution < 2.40 then '2.2-2.4'
when Resolution >= 2.40 and Resolution < 2.60       then '2.4-2.6'
when Resolution >= 2.60 AND Resolution < 2.80 then '2.6-2.8'
when Resolution >= 2.80 and Resolution < 3.00       then '2.8-3.0'
when Resolution >= 3.00 AND Resolution < 3.20 then '3.0-3.2'
when Resolution >= 3.20 AND Resolution < 3.40 then '3.2-3.4'
when Resolution >= 3.40 and Resolution < 3.60       then '3.4-3.6'
when Resolution >= 3.60 AND Resolution < 3.80 then '3.6-3.8'
when Resolution >= 3.80 and Resolution < 4.00       then '3.8-4.0'
when Resolution >= 4.00 and Resolution < 20.0 then '4.0-20.0'
when Resolution >= 20.0 then '20.0'
END)

UPDATE `data set format 1990-1992`
SET `Structure MW` =
(CASE
WHEN `Structure MW` >= 0 AND `Structure MW` < 5000 THEN '0-5000'
WHEN `Structure MW`         >= 5000 and `Structure MW` < 10000 THEN '5000-10000'
when `Structure MW` >= 10000 and `Structure MW` < 15000 THEN '10000-15000'
when `Structure MW` >= 15000 and `Structure MW` < 20000 THEN '15000-20000'
when `Structure MW` >= 20000 and `Structure MW` < 25000        then '20000-25000'
when `Structure MW` >= 25000 and `Structure MW` < 30000        then '25000-30000'
WHEN `Structure MW` >= 30000 and `Structure MW` < 35000     then '30000-35000'
WHEN `Structure MW` >= 35000 and `Structure MW` < 40000 then '35000-40000'
when `Structure MW` >= 40000 and `Structure MW` < 45000 then '40000-45000'
when `Structure MW` >= 45000 and `Structure MW` < 50000 then '45000-50000'
when `Structure MW` >= 50000 and `Structure MW` < 55000 then '50000-55000'
when `Structure MW` >= 55000 and `Structure MW` < 60000 then '55000-60000'
```

```
when `Structure MW` >= 60000 and `Structure MW` < 65000 then '60000-65000'
when `Structure MW` >= 65000 and `Structure MW` < 70000 then '65000-70000'
when `Structure MW` >= 70000 and `Structure MW` < 75000 then '70000-75000'
when `Structure MW` >= 75000 and `Structure MW` < 80000 then '75000-80000'
when `Structure MW` >= 80000 and `Structure MW` < 85000 then '80000-85000'
when `Structure MW` >= 85000 and `Structure MW` < 90000 then '85000-90000'
when `Structure MW` >= 90000 and `Structure MW` < 95000 then '90000-95000'
when `Structure MW` >= 95000 and `Structure MW` < 100000 then '95000-100000'
when `Structure MW` >= 100000 and `Structure MW` < 105000 then '100000-105000'
when `Structure MW` >= 105000 and `Structure MW` < 110000 then '105000-110000'
when `Structure MW` >= 110000 and `Structure MW` < 115000 then '110000-115000'
when `Structure MW` >= 115000 and `Structure MW` < 120000 then '115000-120000'
when `Structure MW` >= 120000 and `Structure MW` < 125000 then '120000-125000'
when `Structure MW` >= 125000 and `Structure MW` < 130000 then '125000-130000'
when `Structure MW` >= 130000 and `Structure MW` < 135000 then '130000-135000'
when `Structure MW` >= 135000 and `Structure MW` < 140000 then '135000-140000'
when `Structure MW` >= 140000 and `Structure MW` < 145000 then '140000-145000'
when `Structure MW` >= 145000 and `Structure MW` < 150000 then '145000-150000'
when `Structure MW` >= 150000 and `Structure MW` < 155000 then '150000-155000'
when `Structure MW` >= 155000 and `Structure MW` < 160000 then '155000-160000'
when `Structure MW` >= 160000 and `Structure MW` < 165000 then '160000-165000'
when `Structure MW` >= 165000 and `Structure MW` < 170000 then '165000-170000'
when `Structure MW` >= 170000 and `Structure MW` < 175000 then '170000-175000'
when `Structure MW` >= 175000 and `Structure MW` < 180000 then '175000-180000'
when `Structure MW` >= 180000 and `Structure MW` < 185000 then '180000-185000'
when `Structure MW` >= 185000 and `Structure MW` < 190000 then '185000-190000'
when `Structure MW` >= 190000 and `Structure MW` < 195000 then '190000-195000'
when `Structure MW` >= 195000 and `Structure MW` < 200000 then '195000-200000'
when `Structure MW` >= 200000 then 'Over 200000'
END)
```

MySQL Query for Data Integration

```
SELECT `data set format 1990-1992`.`Macromol. Type` FROM `data set format 1990-1992`
        GROUP BY `data set format 1990-1992`.`Macromol. Type`;
```

Perl Script for Data Integration

```
#usr/local/bin/perl

$inputfile = "meta convert file.txt";
$outputfile = "data_org_meta.txt";
```

```
print "$inputfile\n";
print $outputfile;
open(FILE, $inputfile);
open(OUTPUTFILE, ">> $outputfile");

my $line;
while($text=<FILE>){
        chomp $text;
        print OUTPUTFILE "Macromol. Type        $text\n";
};

close(FILE);
```