COMPARING X-RAY AND NMR PROTEIN STRUCTURES USING COMPUTATIONAL GEOMETRY

by

Steven Bowers

A Thesis
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Master of Science
Bioinformatics and Computational Biology

Committee:

Dr. Iosif Vaisman, Thesis Director

Dr. Dimitri Klimov, Committee Member

Dr. Alessandra Luchini, Committee Member

Dr. James D. Willett, Director, School of Systems Biology

Dr. Donna M. Fox, Associate Dean for Academic and Student Affairs, College of Science

Dr. Peggy Agouris, Dean, College of Science

Date: April 29, 2014

Spring Semester 2014
George Mason University
Fairfax, VA

Comparing X-RAY and NMR Protein Structures Using Computational Geometry

A Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at George Mason University

by

Steven Bowers
Bachelor of Arts
George Mason University, 1987

Director: Iosif Vaisman, Professor
Department of Bioinformatics and Computational Biology

Spring Semester 2014
George Mason University
Fairfax, VA


This work is licensed under a creative commons attribution-noderivs 3.0 unported license.

DEDICATION

This is dedicated to my loving wife Dena who keeps me out of trouble, and my sons Dennis and Andrew who get me into trouble.

ACKNOWLEDGEMENTS

I would like to thank the facility of the Bioinformatics department and my thesis committee, especially Dr. Vaisman, for all the help on this project and during my time at Mason.

TABLE OF CONTENTS

Page

List of Tables viii

List of Figures ix

Abstract x

Chapter One - Introduction 1

Chapter Two – Methods and Materials 4

- Section 1 - Materials 4
- Section 2 - Methods 4
  - Creating the Datasets 4
  - Determining Simplex Type 6
  - Determining Secondary Structure 7
  - Determining residue and Simplex Exposure 9
  - Determining Simplex consistency 9
  - Determining Log-likelihood Potential 9
  - Finding Common shapes in consistent simplexes 10
  - Assigning simplexes to shapes 13
  - Finding Relationships between Consistency and other properties 14
  - Matching X-ray to NMR files 14
  - Predicting Consistency levels of X-ray files 15

Chapter Three - Results 16

- Consistency of Simplexes 16
- Effect of simplex type and Secondary Structure on Consistency 17
- Consistency and Residue Exposure 27
- Consistency and Log-likelihood Potential 28
- Consistency and simplex shape 31
- Properties of Common Simplex Shapes 32
- Consistency of X-ray Structures 34

Predicting consistency from X-ray structure 36
Chapter Four - Conclusion 38
Possible future work 40
Appendix – Shape descriptions 41
References 45

LIST OF TABLES

Table Page
Table 1- Percentage of each secondary structure for each consistency level 18
Table 2 – Percentage of T4000 with no Secondary structure at each consistency level .. 20
Table 3 - Percentage of each secondary structure for each consistency level for T4000 . 20
Table 4 - Percentage of each secondary structure for each consistency level for T3100 . 21
Table 5- Percentage of each secondary structure for each consistency level for T2200 .. 23
Table 6 - Percentage of each secondary structure for each consistency level for T2110 . 24
Table 7 -- Percentage of each secondary structure for each consistency level for T1111 26
Table 8 - Percentage of T1111 with no Secondary structure at each consistency level ... 26
Table 9 - Percent of residues exposed at each consistency level 28
Table 10 - Average Log-likelihood Potential at each consistency level using 100% consistent simplexes..... 29
Table 11- Average Log-likelihood Potential at each consistency level using all NMR simplexes..... 29
Table 12 - Average Log-likelihood Potential at each consistency level using X-ray simplexes..... 30
Table 13- Average value and Standard Deviation of values in Log-likelihood Potentials tables 30
Table 14 - Distribution of simplexes which match a shape 31
Table 15 - Percent of simplexes at each consistency level which match a shape..... 32
Table 16 - Simplex properties at each consistency level 35
Table 17 - Percent of simplexes with secondary structure at each consistency level..... 35
Table 18- Consistency by type in X-ray NMR pairs 36
Table 19 - Edge Lengths of shapes 42
Table 20 - Secondary Structure of Shapes 43
Table 21- Properties of Shapes 44

LIST OF FIGURES

Figure Page
Figure 1 - Consistency Distribution for all simplexes 17
Figure 2- Distribution of Consistency Scores for T4000 Simplexes 19
Figure 3 - Distribution of consistency scores for T3100 simplexes 21
Figure 4 -Distribution of consistency scores for T2200 simplexes 22
Figure 5 - -Distribution of consistency scores for T2210 simplexes 24
Figure 6- -Distribution of consistency scores for T1111 simplexes 25
Figure 7- Log of ring density of Shape 3110_0 33

ABSTRACT

COMPARING X-RAY AND NMR PROTEIN STRUCTURES USING COMPUTATIONAL GEOMETRY

Steven Bowers, M.S.

George Mason University, 2014

Thesis Director: Dr. Iosif Vaisman

X-ray crystallography is widely used to solve high resolution protein structures, but only when the protein can be crystalized. NMR spectroscopy can be performed in solution, which is much more similar to the natural protein environment in a cell. X-ray and NMR structures of the same proteins, which are available from the Protein Data Bank, are largely similar, but not identical. By understanding the differences between NMR spectroscopy and X-ray crystallography structures, it may be possible to better understand the structure and function of proteins in the cells. Computational geometry analysis of nearest neighbor residues in different conformations of NMR ensemble is first used to identify the consistent parts of NMR structures and the factors which affect this consistency. X-ray and NMR structures of the same protein are then compared to pinpoint the differences and the factors which affect these differences. A number of geometrical and topological factors were identified which are linked to the consistency of

the simplexes across the conformations, including solvent accessibility, simplex residue content, secondary structure, shape of simplex, and type of simplex (based on sequence).

CHAPTER ONE - INTRODUCTION

X-ray crystallography is widely used to solve high resolution protein structures, but only when the protein is in the form of a crystal. NMR spectroscopy can be done in solution, which is much more similar to the natural protein environment in a cell. X-ray and NMR structures of the same proteins, which are available from the Protein Data Bank, are largely similar, but not identical. By understanding the differences in between structures generated NMR spectroscopy and X-ray crystallography structures, it may be possible to better understand the structure and function of proteins in the cell.

In this project Delaunay tessellation will be used to analyses the differences between conformers of the same protein generated through NMR spectroscopy. It will also be used to compare the results of X-ray crystallography and NMR spectroscopy when solving the same protein. The goal is to be able to better understand how proteins solved using X-ray structures will exist in solution.

The use of Delaunay tessellation was first proposed as a way of analyzing structures in 1996 (Singh, Tropsha,Vaisman). Delaunay tessellation is used to determine the nearest-neighbor for protein residues. The protein structure is divided into many non-overlapping tetrahedra. The 4 vertices of the tetrahedra are the alpha carbon of a 4 residue nearest-neighbor Delaunay simplex.

The most common use of Delaunay tessellation has been to predict the effect of the change of a single residue in a protein. Predicting the effect of a change is important in determining if a mutation may be disease causing. It is also important in protein engineering to determine if a change of a residue will make the structure more or less stable.

The program auto-mute (http://proteins.gmu.edu/automute/)(Masso, Vaisman, 2010) is used to predict the effects of residue changes. It works by first taking a set of non-redundant proteins, tessellating the proteins, and then determining the probability of each combination. The ratio of this probability to the probability of the 4 residues coming together randomly (based on the frequency of the residues found in the protein) is computed. The log of this number is used to predict if a residue modification will change the structure's stability, cause a change in protein activity, or cause disease. To make the prediction the difference in log scores of the simplexes is used with other properties (residue exposure, secondary structure, tetrahedra shape and volume) are analyzed using a random tree or Support Vector Machine.

Other programs such as Site Directed Mutator (SDM) (http://mordred.bioc.cam.ac.uk/~sdm/sdm.php), (Topham, Sirnivasan, Blundel, 1997) are also used to predict the effect of residue changes in a protein. SDM determines an environment for each residue based on secondary structure, torsion angles, solvent accessibility, and hydrogen bonding. A table of substitution scores for each environment (similar to a PAM, or BLOSUM table) had been built for each environment based on a training set of proteins. Based on the change in residue a score of the substitution is found

in the substitution table for the residue's environment. From the score, a change in stability is predicted.

Since auto-mute, SDM, and others base their predictions on data obtained from X-ray structures, a better understanding of how proteins exist in environments more similar to the environment of living creatures, could help improve algorithms for determining effects. A better understanding of protein structure could also help in applications such as protein modeling or determining a protein's function from structure.

In the past there have been other works which compared NMR and X-ray structures. A study was done comparing the structures by measuring the root-mean-square deviation (RMSD) (Sikic, Tomic, Carugo, 2010). The study found much higher RMSD between X-Ray and NMR structures, then between 2 X-ray structures of the same protein. The comparison of the differences within secondary structures (defined as helixes, strands, and loops) showed that the differences were highest for loops and smallest for strands. It is not surprising that the difference is highest for loops, but it is harder to explain why strands would have smaller RMSD then helixes.

.

CHAPTER TWO – METHODS AND MATERIALS

Section 1 - Materials

All protein structures for this project were obtained from PDB (http://pdb.org). DSSP structural assignments came from the DSSP database (http://www.cmbi.ru.nl/dssp.html) Open source software qhull was used to tessellate the proteins. A java program simplex (by Zhibin Lu, 2000) was used to format the input and output for the qhull program. Other software written in Python and C for Linux were written and used for this project to analyze the data.

Section 2 - Methods

There a number of steps in processing this data. The datasets must be determined. All the structures must be tessellated. The simplex type, secondary structure, residue exposure, log-likelihood potential, and consistency must be determined for each simplex. Also common shapes of consistent simplexes were found and simplexes were assigned to shapes. Once this was done, the relationship between the consistency and other properties was determined. Finally a program was used which would predict the consistency of the simplexes of an X-ray structure and compare the results to an NMR structure of the same protein.

Creating the Datasets

For this project 3 datasets were used. One is a set of structures which have been solved by both NMR and X-ray. The second is a group of NMR structures.

The first dataset is a group of NMR/X-ray structure chain pairs. The 2 chains in the pairs must have 100% sequence identity. The NMR structure must have at least 10 conformers which can be tessellated by the qhull software. The qhull software must also be able to tessellate the X-ray structure (Some structures in PDB are missing some residues in the middle, which makes it impossible for qhull to tessellate the structure. Structures will tessellate if residues are missing at the start or at the end). All pairs must have less than 30% sequence identity with all the other pairs.

To create this dataset a list of all Solution NMR structures with protein, but no RNA or DNA structures was created by a search of PDB The PDB search is: Macromolecule type – protein=yes, RNA=no, DNA=no, Contains DNA/RNA hybrid=no – Experimental Method=Solution NMR

This gave about 8900 structures. These structures were then tessellated. Structures which had fewer than 10 conformers which could be tessellated were removed from the dataset.

For each structure left in the dataset in a search was done on PDB for structures with 100% identity. For each structure with 100 percent identity a query was sent to PDB to check the experimental method. If the structure was an X-ray structure then the pair was added to the list of potential pairs. Once all the NMR structures were checked a list of potential pairs was created. For each pair in the list, a query was done to PDB to get the 30% sequence identity cluster of the NMR structure (the result contains all structures with greater than 30% identity). Any NMR structures which were returned from the query were removed from the list of potential pairs (The original structure in the query

was not removed). Once all the structures in the list were checked, there were about 300 structure pairs left.

The second dataset is a list of NMR structures with at least 10 conformers which could be tessellated. All of the structures would have a sequence identity of less than 30% with the structures in the first dataset. The dataset would also have less than 30% identity with other structures in the set. After the similar structures were removed from the list there were about 4000 left.

For each NMR structure 10 conformers were chosen. The files were chosen based in the filename of the output filename from the tessellation, and the number of conformers. If 10 – 19 conformers were found, then the first 10 were chosen. For 20-29, the first, third, fifth... nineteenth were chosen. The first 10 files were not chosen, in case the scientist depositing the results had ordered the samples by some criteria (for example least energy). When samples are chosen by almost any criteria, then the samples would have a tendency to be more consistent.

The last dataset was a group of high resolution X-ray structures. A search was done for all X-ray structures with better than 1.5 Angstorms resolution and less than 30% identity from PDB. All the structures which would tessellate were part of the dataset.

Determining Simplex Type

Five types of simplexes have been defined based on residue number in the protein. In this document they will be referred to as T4000 for 4 consecutive residues (for example 4,5,6,7), T3100 for 3 consecutive residues plus one non-consecutive residue (for example 4,5,6,30 or 4,30,31,32), T2200 for 2 groups of consecutive residues (4,5, 30,31),

T2110 for 2 consecutive residues an 2 non-consecutive residues (4,5,30,40), and T1111 for non-consecutive residues (4,30,40,50).

Determining Secondary Structure

Secondary Structure used is based on the DSSP classification, which were obtained from the http://www.cmbi.ru.nl/dssp.html website. Secondary structure was not calculated for each conformer. The values used were obtained from a single query for each PDB ID.

The fact that 2 residues in a simplex have the same secondary structure type does not mean that the residues are in the same secondary structure. For example 2 residues in the same simplex could be in different alpha helices. This section will describe the process used to determine if 2 residues are in the same secondary structure. This is not an exact process as only the information in the simplex is used. The process is a little different for each simplex type.

Type T4000 simplexes will almost always be in alpha helix, a turn, or not in a secondary structure. If the first 2 or last 2 residues are both alpha helix, then this is the start or end of a helix and will be classified as a partial-helix. If all residues are in a helix, then this is a full-helix. If 2 residues in a simplex are turns, then the simplex will be classified as a turn. If 2 or more residues are in a beta strand, then is will be a partial beta strand.

Type T3100 simplex type can be part of an alpha helix, part of a beta sheet, or part of a beta bridge. If 2 consecutive residues are alpha helix, then this is a partial-helix. If two or more residues are beta bridge, then the simplex will be a beta bridge. If all

residues are in a beta strand, then this will be a full beta strand. If 2 or more residues are beta strand, then this will be a partial-beta strand. If 2 residues in a simplex are turns, then the simplex will be classified as a turn.

Type T2200 can have any type of secondary structure. If all 4 are alpha helix and they are separated by one residue (for example if the residue numbers were 4, 5, 7, and 8), then this is a full-helix. This is a common shape in alpha helices. The hydrogen bond is between residue i and (i+4). If the residues are separated by more than one residue, then this is a partial-helix. If 2 consecutive residues are alpha helix, then the simplex is a partial-helix. If 2 non-consecutive residues are beta bridge then the simplex will be considered a beta bridge. If 2 or more residues are beta strand, then this will be a partial-beta strand, or full-beta strand. If 2 residues in a simplex are turns, then the simplex will be classified as a turn.

Type T2210 can also have any type of secondary structure. If all 4 are alpha helix and they are I, i+3, i+4, and i+7, then this will be considered a full-helix. If the 2 consecutive residues are alpha helix, then the simplex will be considered a half helix. If 2 non-consecutive residues are part of a beta bridge then the simplex will be considered a beta bridge. If 4 residues are beta strand, then this will be a full-beta strand. If 2 or 3 residues are beta strand, then this will a partial-beta strand. If 2 residues in a simplex are turns, then the simplex will be classified as a turn.

Type T1111 have a secondary structure of full-helix, partial-beta, beta bridge, or turn. If all 4 are alpha helix and they are I, i+4, i+7, and i+11, then this will be considered a full-helix. If 2 residues are part of a beta bridge then the simplex will be considered a

beta bridge. If 2 or more residues are beta strand, then the simplex will be a partial-beta strand. If 2 residues in a simplex are turns, then the simplex will be classified as a turn.

Determining residue and Simplex Exposure

Each simplex tetrahedron has 4 triangular faces. If the face of a tetrahedron is only included in one simplex, then it is on the edge of the protein and the three residues are surface residues. If a residue is not a surface residue, but is in the same simplex as a surface residue, then it is an undersurface residue. If a residue is neither a surface, nor an undersurface residue, then it is a buried residue.

Since NMR structures have multiple conformers, the buried values will be the average of the values from the 10 conformers. A residue could for example be 10% surface, 80% undersurface, and 10% buried. The values for a simplex will be the sum of the 4 residues.

Determining Simplex consistency

For each NMR protein there will be 10 conformers in the dataset. All the simplexes in all the conformers are found and the number of times the same simplex is found in another simplex is determined. A simplex matches if the 4 residues match. It does not matter which conformer a simplex is found in. For example if the same simplex is found in conformer 2 and 4, then they would both have a consistency level of 2. If a conformer is found in all 10 conformers, then the simplex will have a consistency level of 10.

Determining Log-likelihood Potential

Before determining the potentials for each simplex 3 tables of log-likelihood potentials were created. The first just used only the simplexes which were found in all the

conformers (consistency level 10). The second table of potentials used all the simplexes. If a simplex was found in all 10 conformers, then it would count as 10 simplexes. If it was found in only one conformer, then it would be only 1 simplex. The other table was created from the dataset of high resolution X-ray structures.

To create the tables the residue content of each simplex was determined. The order of the residues does not matter so the simplex AAWW is the same as WWAA. The simplexes are summed to get the number of each combination of residues. The number of simplexes for each combination is divided by the random possibility of getting the simplex times the total number of simplexes. (This is determined by the percentage of each residue in the dataset). For example if a simplex had the residues ABCD, and if the percentages of residues was A=10%, B =20%, C =1%, and D=2%, and there were 1,000,000 simplexes, then the random probability would be ((0.1 * 0.2 * 0.01 * 0.02) * 1,000,000) = 4. The number of ABCD simplexes would be divided by 4, and the log taken. This would be the log-likelihood potential for the residue ABCD. This would be done for all the possible combinations of residues.

Once the potentials tables were created, the potential of each simplex was determined based on the table using only the simplexes found in all the conformers.

Finding Common shapes in consistent simplexes

It is believed that simplexes which are arranged in common shapes will be more consistent. To test this theory, common shapes were determined. The set of 100% consistent quadruplets was used to determine the common shapes. The data was clustered in two ways. The first was to use all the data. The second was to sort the simplexes by

type and only cluster data of the same type together. There are 3 steps in determining the common shapes. First adjust the edge lengths and coordinates so that similar quadruplets will match. Second cluster the edge lengths and determine the center of the cluster. The third step is to determine which clusters show valid shapes.

To adjust the coordinates and edge lengths so that similar quadruplets will match the average length of each edge among the 10 conformers must be determined. Once this is done, the vertices are renumbered so the edges will match. The rules for doing this are:

- (Edge 1-2) is the longest edge.
- (Edge 1-3) is greater than (Edge 1-4), (Edge 2-3), and (Edge 2-4)

Once the vertices are renumbered, then the edge lengths are set. The coordinate of the first vertex is set to (0, 0, 0), the shape is rotated around the Z and Y axis so that vertex 2 is positive and on the X axis. The shape is rotated around the X axis so that point 3 will have a positive value for y and a value of 0 for z. Once the shapes are rotated, the data will be clustered based on areas of high density, but to do this the average density must be determined.

The density of points is the number of points divided by the volume. The volume (V) of space a distance (r) from a point can be calculate as V= C * (r ** N). C is a constant and N is the number of dimensions. For example in 3 dimensional space V = C * (r power 3) – The volume of a sphere (In most cases C would be set to 4/3 pi for a sphere). In this case there are 6 dimensions. (6 edge lengths), so the density will be proportional to the number of points divided by the distance from the center to the 6^{th} power. To determine the average overall density the space over which the density is

calculated must be determined. To do this the average length for each of the edges was calculated. This 6 dimensional point is considered the center of space. The distance from the center of space was then calculated for each quadruplet. The distance where 50% of the quadruplets are closer to the center and 50 are further away was the distance used to calculate the average density. The average density would then be:

D = (n/2) / (C * (r ** 6))

D = density

n/2 = half the number of points

C = Constant

r = distance

Once the average density is found, the data can be clustered based on areas of high density. The steps to clustering the data are as follows:

1. Find the number of quadruplets within a specified Euclidean distance from each quadruplet (A distance 0.2 Angstroms). These points are the initial centroid for the initial list of clusters.
2. Sort all clusters by the number of nearby simplexes.
3. Starting with the quadruplet with the most nearby quadruplets, eliminate all the clusters within a configured Euclidean distance (0.5 Angstroms was used).
4. Next the ring densities of each cluster are found. The ring density is the density of points whose distance from the center of the cluster is between two distances. For example the 0.1-0.2 ring density is the number of points

more than 0.1 Angstroms from the center but less than 0.2 Angstroms divided by the volume of the space between 0.1 and 0.2 Angstroms. Ring densities were calculated for distances up to 1.5 Angstroms at increments of 0.1 Angstroms.

5. Some clusters may appear just outside the minimum range of very dense clusters. To eliminate these clusters, the ring density of the 0.1-0.2 cluster of the less dense cluster is compared to the ring density of more dense clusters at the distance between the two (which must be at least (0.5.-0.6). If the density of the less dense cluster is less than 4 times the large ring density, then the cluster is removed.
6. Adjust the center of the cluster. For each cluster find the new center of the cluster using all the points within a specified distance (0.3 Angstroms was used). Repeat until the location of the point changes less than 0.1 Angstroms or up to 10 times. This moves some low density clusters closer to high density clusters so they can be removed.
7. Repeat steps 3, 4, and 5.

Assigning simplexes to shapes

Once the shapes are found, it must be determined which simplexes match each shape. Many simplexes will not match any shape. The procedure for doing this is as follows.

1. Assume simplexes within 1.5 Angstroms of the cluster center are considered members of the cluster.

2. For each 0.1 Angstrom from 0.1 to 1.5 determine the ring density.
3. Assign simplexes which are in multiple clusters proportionally among all the clusters of which the simplex is a member. The proportion is based on the ring density of the clusters at the distance from the center. For example if a quadruplet were 0.3 Angstroms from cluster A and the ring density of the cluster was 90, and the quadruplet was 0.9 Angstroms from cluster B and the ring density of cluster B was 10 at 0.9 Angstroms, then the quadruplet would be 0.9 in ring A and 0.1 in ring B.
4. Once the simplexes are proportionally assigned, the densities of the clusters change. The densities are recalculated. If the ring density of a ring is less than 2 times the average density, then the ring density is set to 0.
5. Repeat steps 3 and 4 until there is little change in density.

Finding Relationships between Consistency and other properties

Once the properties of each simplex were determined, the each property was checked to see how it affected the consistency. The average log-likelihood potential and average number of exposed residues was determined for each consistency level. For each consistency level the number of simplexes with each secondary structure was determined. The number of simplexes at each level which matched a shape was also determined, and the number of simplexes of each type was determined.

Matching X-ray to NMR files

The simplexes in the X-ray structures were compared to the simplexes for the matching NMR structures and the consistency level of the NMR simplex was found. If

the simplex was not found, then the consistency level was 0. The properties of the X-ray files were compared to the consistency level.

Predicting Consistency levels of X-ray files

A decision tree was created to predict if an X-ray simplex would be below or above a specified consistency percentage. The input to the tree was the range log-likelihood potential of the simplex (there were 4 possible values (P < -0.1), (-0.1 >= P > 0.0), (0.0 >= P > 0.1), and (0.1 < P), the log of the ring density of the shape (possible values (D<1), (1 <= D < 2), or (D>=2)), the number of exposed residues (possible values (n=0), (1<=n<=2), (n=3), and the secondary structure (non-ss, full-helix, partial-helix, full-beta, partial-beta, or turn) .

The dataset of pairs was broken into 10 groups. The decision tree was run 10 times with each of the groups acting as the test set and the other 9 acting as the training set.

CHAPTER THREE - RESULTS

This project has 5 parts. They are to see how simplex type, residue exposure, and secondary structure affect the consistency of NMR structures. Create new log-likelihood potentials and see how each affects the consistency of NMR structures. Generate common shapes from the consistent simplexes, look at the properties of the shapes, and see how each effects the consistency of the NMR structures. The final step is to see how the 5 properties (simplex type, residue exposure, secondary structure, log-likelihood potential, and shape) affect the consistency of the NMR structures which match the X-ray structures.

Consistency of Simplexes

The consistency score of a simplex is between 1 and 10 and is the number of conformers in which the simplex is found (out of the 10 conformers for each structure).

The distribution of consistency scores creates a J shaped curve. (See Figure 1 - Consistency Distribution for all simplexes) The largest number of simplexes has a score of 10 with 34% of the simplexes having this value. The second largest group has a score of 1 at 18%. The smallest groups have a score of 5 and 6, with the groups consisting of only 5.7% and 5.8% of the total simplexes.

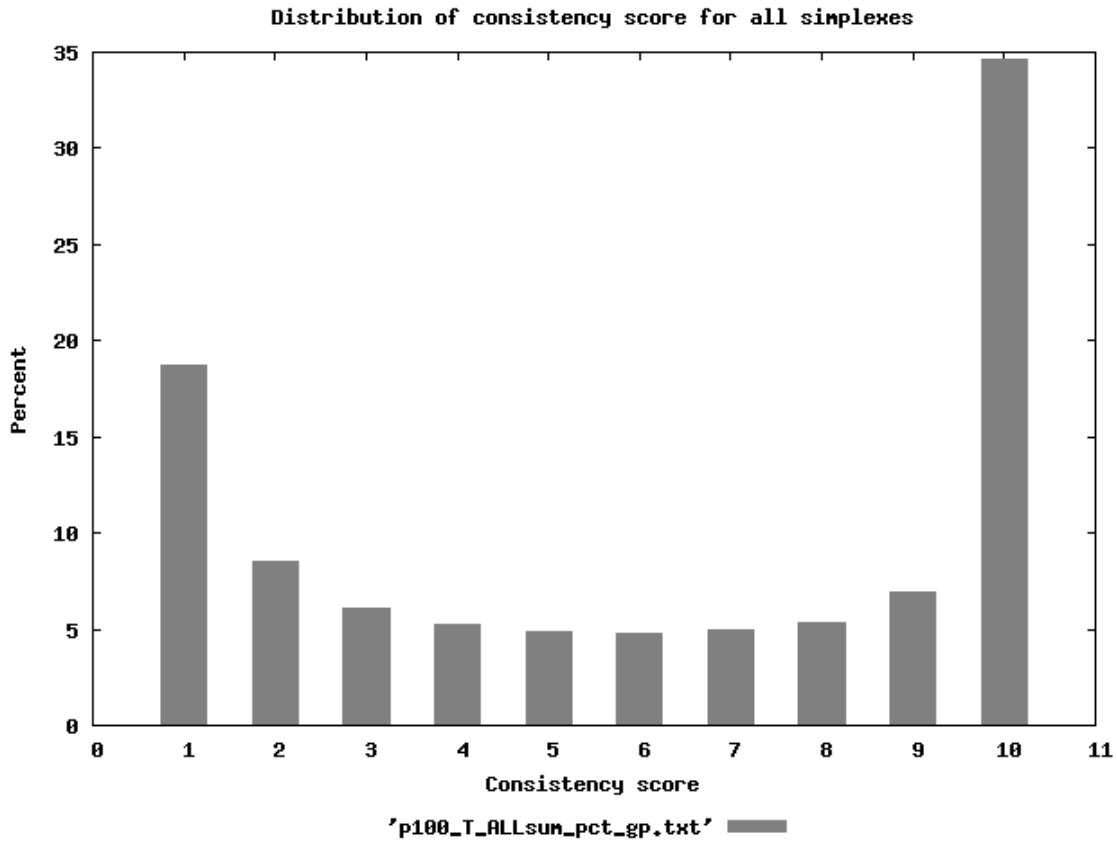

Figure 1 - Consistency Distribution for all simplexes

By looking at the consistency levels broken down by some of the simplex properties, it is possible to better understand the shape of the distribution.

Effect of simplex type and Secondary Structure on Consistency

Secondary structure has a lot to do with the consistency of a simplex. Simplexes classified as full alpha helix (all 4 residues in the same helix) or full beta strands are the most stable. Simplexes with no residues in the same secondary structure are the least consistent. Table 1 shows the makeup of the simplexes by secondary structure type. The table shows that 75% of simplexes with a consistency level of 1 do not have any residues in the same secondary structure. Full helixes and full strands make up a very low

percentage the simplexes at consistency level 1, and a high percentage at level 10. The fact that secondary structures are very consistent is expected. Although at the percentages of simplexes with no secondary structure goes down consistently as the consistency increases, there are still many simplexes which have a consistency level of 10. These must be held together by forces other than hydrogen bonds (Hydrogen bonds are the only criteria DSSP uses for assigning secondary structure).

cons	no_ss	full helix	partial helix	full beta	partial beta	bridge	turn
1	75.47	0.063	7.27	0.516	8.85	0.06	7.77
2	66.16	0.19	10.55	1.19	13.27	0.12	8.51
3	60.98	0.35	12.17	1.81	15.86	0.16	8.67
4	58.018	0.53	13.08	2.43	17.31	0.21	8.43
5	55.40	0.79	13.66	2.89	18.51	0.25	8.49
6	52.89	1.04	14.62	3.62	19.45	0.293	8.09
7	50.35	1.35	15.41	4.21	20.37	0.31	8.00
8	47.31	1.83	16.28	5.15	21.19	0.33	7.92
9	43.148	2.86	17.49	6.54	21.73	0.41	7.83
10	28.64	21.24	18.24	8.74	16.26	0.49	6.39

Table 1- Percentage of each secondary structure for each consistency level

The consistency is also related to simplex type. Type T4000 simplexes have high consistency. The highest percentage of T4000 simplexes has a consistency score of 10 and the number of simplexes keeps decreasing as the consistency decreases. (See Figure 2- Distribution of Consistency Scores for T4000 Simplexes).

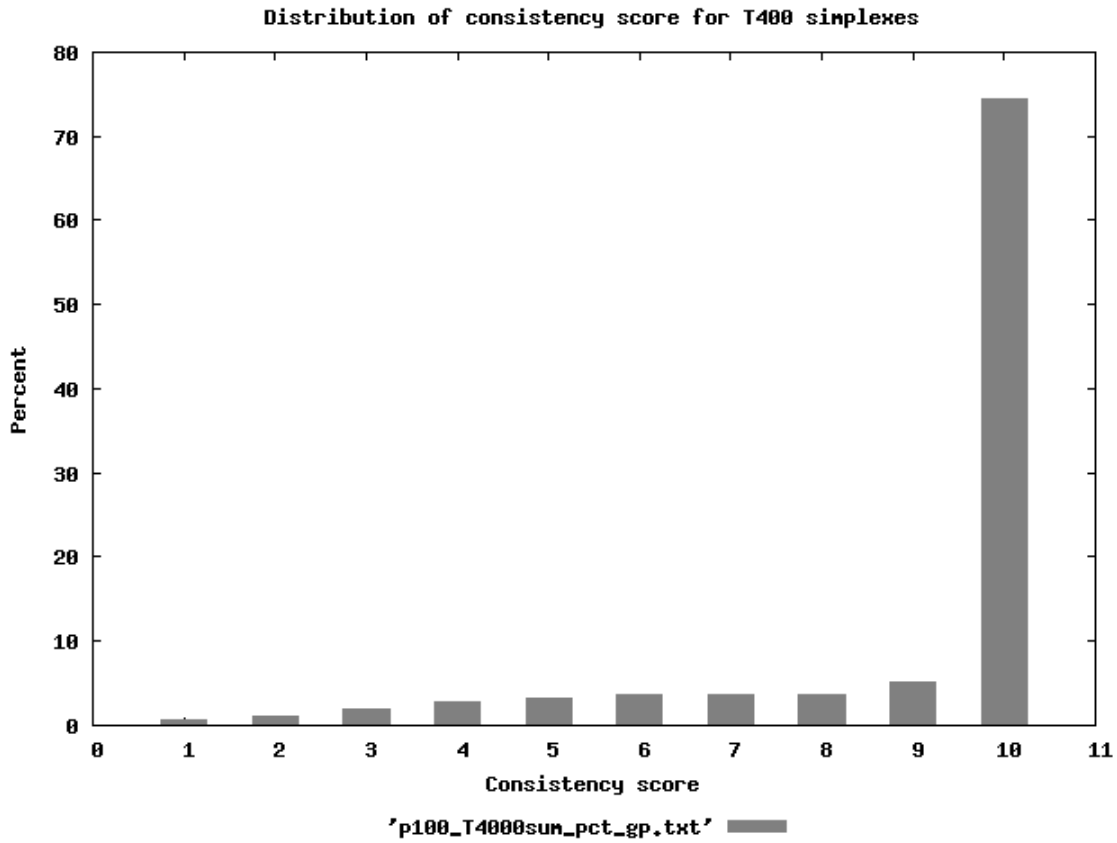

Figure 2- Distribution of Consistency Scores for T4000 Simplexes

One reason T4000 simplexes have such high consistency is that this type has a large number are full alpha helixes, but even simplexes without secondary structure are more consistent than with other simplex types. Table 2 – Percentage of T4000 with no Secondary structure at each consistency level. Table 2 shows over 40% of T4000 simplexes with no secondary structure have a consistency level of 10. Table 3 gives the secondary structure percentages for each consistency level for T4000 simplexes.

Consistency	Pct. simplexes
1	1.44
2	2.93
3	4.99
4	7.32
5	8.74
6	9.31
7	8.61
8	7.63
9	7.98
10	41.04

Table 2 – Percentage of T4000 with no Secondary structure at each consistency level

Cons	no_ss	full helix	partial helix	full beta	partial beta	bridge	turn
1	68.36	1.17	4.13	0	11.73	0	14.61
2	74.60	0.98	3.89	0	7.40	0	13.12
3	77.57	1.03	3.32	0	5.30	0	12.78
4	78.53	1.08	3.69	0	4.14	0	12.56
5	77.81	1.75	4.18	0	3.72	0	12.54
6	75.52	2.51	4.81	0	3.90	0	13.25
7	69.83	3.51	6.00	0	4.26	0	16.40
8	59.70	6.18	9.37	0	4.98	0	19.76
9	45.68	13.21	12.35	0	5.40	0	23.35
10	16.06	54.32	15.49	0	1.26	0	12.87

Table 3 - Percentage of each secondary structure for each consistency level for T4000

The distribution of T3100 simplexes (see Figure 3) is similar to the distribution for all simplexes, many simplexes at the highest and lowest consistencies. Table 4 shows that many full and partial beta sheets account for many of the high consistency simplexes, but there are still a lot of highly consistent simplexes with no secondary structure.

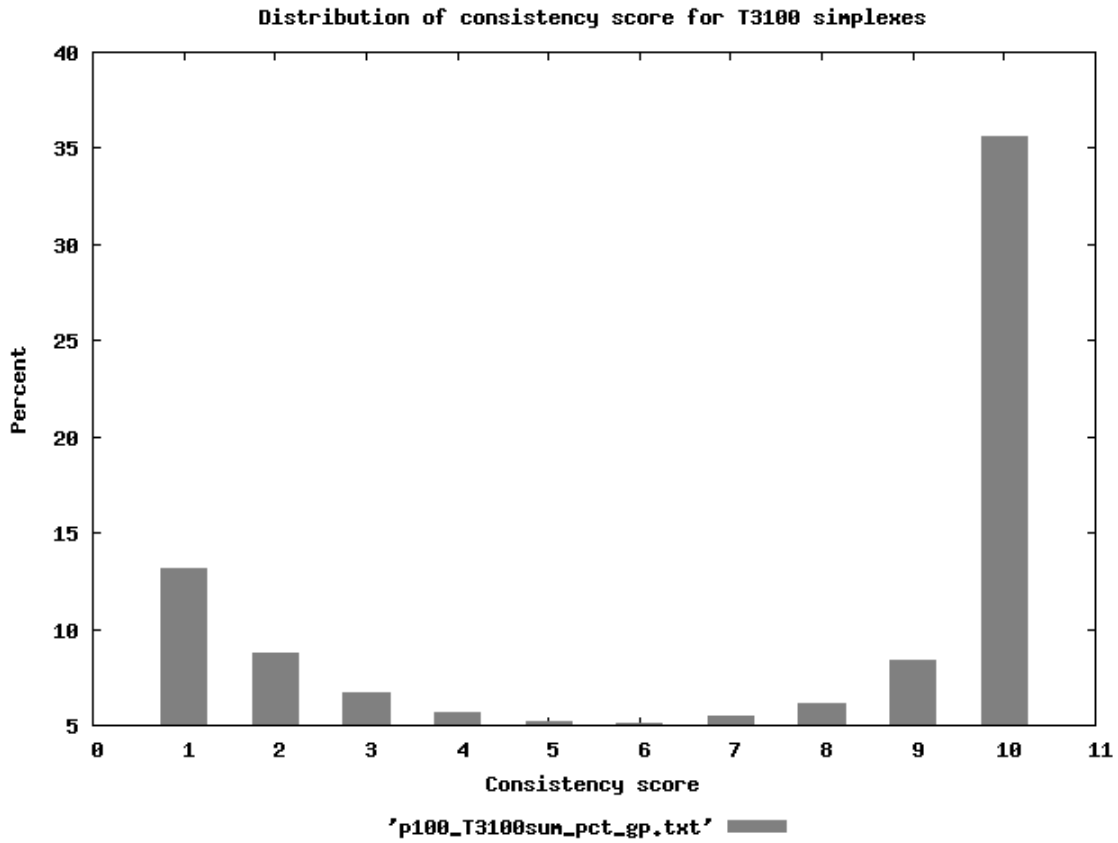

Figure 3 - Distribution of consistency scores for T3100 simplexes

Cons	no_ss	full helix	partial helix	full beta	partial beta	bridge	turn
1	77.20	0	3.81	1.78	8.14	0.11	8.96
2	72.63	0	4.00	2.99	10.83	0.17	9.38
3	67.51	0	4.28	4.47	13.62	0.24	9.89
4	62.61	0	4.46	6.33	16.44	0.32	9.83
5	59.26	0	4.43	7.56	18.29	0.46	10.00
6	55.20	0	4.23	9.81	20.83	0.51	9.42
7	52.44	0	4.16	11.15	22.24	0.53	9.47
8	49.78	0	3.49	13.31	24.38	0.62	8.41
9	45.74	0	3.67	16.43	24.98	0.64	8.54
10	39.49	0	2.64	23.72	26.45	0.83	6.86

Table 4 - Percentage of each secondary structure for each consistency level for T3100

Figure 4 shows that type T2200 simplexes are very consistent. This is due to the large number of simplexes involved in secondary structure, both helixes and strands. The percentage of each secondary level for each consistency level is shown in Table 5.

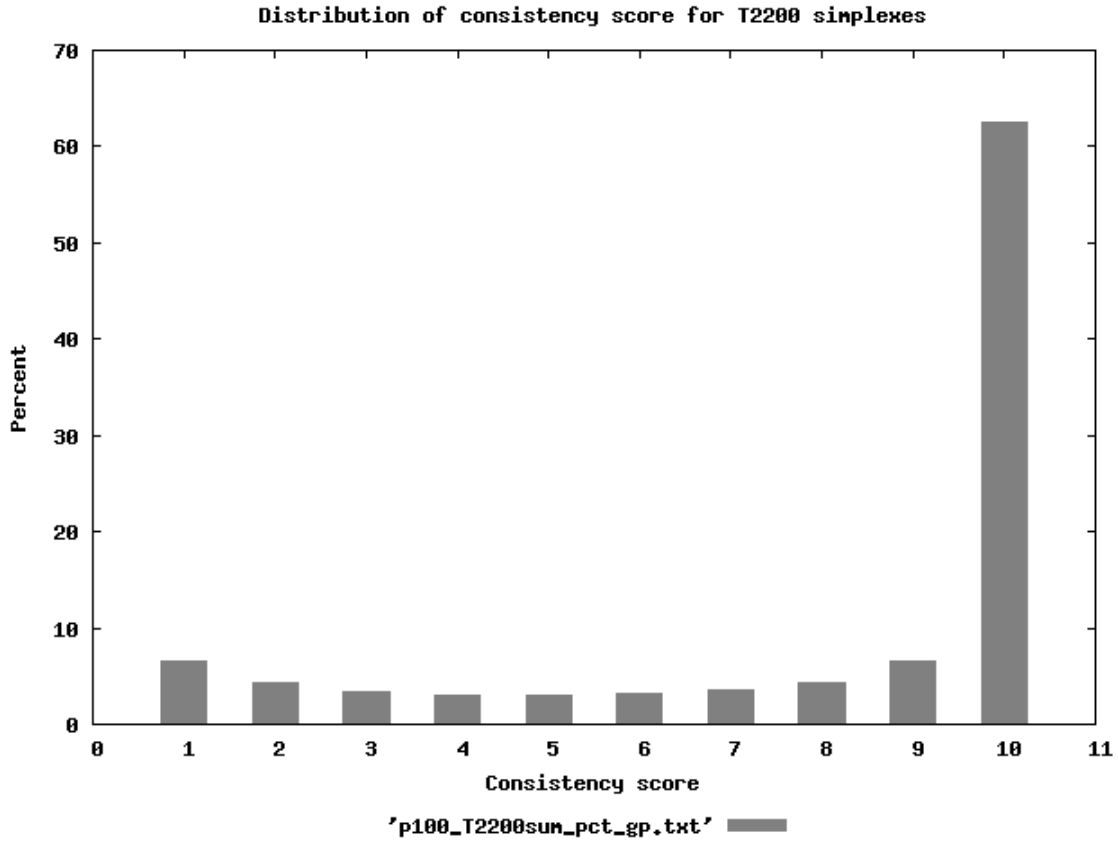

Figure 4 -Distribution of consistency scores for T2200 simplexes

Cons	no_ss	full helix	partial helix	full beta	partial beta	bridge	turn
1	62.97	0.06	16.07	1.39	8.04	0.10	11.37
2	57.63	0.14	17.37	2.99	9.84	0.27	11.76
3	50.94	0.29	19.15	5.12	11.94	0.34	12.21
4	45.84	0.51	20.45	7.36	13.69	0.50	11.66
5	41.92	0.96	20.67	8.87	15.80	0.62	11.16
6	37.43	1.39	22.84	11.38	16.53	0.77	9.66
7	34.80	2.09	21.89	12.88	18.34	0.73	9.27
8	31.25	3.21	21.26	15.64	18.81	0.97	8.85
9	27.13	5.36	21.96	18.35	18.43	1.19	7.58
10	14.64	34.44	17.28	16.18	13.12	1.0	3.33

Table 5- Percentage of each secondary structure for each consistency level for T2200

T2110 simplexes have about equal numbers of simplexes at the minimum consistency and the maximum (see Table 5); with much lower numbers in the middle. Table 6 shows that few type T2110 simplexes have all 4 residues in the same secondary structure, though there are many partial helixes.

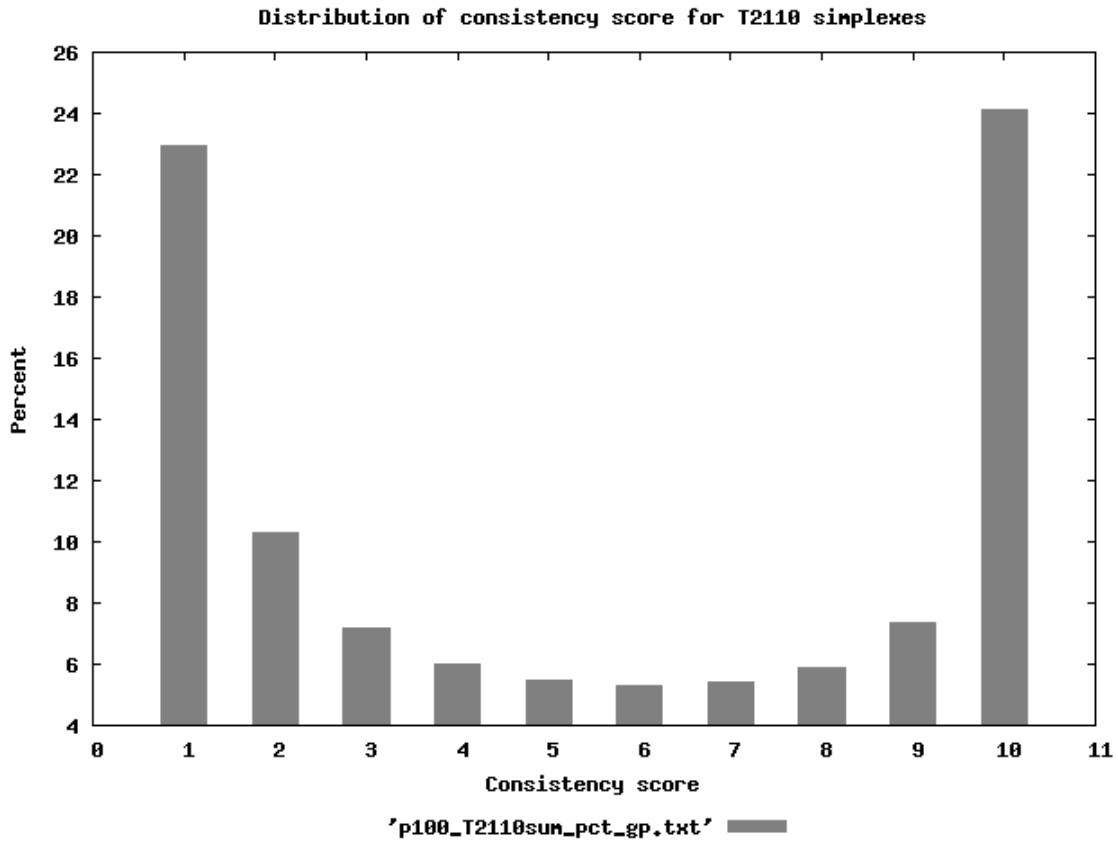

Figure 5 - -Distribution of consistency scores for T2210 simplexes

Cons	no_ss	full helix	partial helix	full beta	partial beta	bridge	turn
1	69.21	0.08	13.53	0.59	7.47	0.10	9.02
2	58.28	0.23	19.58	1.18	10.91	0.17	9.64
3	52.12	0.43	23.10	1.59	12.89	0.22	9.66
4	48.62	0.70	25.19	1.94	14.05	0.28	9.22
5	45.59	1.01	26.61	2.21	15.06	0.30	9.24
6	43.63	1.34	28.28	2.42	15.32	0.33	8.67
7	41.85	1.71	30.04	2.58	15.36	0.36	8.10
8	39.61	2.08	31.59	2.68	15.78	0.29	7.97
9	37.12	2.66	33.75	2.81	15.86	0.32	7.47
10	32.67	3.62	39.76	3.00	14.73	0.35	5.87

Table 6 - Percentage of each secondary structure for each consistency level for T2110

Figure 6 shows that T1111 helixes have the smallest percentage of consistent simplexes. Table 7 shows that this type of simplex has few simplexes in the same secondary structure. Table 8 tells the percentage of simplexes with no secondary structure at each consistency level. Even with no secondary structure, the number of simplexes doubles between consistency level 9 and level 10.

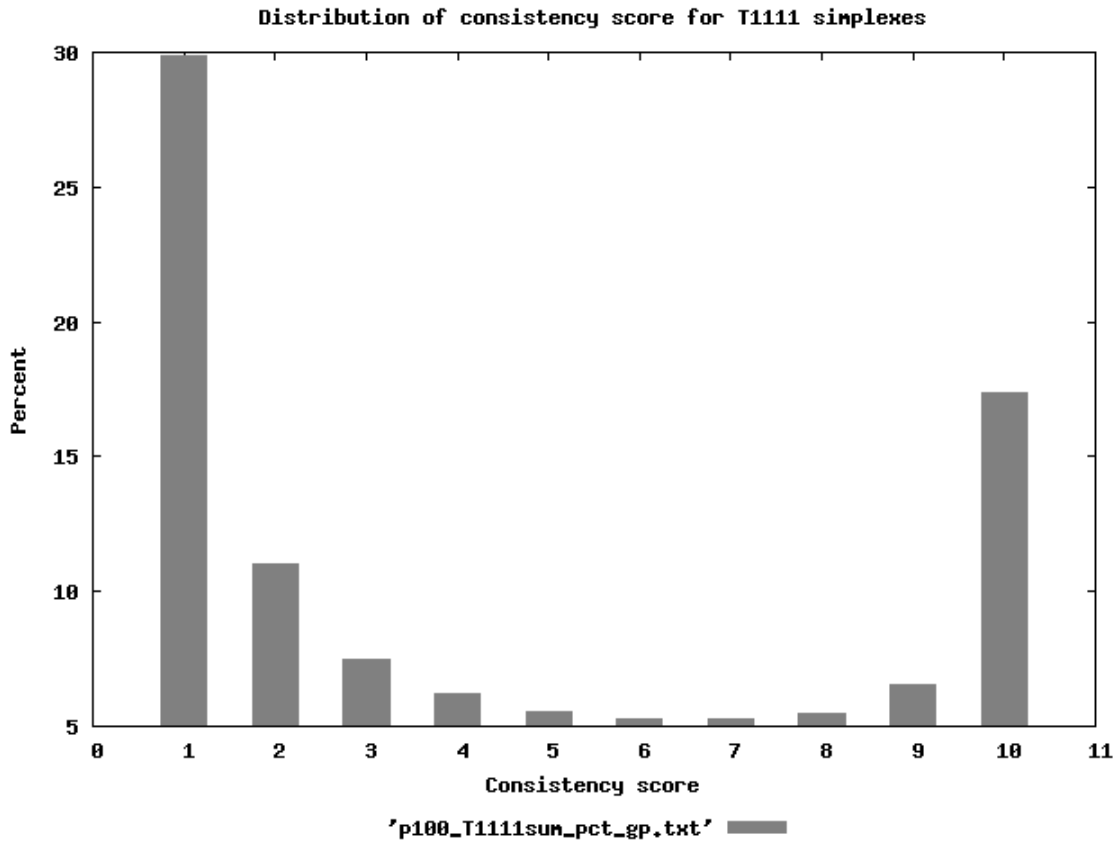

Figure 6- -Distribution of consistency scores for T1111 simplexes

Cons	no_ss	full helix	partial helix	full beta	partial beta	bridge	turn
1	83.76	0.06	0	0	10.65	0	5.53
2	75.52	0.21	0	0	18.63	0	5.65
3	71.11	0.36	0	0	23.20	0	5.33
4	69.03	0.48	0	0	25.53	0	4.96
5	66.80	0.64	0	0	27.67	0	4.89
6	64.97	0.70	0	0	29.70	0	4.63
7	63.24	0.79	0	0	31.67	0	4.30
8	62.23	0.87	0	0	32.80	0	4.10
9	60.00	0.83	0	0	35.58	0	3.59
10	58.18	0.70	0	0	37.84	0	3.28

Table 7 -- Percentage of each secondary structure for each consistency level for T1111

Consistency	Pct. Simplexes
1	35.30
2	11.74
3	7.52
4	6.04
5	5.25
6	4.82
7	4.69
8	4.80
9	5.56
10	14.29

Table 8 - Percentage of T1111 with no Secondary structure at each consistency level

Secondary structure contributes a great deal to a simplex's consistency. Simplexes with all 4 residues in the same secondary structure are almost always stable. They make up over 30% of the simplexes which were found in all 10 of the conformers, but less than 1% of the simplexes were the simplex was found in only one of the conformers.

Although secondary structure is very important, many simplexes are consistent with no secondary structure. Over 40% of T4000 simplexes and 14 % of T1111 simplexes with no secondary structure have a consistency level of 10. If a simplex has all its residues in the same secondary structure, then it is very likely that the simplex will be consistent, but if all the residues are not in the same structure, then it may or may not be consistent.

Consistency and Residue Exposure

As the level of residue exposure increases there is a constant decrease in consistency. The residues which are buried would probably have less freedom of movement since they are surrounded by other residues. Also exposed residues in the solution may interact with water. Table 9 shows that the percentage of exposed residues, undersurface residues (residues in same simplex as an exposed residues), and buried residues (residues not exposed or in same simplex as an exposed residue). As the number of exposed residues increases the consistency level decreases. As the number of buried residues increases the consistency level increases.

Consistency	Pct. exposed	Pct. undersurface	Pct. buried
1	38.88	52.88	8.24
2	34.92	54.76	10.31
3	31.88	56.16	11.96
4	29.79	57.24	12.97
5	28.38	57.85	13.78
6	26.98	58.39	14.62
7	25.67	58.89	15.43
8	24.51	59.33	16.15
9	23.10	59.99	16.91

10	21.61	60.83	17.56

Table 9 - Percent of residues exposed at each consistency level

Consistency and Log-likelihood Potential

The log-likelihood potential is a measure of the propensity of residues to exist in the same simplex. The reasons why the residues come together are not considered.

As the log-likelihood potential increases, the consistency increases. The relationship between the potential and the consistency was looked at using 3 different log-likelihood potential tables. One table was generated the using high resolution (1.5Angstorms or less) X-rays. Another used all the simplexes from the NMR dataset. The third used only the 100% consistent simplexes. In all three groups the potential tended to increase as the consistency increased.

The average potentials were calculated at each consistency level for each of the 3 potential tables. With all the tables, the log-likelihood increased as the consistency increased.

The potentials using all the NMR simplexes showed the least difference in potential as the consistency increased. Table 11 shows the average potential at each consistency level when using the all NMR potentials table. The potentials using the X-ray simplexes showed larger differences in potential as the consistency increased (see Table 12). The 100 % consistent potentials list created the greatest difference in average potentials. Table 10 shows the average potentials for this group. The average potential for the 100% consistent list at a consistency level of 1 was negative. This means that the likelihood of simplexes occurring is less than random.

It should be noted that the dataset used to generate the NMR potentials is the same as the dataset which was checked, but since with the 100% consistent potentials set does not use the non-100% simplexes, the effect of using the same dataset would be small except at 100% consistency level. A few residue combinations were not found at all, so a minimum value of -4 was used for these very rare combinations.

Consistency	Log likelihood
1	-0.113
2	-0.070
3	-0.041
4	-0.022
5	-0.007
6	0.006
7	0.017
8	0.030
9	0.043
10	0.076

Table 10 - Average Log-likelihood Potential at each consistency level using 100% consistent simplexes

.

Consistency	Log likelihood
1	0.023
2	0.029
3	0.032
4	0.037
5	0.042
6	0.045
7	0.047
8	0.048
9	0.050
10	0.046

Table 11- Average Log-likelihood Potential at each consistency level using all NMR simplexes

Consistency	Log likelihood
1	0.008
2	0.016
3	0.021
4	0.026
5	0.030
6	0.032
7	0.034
8	0.036
9	0.037
10	0.029

Table 12 - Average Log-likelihood Potential at each consistency level using X-ray simplexes

LL Table	Avg. potential	Std. Dev
NMR 100	-0.0222	0.371
NMR All	0.0218	0.221
X-ray	0.0143	0.273

Table 13- Average value and Standard Deviation of values in Log-likelihood Potentials tables

Table 13 shows the average potential and standard deviation of the potentials in each of the tables. The standard deviation of the NMR 100 table is much greater than the other tables. This is because it is using a subset of simplexes which are very consistent. Unlikely simplexes which tend to be inconsistent will have even lower values and simplexes which high potential will be even higher.

The average difference in potential when using the 100% table is about 0.2 which is greater than ½ of the standard deviation. For the all NMR and the X-ray potentials, the difference is only about 0.04 or less than 20% of the standard deviation. From this it appears that the NMR 100 table would be a better predictor of consistency.

Since it is logical to believe that the consistency is related to the protein stability, the fact that the 100% NMR potentials dataset shows larger differences than the other datasets may indicate that the 100% NMR potentials dataset may work better for applications using nearest neighbor potentials which try to determine protein stability.

Consistency and simplex shape

Common shapes were found among the 100% consistent simplexes. When a simplex matches a shape it is very likely that the simplex will be very consistent. Table 14 shows the distribution of simplexes which match a shape. The table says that 76% of simplexes which match a shape have a consistency level of 10. Table 15 gives the percentage of simplexes at each consistency level which match a shape. From the 2 tables it is clear that matching a shape is a good indicator that a simplex is consistent, but since many simplexes do not match a shape, the fact that a simplex does not match a shape does not mean that the simplex is not consistent.

Consistency	Pct of shapes
1	1.54
2	1.58
3	1.71
4	1.85
5	2.04
6	2.35
7	2.91
8	3.74
9	6.14
10	76.15

Table 14 - Distribution of simplexes which match a shape

Consistency	Pct of simplexes
1	1.39
2	3.12
3	4.70
4	5.93
5	7.04
6	8.29
7	9.93
8	11.71
9	14.80
10	37.02

Table 15 - Percent of simplexes at each consistency level which match a shape

Properties of Common Simplex Shapes

This section will talk about the properties of the 34 shapes found. About half of the 100% consistent simplexes match one of the shapes. This section will not describe each shape.

For a shape to be considered a common shape there must have a high number of simplexes with very similar shapes. As the Euclidian distance from the shape increases (based on the edge size), the density of points must decline. This is seen in Figure 7 which is the log of the ring density of a typical shape at increasing distances. Since this is for a type T3100 simplex, it might be assumed that 2 of the edges (between consecutive residues) are constant. This would decrease the dimensions of the problem from 6 to 4. Since the volume (inverse of density) is a constant time distance to the 6^{th} power, the differences in density would be much smaller, but the density would still be declining as distance increased. The pattern of declining density is the same for all shapes.

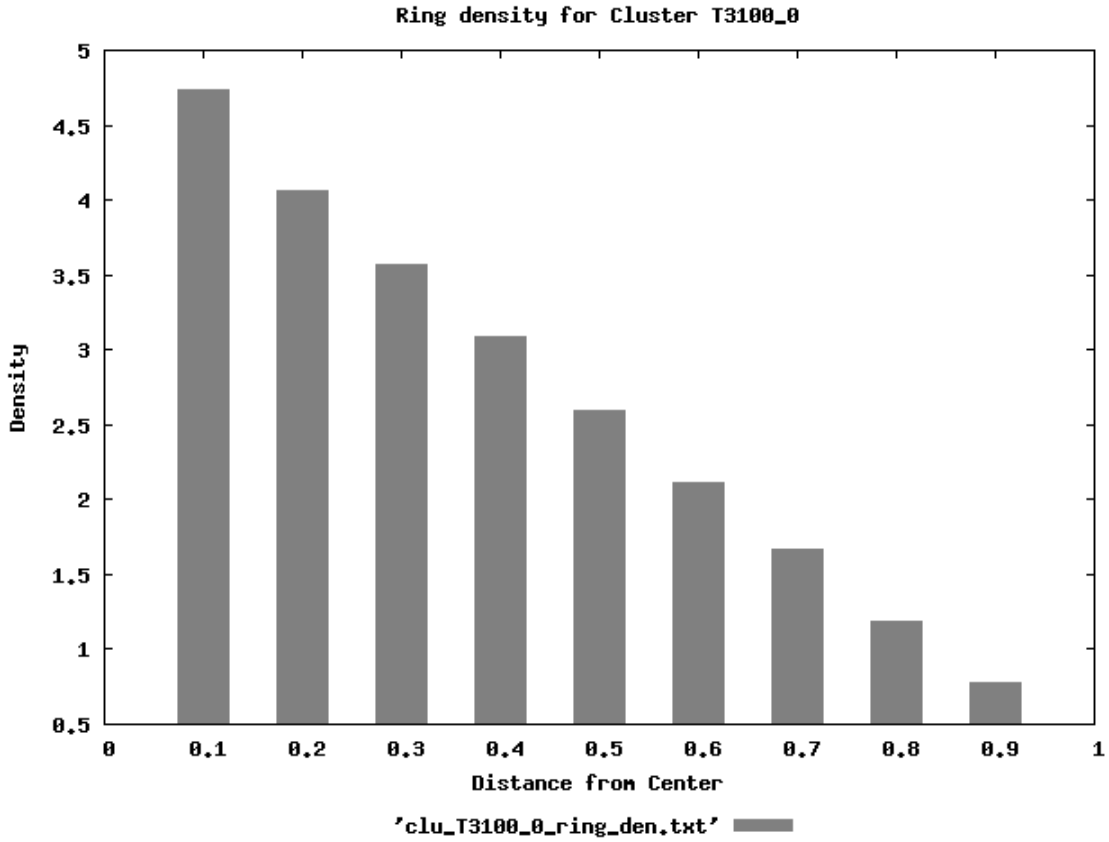

Figure 7- Log of ring density of Shape 3110_0

For each shape a number of properties were determined. They are percent of simplexes in each secondary structure, percent of residues exposed or buried, and mirror image percentage. Since the shapes were determined based on the edge lengths, mirror images would have the same shape. To get the mirror image percentage, all the simplexes are checked for which of the 2 possible shapes they have and the numbers of each are determined. The larger percentage is the mirror image percentage. Each shape will generally have a large number of either alpha helix or beta strand. Few have many of both. A few have shapes are almost completely full alpha helixes, but most have

combinations of full structure, partial structures, and no structures. Descriptions of all the shapes are in the appendix.

Consistency of X-ray Structures

The X-ray structures of proteins were compared to the NMR structure of a number of proteins, and the consistency of each X-ray simplex was determined based on the consistency of the matching NMR protein simplex. Many simplexes did not match, so they had a consistency level of 0.

The results were similar to the results seen processing the NMR data. Type T1111 simplexes had little consistency (43% of the X-ray simplexes did not match any NMR simplex), and Type T4000 had high consistency (74% matched all the simplexes). Simplexes which match one of the shapes were more likely to be consistent. Simplexes which were more exposed were less consistent. Simplexes with secondary structure were more consistent. Simplexes with high log-likelihood potential were likely to be more consistent. Table 16 shows a constant increase in log-likelihood potential and a constant decrease in average residues exposed as the consistency increases. It also shows 66 % of simplexes which match shape are at consistency level of 10.

Consist	Avg LL-pot	Avg exposed	Shape dist pct	Shape cnt	tot
0	-0.011	1.27	9.48	3346	55046
1	-0.010	1.23	2.58	911	10783
2	0.004	1.14	2.27	801	7600
3	0.017	1.13	2.02	715	6029
4	0.020	1.08	1.96	693	5378

5	0.025	1.09	2.02	714	5150
6	0.028	1.07	2.31	814	5158
7	0.035	1.01	2.58	912	5136
8	0.051	0.98	3.25	1146	6132
9	0.049	0.96	5.14	1816	8377
10	0.057	0.87	66.39	23445	56786
total			100	35313	171575

Table 16 - Simplex properties at each consistency level

Table 17 shows the distribution of simplexes with each secondary structure. It for example shows that 85 % of full-helix simplexes had a consistency level of 10, and only 3% were not found. Simplexes without secondary structure were much more likely to not be found in the NMR structure, but still a large number of simplexes with full secondary structure (especially full-beta) were not found. This could indicate a change in secondary structure.

Con	None	full_helix	part_helix	full_beta	part_beta	turn	Num
0	40.79	3.44	27.23	16.67	36.30	37.24	55046
1	7.40	0.93	5.45	4.76	7.14	7.48	10783
2	4.78	0.91	4.07	3.38	5.37	5.06	7600
3	3.97	0.84	3.32	2.99	3.90	4.01	6029
4	3.44	0.68	2.94	2.81	3.58	3.57	5378
5	3.13	0.83	3.23	2.96	3.36	3.10	5150
6	3.20	1.13	3.16	3.26	3.22	2.87	5158
7	2.93	1.17	2.97	3.67	3.40	3.02	5136
8	3.41	1.57	3.93	4.72	3.87	3.38	6132
9	4.37	2.90	5.78	6.48	5.21	4.49	8377
10	22.57	85.59	37.92	48.30	24.67	25.78	56786
total	100	100	100	100	100	100	171575

Table 17 - Percent of simplexes with secondary structure at each consistency level

Table 18 gives the consistency by simplex type. Almost 50% of the T_1111 X-ray simplexes do not exist in the NMR structure.

Consist	T_1111	T2110	T2200	T3100	T4000	total
0	47.21	39.83	13.60	25.78	4.40	55046
10	7.51	7.60	3.97	6.46	1.64	10783
20	5.10	5.09	3.05	4.87	1.72	7600
30	3.84	3.97	2.54	3.99	1.78	6029
40	3.49	3.44	2.33	3.44	1.86	5378
50	3.22	3.18	2.37	3.53	1.93	5150
60	3.00	3.23	2.50	3.44	2.34	5158
70	2.90	3.08	2.61	3.80	2.20	5136
80	3.39	3.66	3.42	4.27	2.84	6132
90	4.10	4.79	5.25	6.16	4.53	8377
100	16.25	22.14	58.35	34.26	74.76	56786
total	100	100	100	100	100	171575

Table 18- Consistency by type in X-ray NMR pairs

Predicting consistency from X-ray structure

A decision tree was created to predict if an X-ray simplex would be stable (> 50% consistency) or unstable (<= 50% consistency).

Consistent correct TP (True positive) 45149

Consistent incorrect FN (False negative) 36395

Inconsistent correct TN (True negative) 76926

Inconsistent incorrect FP (False positive) 16368

Q VAL=0.70

BER VAL=0.31

Although there are a high number of errors, this does show that it is possible to predict the consistency. To better predict the consistency, other methods might work better. A decision tree can only handle a fixed number of values for each parameter. For the log likelihood potential all the possible values were mapped to 0, 1, 2, or 3. Other methods, for example random forest, could handle the continuum of possible values.

CHAPTER FOUR - CONCLUSION

The distribution of simplexes at different levels was determined as well as some of the factors which determine the consistency. Common shapes among the consistent simplexes were found and some of the properties of the shapes. X-ray structure were matched with NMR structures and the properties of the X-ray simplexes which matched each consistency level on the NMR structures were determined as well as properties of simplexes with did not match any NMR simplex. It was shown that it is possible to predict (with some accuracy) the NMR consistency of X-ray simplexes.

It was found that simplexes tend to have either high consistency or low consistency, with very few in the middle. The largest number of simplexes in the NMR data was at 100% consistency (all conformers match) followed by 10% (no conformers match). The lowest values were in the middle. The most consistent simplexes were simplexes with 4 residues in the same alpha-helix. Since simplexes which would be expected to be consistent were, it is reasonable to believe that the differences in NMR and X-ray structures are real differences based on the environment, not products of inaccuracies. It is possible that biases in the methods of computing the structures and choosing the structures would make secondary structures more constant while not effecting random coils. For example when choosing the conformers to submit to PDB,

the criteria of least energy would tend to favor the conformers which show secondary structure, so the secondary structures would be more consistent.

The consistency of residues depends on a number of factors. The factors are number of exposed residues, secondary structure, simplex type, log-likelihood potential, and shape. None of these factors alone can predict perfectly the consistency of a residue. The best predictor is having 4 residues in the same alpha-helix, but still 3% of full alpha-helix simples and 17% of full beta-strands in X-ray structures do not have the same simplex in NMR structures. If the full alpha-helix simplex does not exist, then it is likely that a portion of the alpha-helix does not exist (The helix would be shorter). The missing helix would be an important inaccuracy when trying to understand the relationship between the structure and function. Some beta strands also seem to be missing. The ability to predict changed secondary structure could be very important.

Log-likelihood potentials were generated in a different manor which gave somewhat different results. The log-likelihood potentials generated from only the 100% consistent simplexes seemed to change more as the consistency changed. It would be interesting to see how the results from a programs using the log-likelihood potential (for example auto-mute) would change if the NMR potentials were used.

A number of common shapes were found looking at the 100% consistent data. About half the 100% consistent simplexes were part of one of these shapes. Having these shapes could become very useful. By determining the likely residues at each position of each shape, it might be possible to better predict the effect of a change in residue.

The factors which cause NMR shapes to have low consistency are the same as the factors which cause X-ray structures to not have matching NMR structures. X-ray simplexes which have low log-likelihood potential, high residue exposure, no secondary structure, and do not match a shape are likely to not match an NMR structure. Also type T1111 simplexes are less likely to match. A prediction algorithm was created which used the 5 prediction factors. It had modest success, but proved it was possible and could be improved.

Possible future work

1. Use the new log-likelihood potentials in current software and compare the results.
2. Examine closely the factors which lead to changes in secondary structure between X-ray and NMR structures and hopefully find methods to predict them.
3. Examine more closely the properties of the simplex shapes. Residue content of different shapes may be different, so in different shapes the effect of a residue change could be different even with the same combination of residues.

APPENDIX – SHAPE DESCRIPTIONS

Three tables have information about each shape found. Table 19 - Edge Lengths of shapes gives the edge lengths of each shape. Table 20 - Secondary Structure of Shapes gives the percentage of simplexes with each secondary structure. Table 21- Properties of Shapes tells the percentage of surface, undersurface, and buried residues, as well as the mirror image percentage, and total number of simplexes for each shape.

Clust	Len 1	Len 2	Len 3	Len 4	Len 5	Len 6
T2110_0	10.48	6.229	5.116	5.089	6.041	3.809
T3100_0	6.979	6.674	3.802	4.432	5.708	3.803
T2110_4	7.153	6.066	5.838	3.802	5.673	4.93
T3100_21	6.198	5.771	3.804	5.468	3.803	4.892
T3100_18	5.741	5.561	3.808	5.281	5.35	3.808
T1111_0	16.14	10.45	6.115	6.004	10.26	5.044
T3100_16	7.057	6.414	5.269	3.8	6.246	3.799
T3100_17	6.605	5.753	3.801	4.705	3.803	5.327
T2110_3	7.303	6.857	5.731	4.295	3.805	5.753
T3100_3	6.718	6.509	3.801	4.425	3.802	5.558
T2200_0	6.148	5.229	3.81	3.81	5.01	5.457
T3100_12	5.941	5.354	5.03	3.81	3.81	5.653
T3100_20	6.51	5.842	3.803	5.456	3.802	4.478
T4000_0	5.523	5.138	3.808	3.808	3.808	5.38
T2110_5	6.302	5.557	5.24	5.276	3.808	5.966
T3100_19	7.544	6.472	5.143	3.802	6.303	3.801
T3100_9	5.669	5.361	3.809	5.096	3.806	5.345
T1111_1	16.55	10.69	6.282	6.161	10.5	5.127
T1111_2	15.59	10.2	5.87	5.82	9.976	5.027

T3100_13	7.781	6.732	3.803	4.508	6.249	3.803
T3100_1	7.414	6.637	3.801	4.78	5.988	3.801
T3100_5	8.042	6.649	3.801	5.074	6.222	3.799
T3100_14	6.508	5.852	3.801	5.192	3.804	5.599
T3100_15	6.83	6.596	3.801	5.397	5.649	3.802
T3100_11	7.755	6.76	3.803	5.132	5.824	3.802
T3100_6	6.783	6.585	3.803	4.882	3.803	5.773
T2110_2	10.05	6.058	4.976	4.912	5.824	3.805
T3100_7	6.567	6.359	3.807	5.153	3.808	6.09
T3100_10	6.571	6.317	3.798	4.758	5.917	3.801
T3100_4	7.51	6.566	3.802	5.201	6.242	3.803
T2110_1	10.82	6.519	5.163	5.086	6.278	3.807
T3100_8	6.458	6.11	3.799	4.276	3.799	5.386
T2200_1	6.687	5.526	3.807	3.807	5.06	5.415
T3100_2	7.022	6.526	3.802	5.259	6.094	3.802

Table 19 - Edge Lengths of shapes

Clust	No ss	Full-hel	Part-hel	Full-beta	Part-beta	Turn
T2110_0	0.388	88.81	10.5	0.032	0.083	0.179
T3100_0	5.895	0	0.021	65.9	27.82	0.349
T2110_4	22.26	0	41.68	7.801	24.18	4.068
T3100_21	28.1	0	0.132	27.58	40.81	3.362
T3100_18	27.4	0	2.765	20.93	35.04	13.84
T1111_0	14	85.6	0	0	0	0.388
T3100_16	27.98	0	0.862	29.03	37.81	4.311
T3100_17	24.83	0	0.264	24.22	48.87	1.799
T2110_3	16.02	0	36.6	15.92	28.76	2.68
T3100_3	6.826	0	0	59.54	33.48	0.141
T2200_0	0.975	75.66	16.14	3.227	2.95	1.043
T3100_12	27.1	0	6.552	9.814	34.44	22.08
T3100_20	27.11	0	0.05	36.01	35.42	1.398
T4000_0	1.884	74.61	14.28	0	1.376	7.841
T2110_5	24.57	0.024	35.86	7.431	28.2	3.892
T3100_19	29.86	0	0.561	28.95	37.51	3.098
T3100_9	34.42	0	3.331	10.31	34.01	17.91
T1111_1	18.17	81.15	0	0	0	0.672
T1111_2	14.48	84.84	0	0	0.675	0

T3100_13	8.45	0	0	60.71	30.5	0.334
T3100_1	7.538	0	0.026	68.58	23.71	0.128
T3100_5	10.12	0	0	63.53	25.89	0.442
T3100_14	26.66	0	0.19	23.16	47.95	2.026
T3100_15	22.66	0	0.068	32.76	43.05	1.452
T3100_11	13.17	0	0	52.49	33.85	0.471
T3100_6	14.43	0	0.003	47.26	37.92	0.378
T2110_2	2.183	77.4	16.25	1.002	1.823	1.328
T3100_7	28.01	0	0.111	28.28	42.64	0.938
T3100_10	13.47	0	0.199	52.46	32.83	1.027
T3100_4	12.12	0	0.004	57.43	29.99	0.433
T2110_1	1.743	79.46	16.99	0.243	0.921	0.627
T3100_8	13.15	0	0.678	42.44	41.89	1.826
T2200_1	2.73	61.08	23.31	5.761	5.282	1.828
T3100_2	13.07	0	0.122	51.52	35	0.276

Table 20 - Secondary Structure of Shapes

Clust	Surf	U-surf	Buried	Mirror	Num
T2110_0	201.5	194.6	3.814	99.83	2236
T3100_0	32.27	236.5	131.1	96.1	6351
T2110_4	43.17	270	86.78	59.68	6348
T3100_21	70.56	245.7	83.66	62.06	674
T3100_18	112.6	236.8	50.53	60.24	1539
T1111_0	213.9	183.5	2.427	99.94	234
T3100_16	66.64	249.9	83.43	80.4	2360
T3100_17	87.99	243	68.93	72.89	1895
T2110_3	38.2	263	98.75	61.64	5946
T3100_3	64.8	242.4	92.77	97.3	1709
T2200_0	85.54	252.4	62.04	98.66	48691
T3100_12	117.9	238.8	43.16	61.69	2891
T3100_20	67	243.2	89.77	59.79	2311
T4000_0	101.1	245.2	53.54	99.03	77983
T2110_5	52.4	266	81.57	59.72	4312
T3100_19	57.75	251.1	91.09	85.32	2379
T3100_9	124.1	232.9	42.85	53.14	1867
T1111_1	245.2	151	3.714	99.14	197
T1111_2	148.3	250.4	1.287	98.95	162

T3100_13	40.81	237.8	121.3	93.93	1566
T3100_1	31.26	218.1	150.6	95.28	2591
T3100_5	44.8	237	118.1	92.39	2303
T3100_14	74.28	253.2	72.48	73.18	3999
T3100_15	63.31	247.6	89	75.18	3429
T3100_11	55.65	243.9	100.3	78.41	1577
T3100_6	71.14	238.9	89.85	92.15	1021
T2110_2	149.8	239	11.09	98.94	1980
T3100_7	67.24	253.9	78.8	90.44	1682
T3100_10	40.45	240.6	118.8	92.8	2453
T3100_4	43.79	231.7	124.4	94.32	2067
T2110_1	232	164.3	3.575	99.63	2475
T3100_8	67.73	243.9	88.31	90.97	2384
T2200_1	100.7	239.3	59.89	98.02	17289
T3100_2	45.68	241.5	112.8	94.94	2831

Table 21- Properties of Shapes

Table 21- Properties of Shapes gives the percentage of residues which are surface, undersurface, or buried. The total of the 3 added up to 400 since there are 4 residues. It also gives the percentage of the simplexes which have the same mirror image (the higher of the 2 numbers). Also it gives the total number of simplexes found with the shape.

REFERENCES

Singh, R.K., Tropsa, A, and Vaisman, I (1996) "Delaunay Tessellation of Proteins: Four Body Nearest-Neighbor Propensities of Amino Acid residues". J.Vomput.Biol.,3, 212-221

Joosten RP, Te Beek TAH, Krieger E, Hekkelman ML, Hooft RWW, Schneider R, Sander C, Vried G, "A series of PDB related databases for everyday needs", NAR 2010; odi: 10.1093/nar/gkq1105.

Kabsch W, Sander C, "Dictionary of protein secondary structure: pattern recongnition of hydrogen-bonded and geometrical features", Biopolymers, 1983 22 2577-2637. PMID:6667333; UI:84128824

Hekkelman ML, Vried G., "MRS: a fast and compact retrieval system for biological data:", Nucleic Acids Research 2005 33(Web Server issue): @766-W769; doi:10.1093/nar/gki422.

Majid Masso and Iosif I. Vaisman, "AUTO-MUTE: web-based tools for prediction stability changes in proteins due to single amino acid replacements". Protein Engineering, Design & Selection vol.23 no.8 pp. 683-687, 2010

Kresimir Sikic, Sanja Tomic and Oliviero Carugo, "Systematic Comparison of Crystal and NMR Protein Structures Deposited in the Protein Data Bank". The Open Biochemistry Journal, 2010, 4, 83-95.

Overington John, Donnelly Don, Johnson Mark S, Sali Andre, and Blundell Tom (1992) "Environment-specific amino acid substitution tables: Tertiary templates and prediction of protein folds", Protein Science (1992), I, 216-226

Topham CM, Srinivasan N, and Blundell TL (1997) "Prediction of protein mutants based on structural environment-dependent amino acid substitution and propensity tables". Protein Eng. 10,7.21

BIOGRAPHY

Steven Bowers graduated from T.C. Williams High School, Alexandria, Virginia, in 1977. He received his Bachelor of Science in Civil Engineering from University of Virginia in 1981.