Big Molecules Meet Big Data

Predicting Protein Tertiary Structure from Combinatorial Chemistry Data

1 ABSTRACT

Much has been written in both the popular and business press about Big Data and its use. Less well reported, but just as revolutionary, has been the development of Combinatorial Chemistry methods used to carry out large-scale sequencing projects (such as the Human Genome Project). These projects have produced massive amounts of data relating to protein structure. The experimental determination of the protein structure is typically time consuming and relatively expensive. Hence, statistical methods for the prediction of some structural parameters of these proteins would be valuable. This paper discusses the prediction of one property, the Root Mean Square Deviation, from some measured and calculated protein attributes using the R statistical programming language.

2 BACKGROUND

We will begin the analysis with a very brief explanation of Combinatorial chemistry and how the data created by it is collected and stored. We will next discuss protein structure and its importance, giving two medical examples of each. Finally, we will analyze a large, real world protein data set, originally posted to the World Wide Protein Data Bank (wwPDB.org).

2.1 COMBINATORIAL CHEMISTRY

Combinatorial chemistry comprises synthetic methods that make it possible to prepare large numbers of compounds (compound libraries) in an integrated, continuous, multi-step process, making heavy use of robotics and automated data collection. It can be used for the polymerization of amino acids into peptides and finally into proteins. Combinatorial chemistry has found widespread use not only in academic research but also in the pharmaceutical industry where it is used for drug discovery and development. None of this would have been possible without the implementation of methods for collecting, storing, retrieving and analyzing large amounts of data. The process is summarized in Figure 1, below.

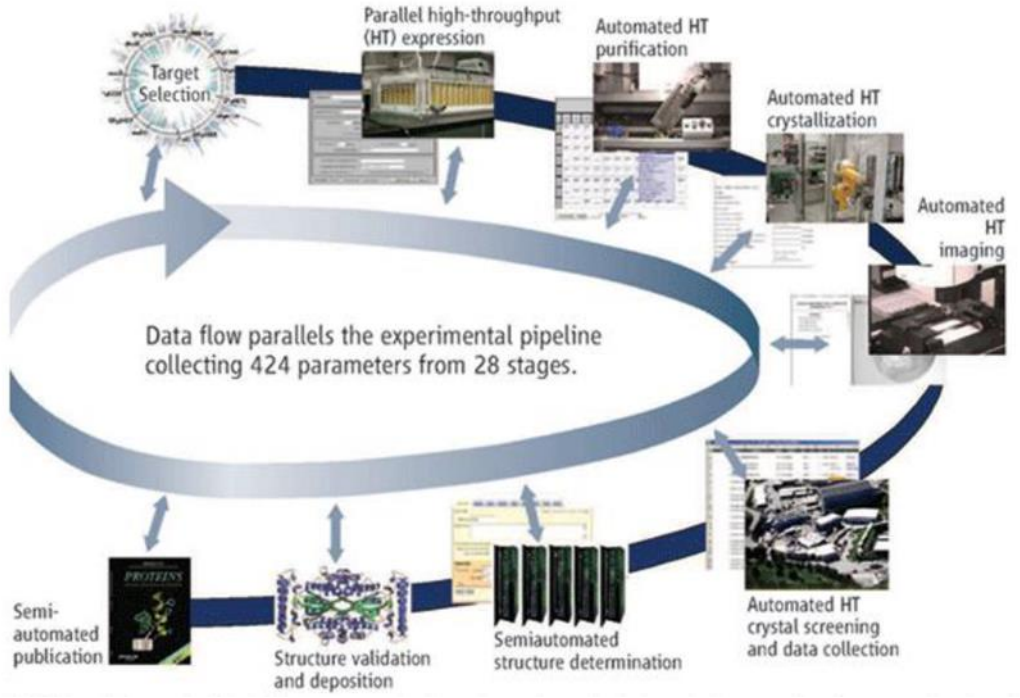

Figure 1
Automated Protein Synthesis and Data Collection

2.2 PROTEIN STRUCTURE

Proteins are composed of a specific sequence of amino acids, linked head-to-tail in long, linear molecules. Due to intra-molecular and inter-molecular interactions, these molecules form complex, three-dimensional conformations described by their primary, secondary, tertiary and quaternary structures. These structures are defined as follows:

- Primary – the linear sequence of its constituent amino acids
- Secondary – the local three-dimensional arrangement of the backbone atoms, without regard to the conformation of the side chains
- Tertiary – the three dimensional structure, or conformation, of the entire molecule
- Quaternary – the three dimensional arrangement of two or more associated molecules

This is depicted in figure 2, below, where the molecule forms a spiral secondary structure which then folds into tertiary structure, which then joins with three other molecules into a final, globular protein.

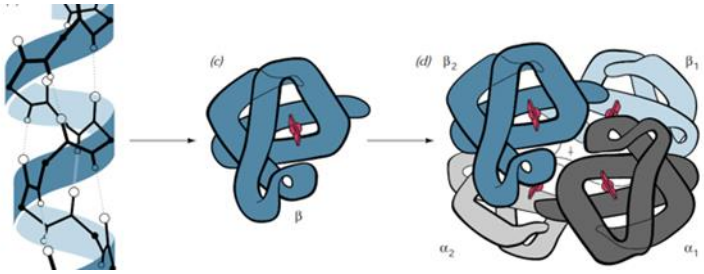

Figure 10: Secondary to Tertiary to Quaternary Structure

2.3 MECHANISMS AND FORCES

There are a number of chemical mechanisms which can contribute to a protein's stability and conformation, including hydrogen bonding, ion pairing and hydrophobicity. However, when considering tertiary structure, the first two are relatively unimportant. They have been for the most part "used up" in the forming the Secondary structure. That leaves hydrophobicity as the primary determinant of Tertiary structure. The greater a side chain's hydrophobicity, the more likely it is to occupy the interior of a protein. However, not all of these hydrophobic side chains can make their way into the interior and some will be left exposed. This situation is described by measured and calculated parameters such as Total Surface Area and Non polar Exposed Area (sic). We will use these parameters as factors in our study.

2.4 REAL WORLD RELEVANCE

All this is of more than just theoretical importance. For example, Alzheimer's disease is characterized by the precipitation of single protein molecules. These protein molecules are misfolded versions of molecules that are normally present in the same tissues.

Probably the best-known example however occurs due to the substitution of a hydrophobic amino acid (Valine) for an acidic, hydrophilic amino acid (Glutamic Acid) in the β-chain of hemoglobin. This change of a single amino acid alters the structure of the hemoglobin molecule in such a way that precipitation occurs within the red blood cell, leading to a condition known as Sickle Cell Anemia. The structure of the Hemoglobin molecule is given below as Figure 3 and the two amino acids are given in Figure 4.

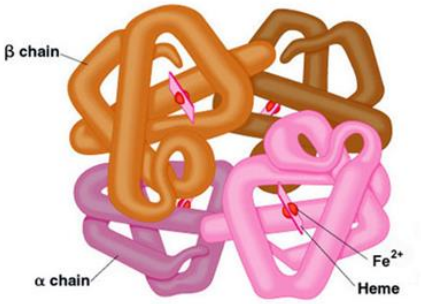

Figure 3: Quaternary Structure of Hemoglobin

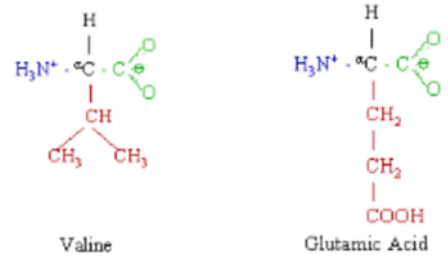

Figure 4: Substitution of Valine for Glutamic Acid results in Sickle Cell Anemia

2.5 ROOT-MEAN-SQUARE-DEVIATION OF ATOMIC POSITIONS

The above obviously results in some very complex issues for analysis and characterization of proteins. One summary statistic which has come into use is the Root-Mean-Square-Deviation or the RMSD of the protein. It is defined mathematically as

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2}$$

Figure 5: Root-Mean-Square-Deviation

where delta is distance between corresponding alpha Carbon atoms on superimposed molecules and is given as units of length, usually Angstrom units or 1 x 10exp -10 meters. The RMSD is apparently quite useful as it appears in 5.8% of all PDB entries. The concept is illustrated in Figure 13, below, with the side chains omitted.

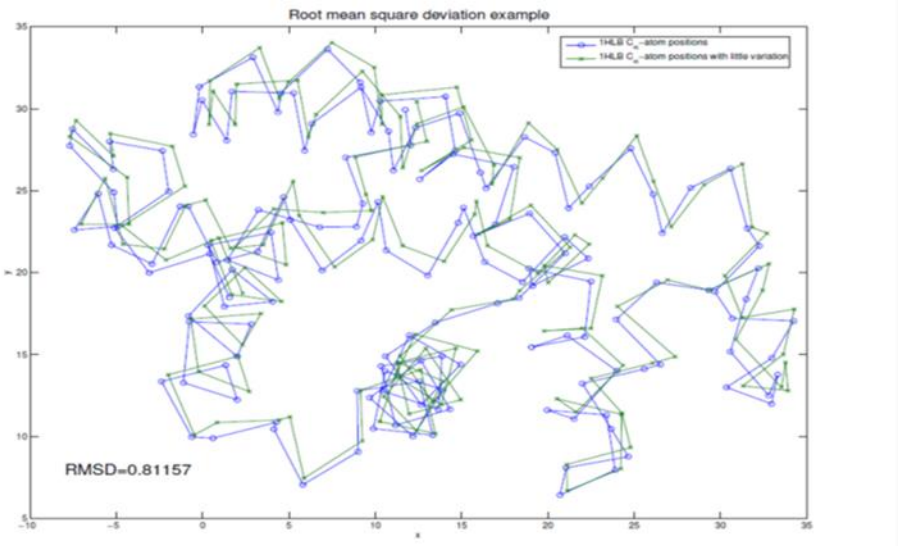

Figure 5: Example of Root-Mean-Square Deviation

We are now finally read to do the actual analysis.

3 DATA SET

The data set used in this study is the "Physicochemical Properties of Protein Tertiary Structure Data Set" from the Machine Learning Repository, University of California Irvine: http://archive.ics.uci.edu/ml/datasets/Physicochemical+Properties+of+Protein+Tertiary+St ructure. It was originally published in CASP 5-9 (Critical Assessment of protein Structure Prediction) and then posted to wwPDB.org.

This data set was compiled from direct measurements and from calculated values taken under controlled, laboratory conditions. Hence, little cleaning was required. This data set has however two issues which will make the extraction of useful information a bit difficult from an analyst's point of view.

The first issue is that there was no documentation supplied with the data set. Therefore, variables such as "fractional area of exposed non polar residue" had to be researched. The second issue was that several of the variables contained some reference to "exposed area" and hence were suspected to be correlated to at least some degree. (This was later found to be true, see below.)

A table of the independent variables (factors) and their description is given below as Figure 6.

F1 - Total surface area
F2 - Non polar exposed area
F3 - Fractional area of exposed non polar residue
F4 - Fractional area of exposed non polar part of residue
F5 - Molecular mass weighted exposed area
F6 - Average deviation from standard exposed area of residue
F7 - Euclidian distance
F8 - Secondary structure penalty
F9 - Spacial Distribution constraints (N,K Value)

Figure 6: Table of Factors

Other than the concerns mentioned above, none of the independent variables was considered to be more significant than the others, nor were there any apparent limitations to answering the fundamental question posed, i.e., predicting the RMSD from the chemical structures and properties.

4 ANALYSIS

The following description of the analysis is given without the complete R code, which can be found on GitHub (R code - GitHub) and as an addendum to this report.

4.1 READ IN AND EXAMINE DATA SET; CHECK FOR MISSING VALUES

The data was read into Rstudio using read.csv(), then checked for missing values using na.omit() and length().

4.2 Rename Column HEADINGS AND Create Table to Link to Definitions

The independent variables were changed to conform with R naming conventions, i.e., begin variable names with a lower case letter using names().

The result is displayed below as Figure 7, below.

Variable	Description
RMSD	Root Mean Square Deviation - dependent variable
f1	Total surface area
f2	Non polar exposed area
f3	Fractional area of exposed non polar residue
f4	Fractional area of exposed non polar part of residue
f5	Molecular mass weighted exposed area
f6	Average deviation from standard exposed area of residue
f7	Euclidian distance
f8	Secondary structure penalty
f9	Spacial Distribution constraints (N,K Value)

Figure 7: Variable Names

4.3 Exploratory Data Analysis

Preliminary investigations showed that the data set was clean and tidy and had no missing values.

4.3.1 Plot all two-dimensional combinations of all variables and display in a grid

From past experience we knew that generating lattice plots from the entire data set would yield graphs which would be too dense (even using the alpha function) to be useful. So we selected a random sample of 100 using the sample() function. We next used the plot() function to yield a lattice plot giving us the visual representation of our two factor relationships in Figure 8, below.

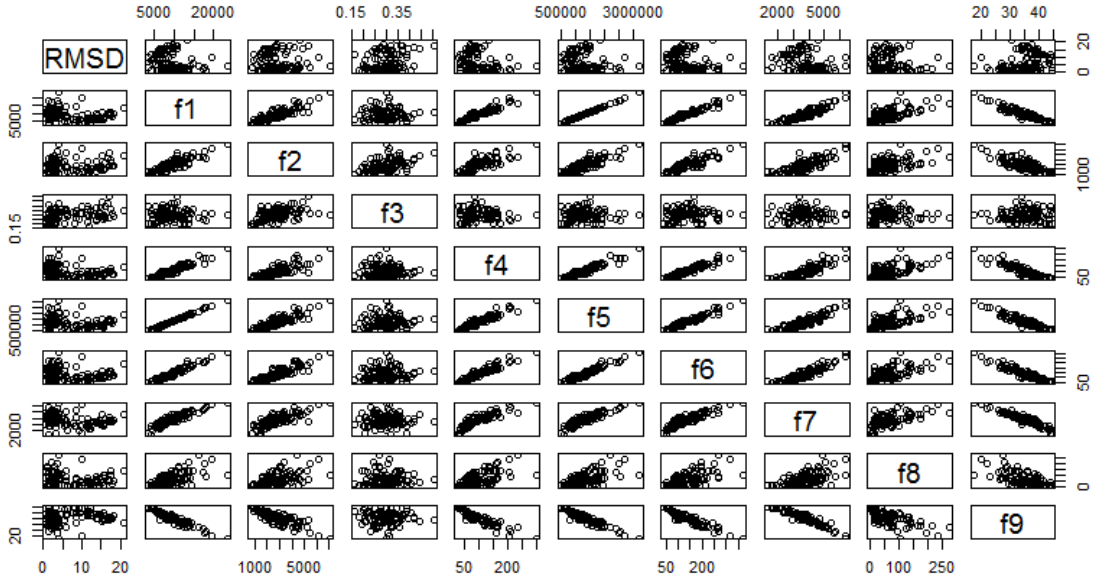

Figure 8: Lattice Plot of Two Factor Relationships

We noted the high degree of correlation of all "surface" variables and that the Euclidian distance (f7) correlations always appeared in a tight band. An examination of correlation coefficients using cor() did indeed find a high degree of correlation between several variable pairs, with absolute values well above 0.8. A copy of the correlation matrix is given as Figure 8, below.

	RMSD	f1	f2	f3	f4	f5	f6	f7	f8	f9
RMSD	1.00	-0.02	0.16	0.37	-0.17	-0.01	-0.04	0.00	0.00	0.06
f1	-0.02	1.00	0.91	0.13	0.93	1.00	0.97	0.55	0.65	-0.90
f2	0.16	0.91	1.00	0.50	0.79	0.90	0.91	0.52	0.58	-0.79
f3	0.37	0.13	0.50	1.00	0.03	0.12	0.20	0.08	0.10	-0.07
f4	-0.17	0.93	0.79	0.03	1.00	0.93	0.94	0.49	0.68	-0.89
f5	-0.01	1.00	0.90	0.12	0.93	1.00	0.96	0.55	0.64	-0.90
f6	-0.04	0.97	0.91	0.20	0.94	0.96	1.00	0.54	0.66	-0.88
f7	0.00	0.55	0.52	0.08	0.49	0.55	0.54	1.00	0.35	-0.52
f8	0.00	0.65	0.58	0.10	0.68	0.64	0.66	0.35	1.00	-0.64
f9	0.06	-0.90	-0.79	-0.07	-0.89	-0.90	-0.88	-0.52	-0.64	1.00

Figure 9
Correlation Matrix

We also noted that RMSD gives a bifurcated plot with two distinct areas (which may be due to interactions between factors) and is not easily described by the usual functions, e.g., linear polynomial, log, etc.

When plotted as a function of, for example, total surface area (f1), using qplot() from the ggplot2 package, the result was Figure 10, below.

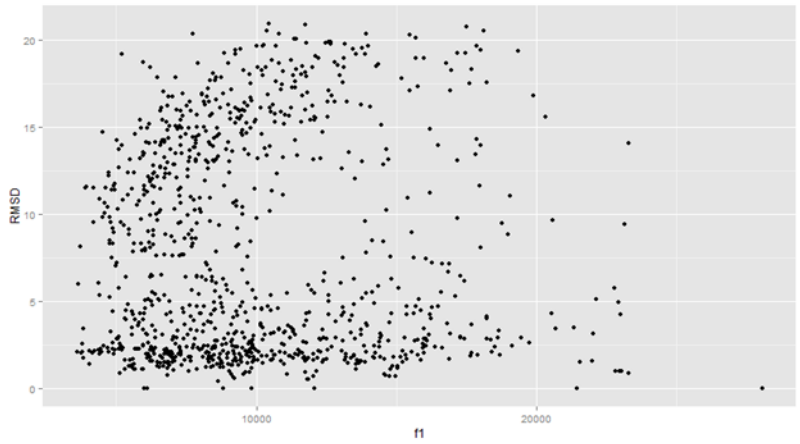

Figure 10: RMSD as a function of Total Surface Area (f1) of a 1000 record random sample

Apparently, there are a set of conditions which give a group of RMSDs with a mean of about 2 and a slope of about 0, and another set which gives a roughly logarithmic relationship with a mean of about 10 and a slope of about 1. This situation may be more suited to a Classification Tree or K Nearest Neighbors analysis.

4.3.2 Develop and Examine Descriptive Statistics

Next we examined the same data by using the describe() function from the psych package, followed by the select() function from the dplyr package. The result is given as Figure 12, below.

Variable	mean	sd	skew	kurtosis	median	range
RMSD	7.75	6.12	0.57	-1.14	5.03	21.00
f1	9871.60	4058.14	1.09	1.31	8898.81	37642.85
f2	3017.37	1464.32	1.19	2.00	2668.15	14908.50
f3	0.30	0.06	0.24	0.09	0.30	0.49
f4	103.49	55.42	1.23	1.33	87.74	359.01
f5	1368299.02	564036.69	1.06	1.18	1237219.06	5152521.19
f6	145.64	70.00	1.12	1.08	126.18	566.44
f7	3989.76	1993.57	20.76	804.58	3840.17	105948.17
f8	69.98	56.49	1.68	3.31	54.00	350.00
f9	34.52	5.98	-0.47	-0.25	35.30	40.07

Figure 11: Some Descriptive Statistics of the protein Data Set

Note the very high kurtosis of over 800 for Euclidean distance (f7), two orders of magnitude over any reasonable value. We examine the situation below.

4.3.3 Investigate High Kurtosis for f7, Euclidian Distance

First we checked distribution of the entire data set using histogram(), which gives the graph displayed as Figure 10, below.

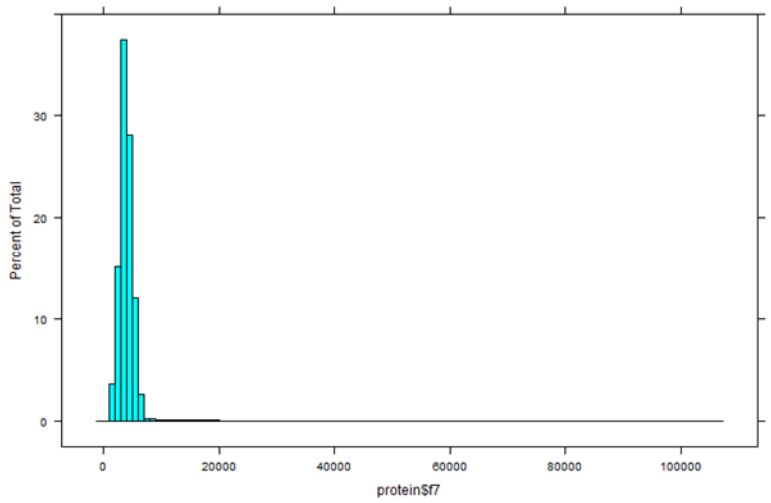

Figure 12: Distribution of Euclidian Distance (f7)

We see a very small percentage of outliers between about 10, 000 and 20,000 and apparently extending beyond 100,000 as well.

We next took a random sample using the sample() function which yielded the Graph in Figure 13, below.

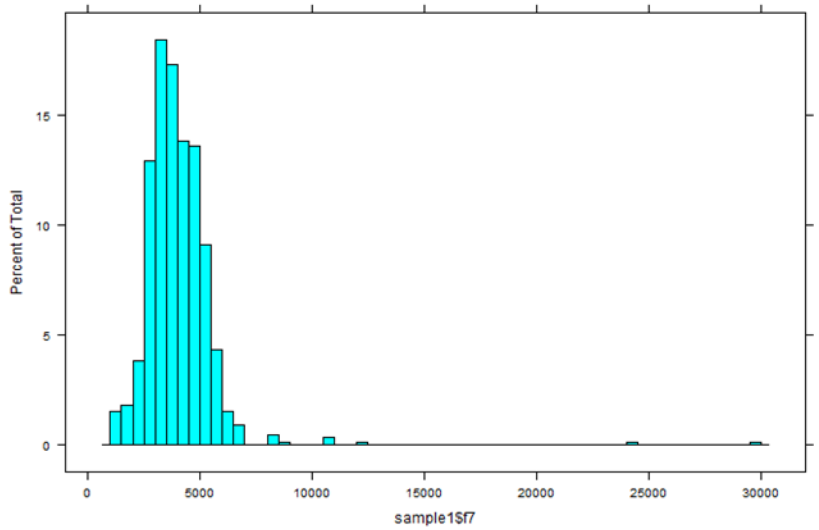

Figure 13: Distribution of f7 Euclidian Distance Factor – sample size of 1000

This confirms our suspicions, but since the percentage is small, we will ignore them for this stage of our investigation.

4.3.4 Investigate RMSD = 0

As noted previously there are points where RMSD is 0. According to the wwPDB, this is an allowed value. However, we checked some of those points to verify that there are no 0 values in another independent variable which could lead to an erroneous 0 value in the RMSD. We did this by sampling and using the subset() function. An abbreviated table of the results are given below as Figure 15, below. While not an exhaustive test this, when coupled with the knowledge that 0 is an allowed value, gives us confidence that these values are indeed real.

Row	RMSD	f1	f2	f3	f4	f5	f6	f7	f8	f9
25418	0.00	6009.31	1376.35	0.23	46.73	820019.70	77.10	3164.74	25.00	46.55
42697	0.00	12059.70	2312.27	0.19	158.69	1692984.90	170.14	4679.78	141.00	29.44
21794	0.00	8825.95	2563.49	0.29	85.48	1213045.50	145.09	3343.21	9.00	36.59
39195	0.00	28038.20	10897.00	0.39	313.07	3734152.50	430.70	8165.10	221.00	19.68
41418	0.00	6107.61	2095.28	0.34	57.55	854829.60	90.59	3550.41	27.00	37.12
9327	0.00	9812.62	2009.95	0.20	101.20	1424683.70	132.95	3770.70	61.00	46.55
1461	0.00	21427.30	8183.98	0.38	202.44	2828193.50	360.80	6136.42	220.00	22.19

Figure 14: Investigation of RMSD = 0

5 REGRESSION ANALYSIS

We started with the factor which has the highest correlation with dependent variable RMSD see section 4.3) and added factors in order of decreasing correlation coefficients until the Adjusted R-squared levels off. Next we tried some non-linear functions, followed by a two-factor interaction term of the most significant factors. All models but one gave coefficients of Pr(>|t|) <2e-16 and so were highly significant. The results are summarized in the Figure 15, below.

Model	Type	Factors	Adj R-square
rmsd1	linear	f3	0.14
rmsd2	linear	f3 + f4	0.17
rmsd3	linear	f3 + f4 +f2	0.26
rmsd4	linear	f3 + f4 + f2 + f9	0.26
rmsd5	linear	f4 + f2	0.26
rmsd6	square	f2^2	0.14
rmsd7	log	log(f3)	0.12
rmsd8	interaction	f4 + f2 + f4*f2	0.26

Figure 15: Summary of Regression Analysis.

The model chosen as the best model is rmsd5 since it gives an R-square of 0.26 and has the fewest factors.

rmsd5: y = 6.90 – 8.76 f4 + 3.28 f2

6 RESULTS AND DISCUSSION

Our results may be summarized as follows:

- Significant factors were all some measurement of an exposed non-polar surface area.
- The highest R-squares (0.026) were achieved by four models, all of which used some combination of f2, f3 and f4.
- The chose model was rmsd5 since it yielded an R-square of 0.26 and was a function of only two factors, f2 and f4.
- Adding an interaction term did not add any predictive value.
- In comparison to studies of other large molecules, e.g., commercial polymers which can yield R-squares of over 0.90, the results may seem less than satisfactory. However, considering the complexity of the protein molecules under investigation and the fact that five of the nine factors were really measurements of some non-polar surface area, the results were quite good. We were able to predict over one quarter of the variation from the mean of a diverse group of very large protein molecules with very little information.

7 FUTURE WORK

Since the focus of this work was not only to predict the RMSD from the factors provided, but also to understand the physical and chemical forces which caused those RMSDs, more domain research is required. In particular:

- Determination of exactly how the various non-polar exposed surface area factors were measured and calculated. There may be more basic factors “upstream” from the factors used which would be better predictors of RMSD.
- More research into the vast data store that is the wwPDB and its specialized sub-domains.
- More literature research for previous work on RMSD prediction.

Note that the above recommendations deal with specific domain research. This is more important in some areas than in others.

Finally, investigate the underlying nature of the bifurcation of the RMSDs as noted in Section 4.3.1. One method would be to use the caret package to split the data, and then perform classification and regression testing on each individual group. This could lead to results which are not only statistically satisfying but also useful in predicting protein tertiary structure.

8 ACKNOWLEDGEMENTS

I would like to thank Matthew Kent for his advice regarding this study, especially for suggesting the caret package.

9 ADDENDUM – R CODE

```
# Read in and Examine Data Set #
protein <- read.csv("~/R_projects/Data/Capstone Project/Protein Tertiary Structure.csv")
head(protein, 20)
str(protein)
length(protein$RMSD)

# Replace Column Names for ease of typing and conform with naming standards #
protein_names_new <- c('RMSD', 'f1', 'f2', 'f3', 'f4', 'f5', 'f6', 'f7', 'f8', 'f9')
protein_names_new
names(protein) <- protein_names_new
head(protein)


# Create an internal reference table (for humans) for column headings and their definitions #
# Note: two typos from original were carried through to maintain consistency #

# Create two vectors, variable and definition then cbind into a description table #
variable <- c('f1', 'f2', 'f3', 'f4', 'f5', 'f6', 'f7', 'f8', 'f9')
definition <- c('Total surface area', 'Non polar exposed area',
                'Fractional area of exposed non polar residue',
                'Fractional area of exposed non polar part of residue',
                'Molecular mass wieghted exposed area',
                'Average deviation from standard exposed area of residue',
                'Euclidean distance', 'Secondary structure penalty',
                'Spacial Distribution constraints(N,K, Value)')
descriptions_table <- cbind(variable, definition)
descriptions_table

# Check for missing values and visually inspect length #
protein1 <- na.omit(protein)
length(protein$RMSD)
length(protein1$RMSD)
# Both are of length 45,730 #


# Collect sample2 as per sample 1, but with fewer points, for quick, exploratory plots #
# Use sample size rather than alpha = for dense data #
sample1 <- protein[sample(1:nrow(protein), "1000", replace = FALSE), ]
sample2 <- protein[sample(1:nrow(protein), "100", replace = FALSE), ]

# Get descriptive statistics on Columns #
# Select statistics most useful for this analysis #
df.describe <- describe(protein)
df.describe
```

```
df.describe_select <- select(df.describe, mean, sd, skew, kurtosis, median, range)
df.describe_select
# Note the very high kurtosis for f7, Euclidian distance #

# Examine f7 distribution with histograms #
histogram(x = sample1$f7, breaks = 50)
histogram(x = sample2$f7, breaks = 50)
# Note outliers at 8,000 out to 24,000 A.  Small % so ignore #
# Look at f7 over the entire data set #
histogram(x = protein$f7, breaks = 100)
# Very "sharp", centered distribution with a few outliers, so ignore for now #

# START ANALYSIS #

# Plot all variable two-dimensional conbinations and display in a grid #
# Use random samples to simplify and clarify #
plot(sample1) # too dense
plot(sample2)
# Note: high degree of correlation of all "surface" variables, Euclidian distance correlations
# always in a tight band #
# Note: RMSD gives a bifurcated plot with two distinct areas; two factor interaction. #

# Plot RMSD for further examination #
qplot(x = f1, y = RMSD, data = sample1, geom = 'point')
qplot(x = f1, y = RMSD, data = sample2, geom = 'point')

# Check for anomalies in records from Sample1 where RMSD  = 0 #
# Two ways for 0, RMSD actually is 0, missing value which leads to a 0 value #
sample1_zero <- subset(sample1, RMSD == 0)
sample1_zero


# Now Check descriptive statistics #
sample1_zero_describe <- describe(sample1_zero)
sample1_zero_describe_select <- select(sample1_zero_describe, mean, sd, skew, kurtosis, median,
range)
sample1_zero_describe_select
# Kurtosis is now "normal" which makes sense since RMSD = 0.  Notice NAN #

# Do same for entire data base, protein
protein_zero <- subset(protein, RMSD == 0)
head(protein_zero)
protein_zero_describe <- describe(protein_zero)
protein_zero_describe_select <- select(sample1_zero_describe, mean, sd, skew, kurtosis, median,
range)
protein_zero_describe_select
# Note: same results as sample 1 #
```

```
# Get correlation matrix in preparation for regression #
cor(protein[c('RMSD', 'f1', 'f2', 'f3', 'f4', 'f5',
          'f6', 'f7', 'f8', 'f9')])

# General Linear Regression #
# Start with factors which have higher correlation with dependent variable RMSD #
# Adjusted R-squared copied to Excel spreadsheet #
rmsd1 <- lm(RMSD ~ f3, data = protein)
summary(rmsd1)

rmsd2 <- lm(RMSD ~ f3 + f4, data = protein)
summary(rmsd2)

rmsd3 <- lm(RMSD ~ f3 + f4 + f2, data = protein)
summary(rmsd3)

rmsd4 <- lm(RMSD ~ f3 + f4 + f2 + f9, data = protein)
summary(rmsd4)

rmsd5 <- lm(RMSD ~ f4 + f2, data = protein)
summary(rmsd5)

# General Linear Regression with non-linear Terms #
rmsd6 <- lm(RMSD ~ f2^2, data = protein)
summary(rmsd6)

rmsd7 <- lm(RMSD ~ log(f3), data = protein)
summary(rmsd7)

rmsd8 <- lm(RMSD ~ f4 + f2 + f4*f2, data = protein)
summary(rmsd8)

rmsd9 <- lm(RMSD ~ f3 + f3*f2 + f3*f4, data = protein)
summary(rmsd9)

rmsd10 <- lm(RMSD ~ f3 + f7 + f3*f7, data = protein)
summary(rmsd10)
```