GSP: A SOFTWARE SYSTEM FOR GLYCOSYLATION SITES PREDICTION

by

LIREN DING

(Under the Direction of Krzysztof J. Kochut)

ABSTRACT

The attachment of the sugar N-acetylglucosamine (GlcNAc) to proteins occurs after the protein is folded into its 3D shape. The GlcNAc is transferred to the surface of the protein by the enzyme O-GlcNAc transferase (OGTase). This process is called glycosylation. Glycosylation is a normal process in mammals, including humans. However, when it becomes hyperactivated, it leads to health problems, including type II diabetes. If we could identify the requirements for glycosylation, it is possible that a drug could be derived that could be used to mediate this interaction and so treat diseases such as diabetes. In this thesis, we propose a method which adapt RMSD algorithm to identify the structural features in the vicinity of serine or threonine residues, which are responsible for directing the enzyme to a particular position in the protein.

INDEX WORDS: Glycosylation site prediction, RMSD

GSP: GLYCOSYLATION SITES PREDICTION

by

LIREN DING

B.S., North China University of Technology, China, 2002

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2008

© 2008

Liren Ding

All Rights Reserved

GSP: GLYCOSYLATION SITES PREDICTION

by

LIREN DING

Major Professor:	Krzysztof J. Kochut
Committee:	Robert J. Woods
	Liming Cai

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2008

ACKNOWLEDGEMENTS

I would like to thank everyone who helped make this thesis possible, especially, Dr. Kochut for his kindness, and his efforts to always make the project better, Dr. Woods for suggesting an expansion of the research to higher dimensions, and Dr. Cai for answering so many of my questions. And last, but far from least, I would like to thank all my friends at the Computer Science department and CCRC for making it fun to spend time with them.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

Computational biology and computational chemistry involve the use of computer science techniques in solving biological and chemistry problems. From late 1980s, with the rapid growth of computer applications, computer techniques began to change the field of biology and chemistry. Bioinformatics would not be possible without advances in computing hardware and software. Fast and high-capacity storage media are essential even to maintain the archives. Information retrieval and analysis require programs, some fairly straightforward and others extremely sophisticated. Distribution of the information requires the facilities of computer networks and the World Wide Web[1]. Major research fields in computational biology and computational chemistry include sequence alignment, gene finding, genome assembly, protein structure alignment, protein structure prediction, prediction of gene expression and protein-protein interactions. Lots of computing tools became widely available to the biology and chemistry community. Homology-modeling software and web-based database search services have been developed to provide many services such as multiple alignment, genetic analysis, and motif identification. Researchers can find dozens of sequence matches in seconds using sequence-alignment programs such as BLAST [2] and FASTA [3]. At the same time, Web based

search services, such as RCSB [4] or NCBI [5], provide a variety of Web services and resources for studying the structures of biological macromolecules and their relationships to sequence, function, and disease.

Developing analytical tools to discover knowledge in data is one of the main goals of computer science research and application development in biology and chemistry. There are many levels at which biological information is used. For example, comparing sequences to develop a hypothesis about the function of a newly discovered gene; breaking down known 3D protein structures into free amino acids to find patterns that can help predict how the protein folds; or modeling how proteins and metabolites in a cell work together to make the cell function. For example, multiple sequence alignment tools have been proved one of the most powerful computational tools available to the molecular biologist. Where one sequence is of unknown structure and function, it's alignment with other sequences that are well characterized in all structure and function immediately reveals the structure and function of the first sequence [6]. Another goal of developing analytical programs is to develop predictive methods that allow researchers to model the function of an organism based only on its genome sequence or shape.

Glycosylation prediction is a sub field of employing analytical computing techniques to develop predictive methods. In this field, various methods are implemented to predict the potential glycosylation sites on proteins. The original prediction method is based on searching the consensus sequence patterns. These patterns have always been inflexible and have not allowed for sequence substitutions. Weight matrices have been used as a better prediction method, because it allows for more diverse patterns and provides the opportunity for a scoring

scheme to rank potential hits. Currently, more complex methods, such as the Hidden Markov Models (HMMs) and Artificial Neural Networks (ANNs) are used to do Glycosylation site prediction. These methods can classify complex motifs containing positional correlations with functional sites.

CHAPTER 2

BACKGROUND

2.1 OVERVIEW

This chapter introduces the concept of Glycosylation and the importance of Glycosylation prediction. Current glycosylation sites prediction approaches are discussed following that, with the focus on the limitations of the traditional prediction methods.

2.2 GLYCOSYLATION

Glycosylation is the process or result of addition of saccharides to proteins and lipids. Many proteins in eukaryotic cells are glycoprotein as they contain oligosaccharide chains covalently linked to certain amino acids [7]. This process is one of the four principal co-translational and post-translational modification steps in the synthesis of membrane, secreted proteins, and the majority of proteins synthesized in the rough ER undergo Glycosylation. It is an enzyme-directed site-specific process, as opposed to the non-enzymatic chemical reaction of glycation. Enzymes are bimolecular that catalyzes (i.e. increase the rates of) chemical reactions [8]. Almost all enzymes are proteins. In enzymatic reactions, an enzyme provides a specific

environment within which a given reaction is energetically more favorable. The molecules at the beginning of the process are called substrates. The distinguishing feature of an enzyme-catalyzed reaction is that it occurs within the confines of a pocket on the enzyme called the active site. The surface of the active site is lined amino acid residues whose substituent groups bind the substrate and catalyze its chemical transformation [9]. There are various mechanisms for glycosylation, although all share several common features[10]:

- Glycosylation is an enzymatic process,
- The donor molecule is an activated nucleotide sugar [10]
- The process of glycosylation is site-specific.

Glycosylation plays an important role in biological processes ranging from protein folding and subcellular localization, to ligand recognition and cell-cell interactions [10]. It is known to affect protein folding, localization and trafficking, protein solubility, antigenicity, biological activity and half-life, as well as cell-cell interactions [7]. We can see the structure of a glycoprotein in Figure 2.1.

Protein glycosylation can be divided into four main categories mainly depending on the linkage between the amino acid and the sugar. These are N-linked glycosylation, O-linked glycosylation, C-mannosylation and glycophospharidlyinositol (GPI) anchor attachments [11]. N-glycosylation is characterized by the addition of a sugar to the amino group (NH_2) of an asparagines. In O-linked Glycosylation, a sugar is attached to the hydroxy oxygen of serine and threonine side chains [10]. GPI anchors refers to glycophospharidly-inositol groups attached near the C-terminal of a protein chain, that anchor the protein to the cell membrane. C-mannosylation

is the attachment of an a-mannopyranosyl residue to the indole C2 of tryptophan via a C-C link, and occurs on the first tryptophan in the consensus sequence W-X-X-W (or in some cases, W-X-X-C and W-X-X-F) [7].

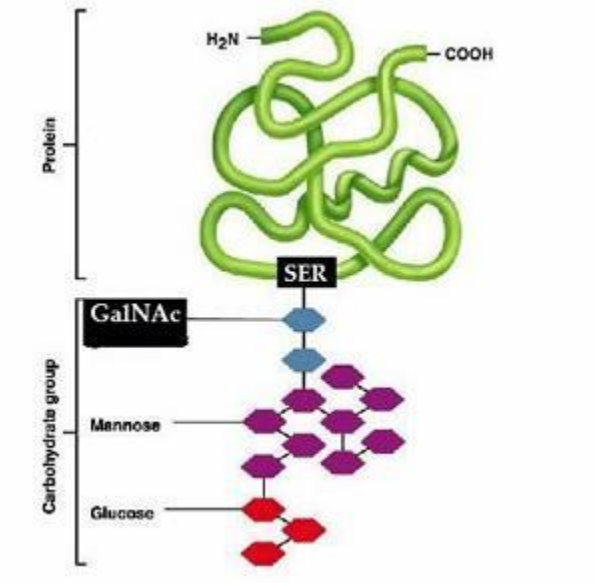

Figure 2.1 Glycoprotein

The attachment of the sugar N-acetylglucosamine (GlcNAc) to the proteins occurs after the proteins are folded into its 3D shape. The GlcNAc is then transferred to the side chains of serine or threonine residues on the surface of the proteins by the enzyme O-GlcNAc transferase (OGTase). This process is known as O-Glycosylation. O-linked glycosylation reactions may happen at two cellular locations in the cell [7]. Those taking place in the Golgi are initiated by the addition of various reducing terminal linkages such as N-acetylgalacosamine,

N-acetylglucosamine, mannose, fucose, phosphodiester linked N-acetylglucosamine, glucose, galactose or xylose to hydroxylamino acids (usually serine or threonine) [7].

However, not all serine or threonine residues on a protein surface are modified in this way. It is not known how the OGTase discriminates between the serine and threonine residues. Which serine or threonine residues in the protein will have GlcNAc added to them cannot be accurately predicted according to the primary amino acid sequences of proteins. There is no acceptor motif defined for O-linked glycosylation [7]. The only common characteristic among most O-glycosylation sites is that they occur on serine and threonine residues in close proximity to proline residues, and that the acceptor site is usually in a beta-conformation [7].

2.3 THE IMPORTANCE OF THE RESEARCH ON GLYCOSYLATION PREDICTION

The process of O-Glycosylation is a normal process in mammals, including humans. Normally, only 4.5% - 6% of blood glucose is covalently linked to the red blood cells in the hemoglobin of the non diabetes population. This value is commonly referred to as glycosylated hemoglobin or more specifically hemoglobin A1c [12]. Nevertheless, the increased amount and duration of glucose in the blood allows more Glycosylation to occur, not only with the hemoglobin, but with the proteins, and this can have systemic ramifications [13]. The excessive cleavage of glucose, especially with important protein amino groups, can affect the cell function and its structure and thus disequilibrate the existing balance, which leads to cell destabilization [14, 15]. This destabilization can lead to health problems, including type II diabetes. If we could

identify the requirements for Glycosylation, then it is possible that a drug could be developed to mediate this interaction and so treat those diseases such as diabetes.

2.4 TRADITIONAL GLYCOSYLATION PREDICTION METHODS

Experimental determination of glycosylated sites in proteins is an expensive and laborious process. Hence, there has been a significant interest in plenty of computational approaches to reliably predict the glycosylated sites from an amino acid sequence. Artificial Neural Network currently offers the most cost-effective approaches to construct the predictive models in applications where representative training data are available. Predicting Glycosylation sites, such as NetOGlyc [16], and YinOYang [17], provide Web accessible services for O-glycosylated sites prediction.

Artificial Neural Network (ANN) is a computational model based on the biological neural networks. It consists of an interconnected group of artificial neurons and it processes information using a connectionist approach of computation. In most cases, an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase [10]. AI techniques, including of complex representations, pattern recognition, search, and machine learning have been applied to the task of inferring and recognizing structural patterns, associated with molecular function. Back propagation ANNs are the most commonly used method for the prediction tasks. The networks are trained with information contained in the known protein sequences. The O-glycosylation methods of

predictions use back propagation for adjusting the weights to a set threshold value based on the surface accessibility of the amino acid residues.

2.5 LIMITATIONS OF CURRENT GLYCOSYLATION PREDICTION

Current prediction methods, such as prediction with ANN, are all based on the sequence pattern discovery. Nevertheless, it has been observed that one enzyme is responsible for the transfer of GlcNAc to many different proteins. This means that the topological structure of protein might affect the glycan binds. The interaction is greatly affected by protein structure and is often accompanied by conformational changes. For example, the specificity with which heme binds its various ligands is altered when the specificity with which heme binds its various ligands is altered when the heme is a component of myoglobin [9]. Normally there are more than one serine or threonine in a protein, but not all of them can be proven to be active sites to glycans. The Glycosylation may depend on the topological features of the vicinity near the glycosylation sites. Residues in vicinity nearby the glycosylation positions are not necessarily the neighbors of Glycosylation sites in the sequence or secondary domain. For example, in Figure 2.2, the vicinity nearby the Glycosylation site includes residues in more than one domain and the residues are not necessarily neighbors in the primary sequence.

The analysis based on the sequence pattern reorganization is unsuitable for the analysis of the topological features around the glycosylation site. Artificial neural networks based on the sequence pattern discovery can hardly make accurate identification of glycosylation sites. A new

method is expected to identify the topological features in the vicinity of serine or threonine residues, which are responsible for targeting the O-Glycosylation of those residues.

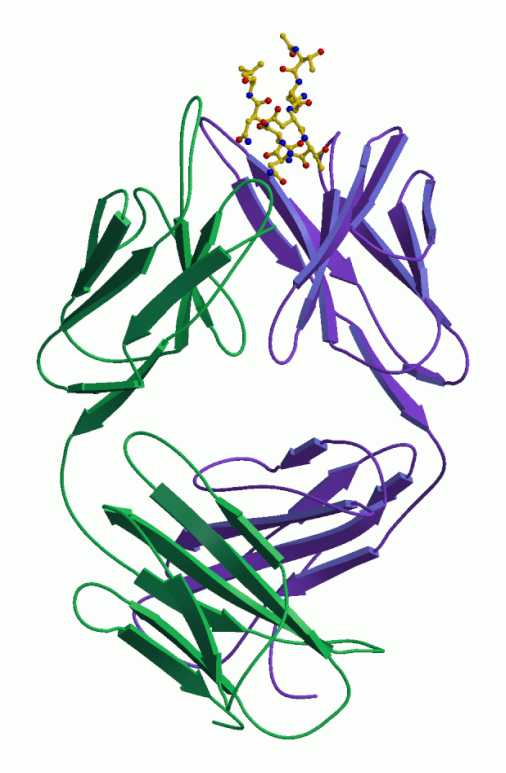

Figure 2.2 Vicinity nearby the glycosylation site

CHAPTER 3

GSP DESIGN

3.1 OVERVIEW

This chapter introduces a new glycosylation site prediction (GSP). The new method performs analysis and prediction method based on the comparison of a variety of protein topological features.

3.2 GLYCOSYLATION PREDICTION BY COMPARING TOPOLOGICAL FEATURES

The analysis of the topological features of proteins is based on protein models in Protein Data Bank format. The Protein Data Bank (PDB) format is a textual file format describing the three dimensional structures of molecules held in the Protein Data Bank. Most of the information in that database pertains to proteins, and the pdb format accordingly provides for rich description and annotation of protein properties [18], including the position information of all residues in the protein model. Figure 3.1 shows part of a protein model in the format of PDB.

```
356  ATOM     47  O   LYS A   6      28.213  36.753  16.411  1.00  5.76           O
357  ATOM     48  CB  LYS A   6      26.219  37.684  14.307  1.00  7.45           C
358  ATOM     49  CG  LYS A   6      25.884  39.139  14.615  1.00 11.12           C
359  ATOM     50  CD  LYS A   6      24.348  39.296  14.642  1.00 14.54           C
360  ATOM     51  CE  LYS A   6      23.865  40.723  14.749  1.00 18.84           C
361  ATOM     52  NZ  LYS A   6      22.375  40.720  14.907  1.00 20.55           N
362  ATOM     53  N   THR A   7      29.426  38.430  15.446  1.00  7.41           N
363  ATOM     54  CA  THR A   7      30.225  38.643  16.662  1.00  7.48           C
364  ATOM     55  C   THR A   7      29.664  39.839  17.434  1.00  8.75           C
365  ATOM     56  O   THR A   7      28.850  40.565  16.859  1.00  8.58           O
366  ATOM     57  CB  THR A   7      31.744  38.879  16.299  1.00  9.61           C
367  ATOM     58  OG1 THR A   7      31.737  40.257  15.824  1.00 11.78           O
368  ATOM     59  CG2 THR A   7      32.260  37.969  15.171  1.00  9.17           C
369  ATOM     60  N   LEU A   8      30.132  40.069  18.642  1.00  9.84           N
370  ATOM     61  CA  LEU A   8      29.607  41.180  19.467  1.00 14.15           C
371  ATOM     62  C   LEU A   8      30.075  42.538  18.984  1.00 17.37           C
372  ATOM     63  O   LEU A   8      29.586  43.570  19.483  1.00 17.01           O
373  ATOM     64  CB  LEU A   8      29.919  40.890  20.938  1.00 16.63           C
374  ATOM     65  CG  LEU A   8      29.183  39.722  21.581  1.00 18.88           C
375  ATOM     66  CD1 LEU A   8      29.308  39.750  23.095  1.00 19.31           C
376  ATOM     67  CD2 LEU A   8      27.700  39.721  21.228  1.00 18.59           C
377  ATOM     68  N   THR A   9      30.991  42.571  17.998  1.00 18.33           N
378  ATOM     69  CA  THR A   9      31.422  43.940  17.553  1.00 19.24           C
379  ATOM     70  C   THR A   9      30.755  44.351  16.277  1.00 19.48           C
380  ATOM     71  O   THR A   9      31.207  45.268  15.566  1.00 23.14           O
381  ATOM     72  CB  THR A   9      32.979  43.918  17.445  1.00 18.97           C
382  ATOM     73  OG1 THR A   9      33.174  43.067  16.265  1.00 20.24           O
383  ATOM     74  CG2 THR A   9      33.657  43.319  18.672  1.00 19.70           C
384  ATOM     75  N   GLY A  10      29.721  43.673  15.885  1.00 19.43           N
385  ATOM     76  CA  GLY A  10      28.978  43.960  14.678  1.00 18.74           C
386  ATOM     77  C   GLY A  10      29.604  43.507  13.393  1.00 17.62           C
```

Figure 3.1 Format of PDB models

The glycosylation prediction by comparing topological features includes two steps:

1. Selection of criterion which can be used to find the difference between topological features of proteins;

2. Assigning common amino acids into different groups and then calculating the similarity of the topological features nearby potential Glycosylation sites in each group.

In this experiment, Root Mean Square Deviation (RMSD) [10] is used as the criterion of measuring the similarity of the topological feature of different proteins. RMSD is a frequently-used measurement of the differences between the values predicted by a model or an estimator and the values actually observed from the object being modeled or estimated. These differences are also called residuals, and the RMSD serves to aggregate them into a single measurement of predicting power [10]. We use the RMSD to measure the difference between proteins. This means we attempt to find similar topological features around glycosylation sites on various proteins, but our method is not limited to the comparison among the primary sequences or secondary domains. The compared objects can be expressed as random vectors as shown in Figure 3.2. The corresponding RMSD formula is shown in Figure 3.3. The θ1 and θ2 are coordinates of chosen residues around the Glycosylation sites in the referece protein and the target protein. A researcher assumes residues belong to different classes and each combination of these classes brings out one RMSD value. At the same time, according to the distance between the residue and the glycosylation site, the residues can be classified into different scopes. The comparison should be proceeding in all combination of classes and in different scopes.

RMSD is an efficient metric which can be used to establish the level of similarity between proteins. If we compare a protein model with another identical protein model, then the RMSD value of comparison should be zero. If in the comparison, a group of residues induce a quite

smaller RMSD value, then these residues have a greater chance of composing a common topological feature that may affect the process of glycosylation.

$$\theta_1 = \begin{bmatrix} x_{1,1} \\ x_{1,2} \\ \vdots \\ x_{1,n} \end{bmatrix} \qquad \text{and} \qquad \theta_2 = \begin{bmatrix} x_{2,1} \\ x_{2,2} \\ \vdots \\ x_{2,n} \end{bmatrix}$$

Figure 3.2 Ramdom Vectors that express the compared objects

$$\mathrm{RMSD}(\theta_1, \theta_2) = \sqrt{\mathrm{MSE}(\theta_1, \theta_2)} = \sqrt{\mathrm{E}((\theta_1 - \theta_2)^2)} = \sqrt{\frac{\sum_{i=1}^{n} (x_{1,i} - x_{2,i})^2}{n}}$$

Figure 3.3 RMSD formula

3.3 CLASSIFICATION OF AMINO ACIDS

There are a total of twenty common amino acids existing. Calculating RMSD based on the combination of all the amino acids produces enormous data. Since the twenty amino acids can be classified into a few classes and the amino acids in the same class have the same features in the process of glycosylation, a better choice is to calculate the RMSD of the combination of amino acid classes. In this experiment, the RMSD is calculated based on the combination of amino acid classes and those amino acids in the same class are considered as the same group of

residues during the calculation. This experiment defines 5 classes (see Table 3.1): Aliphatic, Aromatic, Polar uncharged, Positive charge, and Negative charge. There are a total of 31 subsets of the class combination.

Table 3.1 Residues and corresponding classes

Amino acids	**Classes**	**Amino acids**	**Classes**
ALA	*Aliphatic*	ARG	*Positive charge*
VAL	*Aliphatic*	SER	*Aromatic*
PHE	*Polar uncharged*	THR	*Aromatic*
PRO	*Aliphatic*	TYR	*Polar uncharged*
MET	*Aliphatic*	HIS	*Positive charge*
ILE	*Aliphatic*	CYS	*Aromatic*
LEU	*Aliphatic*	ASN	*Aromatic*
ASP	*Negative charge*	GLN	*Aromatic*
GLU	*Negative charge*	TRP	*Polar uncharged*
LYS	*Positive charge*	GLY	*Aliphatic*

3.4 RATED ROOT MEAN SQUARE DEVIATION

During the topological feature comparison, the positions of the same type of amino acids are considered more important than those in other classes. For instance, there is a common characteristic among most O-glycosylation sites is that they occur on serine and threonine residues in close proximity to proline residues, and that the acceptor site is usually in a beta-conformation. RMSD will not tell such information. If we simply use RMSD as the measurement, we cannot tell which residues are more important than others. So, we should modify the formula of RMSD to set the priority for the residues in the same class when we do the comparison between difference proteins. Considering the importance of same-class amino acids in the topological comparison, the root mean square deviation is modified to r-RMSD. We changed the RMSD formula by adding the rate of "number of same-type residues" to "number of all residues" in observed scopes. The r-RMSD fomula is shown in Figure 3.4. R-RMSD can help to find more information rather than just the topological similarity between proteins.

$$\mathrm{r-RMSD} = \mathrm{RMSD} * \log_2 \sqrt{\frac{number\ of\ same\ type\ residue}{(number\ of\ all\ residues)^2} + 2}$$

Figure 3.4 R-RMSD formula

3.5 COMPARISON IN VARIOUS SUBSETS OF AMINO ACIDS

Amino acids on a protein surface are classified into different subsets in order to perform the comparison. First, the amino acids are classified into 31 groups according to their classes. Then the amino acids in each group are distributed into 3 subsets according to the distance between the amino acids themselves and the Glycosylation site. The distance subsets are set as the scopes of 5 angstroms, 10 angstroms and 15 angstroms. Therefore, there are a total of 93 subsets for each comparative protein, and each subset will produce an r-RMSD value.

3.6 SELECTION OF RESIDUES IN COMPARISON

Because the GlcNAc is only transferred to the side chains of serine or threonine residues on the surface of the protein, we only do comparison among residues on the surface of protein models. We use a tool named Michel Sanner's Molecular Surface (MSMS) to choose the residues on the surface of a protein. MSMS allows us to very efficiently compute triangulations of Solvent Excluded Surfaces [19]. Figure 3.5 shows a computing result of Solvent Excluded Surfaces of a protein model.

Atom#	ses_0	sas_0	
0	32.9783	67.9329	N_SER_2
1	7.7325	8.7802	CA_SER_2
2	1.3816	0.0416	C_SER_2
3	5.5840	3.4864	O_SER_2
4	24.5687	50.1580	CB_SER_2
5	10.4796	26.3680	OG_SER_2
6	10.1687	8.5838	N_GLU_3
7	9.8727	17.9169	CA_GLU_3
8	2.2535	1.4177	C_GLU_3
9	13.1965	21.3512	O_GLU_3
10	17.3341	28.2670	CB_GLU_3
11	12.5045	12.8475	CG_GLU_3
12	18.6119	32.3155	CD_GLU_3
13	11.3435	32.5419	OE1_GLU_3
14	12.0817	33.7669	OE2_GLU_3
15	2.1569	0.0762	N_ARG_4
16	3.5724	0.7536	CA_ARG_4
17	0.3371	0.0095	C_ARG_4
18	15.1417	5.0411	O_ARG_4
19	14.5829	12.5307	CB_ARG_4
20	14.9512	17.5193	CG_ARG_4
21	9.6564	4.6594	CD_ARG_4
22	12.2160	5.7776	NE_ARG_4
23	0.0000	0.0000	CZ_ARG_4

Figure 3.5 Solvent Excluded Surfaces

CHAPTER 4

GSP IMPLEMENTATION

4.1 OVERVIEW

This chapter describes how to use the GSP program for RMSD and r-RMSD calculation and how to view and analyze the comparison results.

4.2 THE IMPLEMENTATION OF PROGRAM

To do glycosylation analysis with the GSP program, users are expected to follow the steps below:

1. Choose the protein models and specify the expected active sites;
2. Run the comparison between the reference protein model and each of the comparative protein models provided in the step one;
3. View and analyze the comparison results.

In the first step, the user needs to generate an input file for r-RMSD calculation. The input file contains protein file paths and related glycosylation sites. The syntax of the input is:

Path of protein—glycosylation site

The input file should have an extension ".path". Below is the content of an example input file:

D://aladdin_data/projects/GSS/rmsd/source_pdb/1PK8_Ser261.pdb--261

D://aladdin_data/projects/GSS/rmsd/source_pdb/1svmf_Ser532.pdb--532

D://aladdin_data/projects/GSS/rmsd/source_pdb/Thr1889.pdb—1889

In this file, the first line specifies the reference protein model, and following lines specify the comparative protein models. The program performs the comparison between the reference model and each of the comparative models.

After the program starts up, the user can start the r-RMSD calculator through the second item of the "Tools" menu. As shown in Figure 4.1, there are three buttons on the top of the calculation panel. The first button with a folder icon is used to set the path of the RMSD input file, which is generated in the first step. The second button with the save icon is to designate the path to save calculation results. The third button "Calculate RMSD" starts the calculation process. When the calculation is done, a message box is displayed to notify the user. The comparison result is stored in a file with the extension ".result".

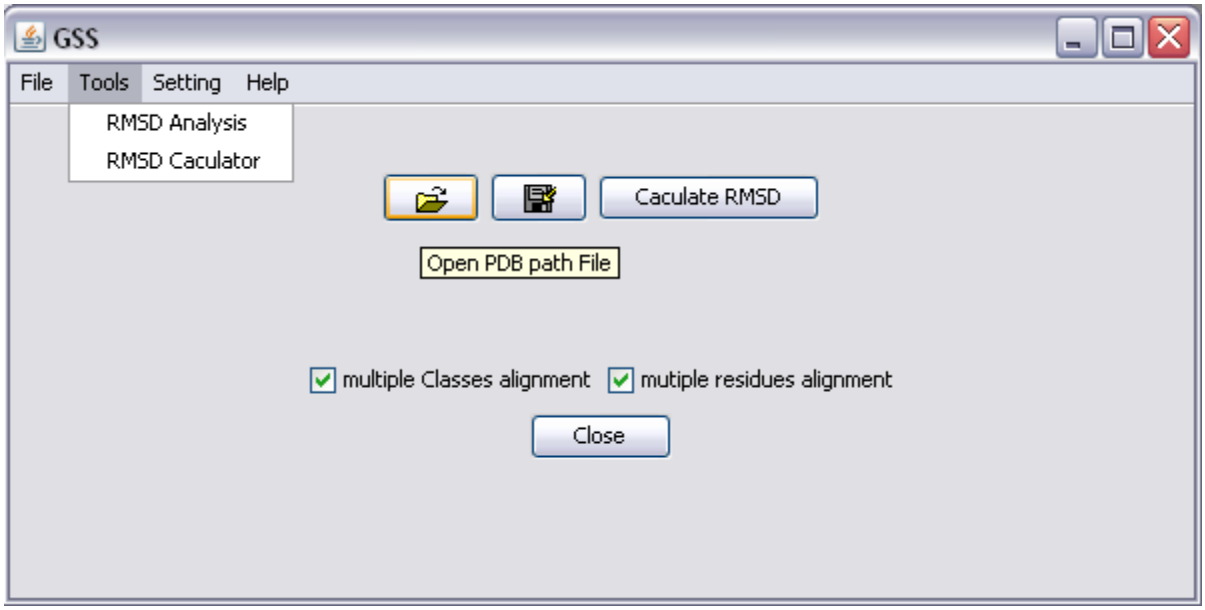

Figure 4.1 The interface of RMSD calculation tool

The user can start the analysis program by clicking the first item in the "Tools" menu. The interface of the analysis program is shown in Figure 4.2. The user needs to click on the two top left buttons to set the path of two files. The first button specifies the input file generated in the first step and the second button sets the path of the output file generated in the second step. After setting the paths of two files, the comparison results will be displayed in the window below. Figure 4.3 shows an instance of table which contains comparison results. In this table, the user can view the RMSD value, the r-RMSD value and related information such as the chosen scope, classes combination, number of residues involved, and the angle used in the comparison. The user can also click on the column title to sort the records to have a better view of the data. Each row in the sheet shows on an r-RMSD value with the corresponding data set.

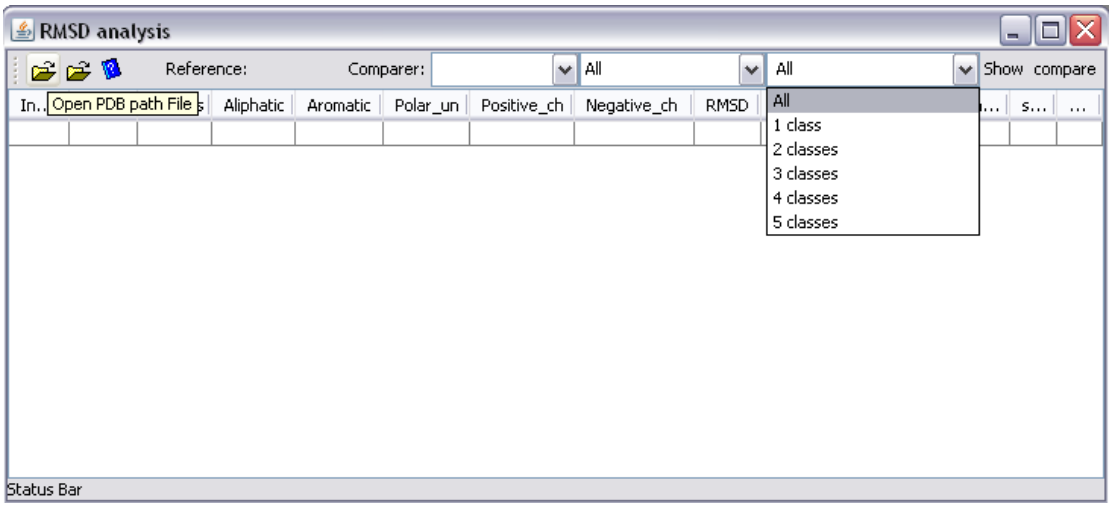

Figure 4.2 The interface of RMSD analysis frame

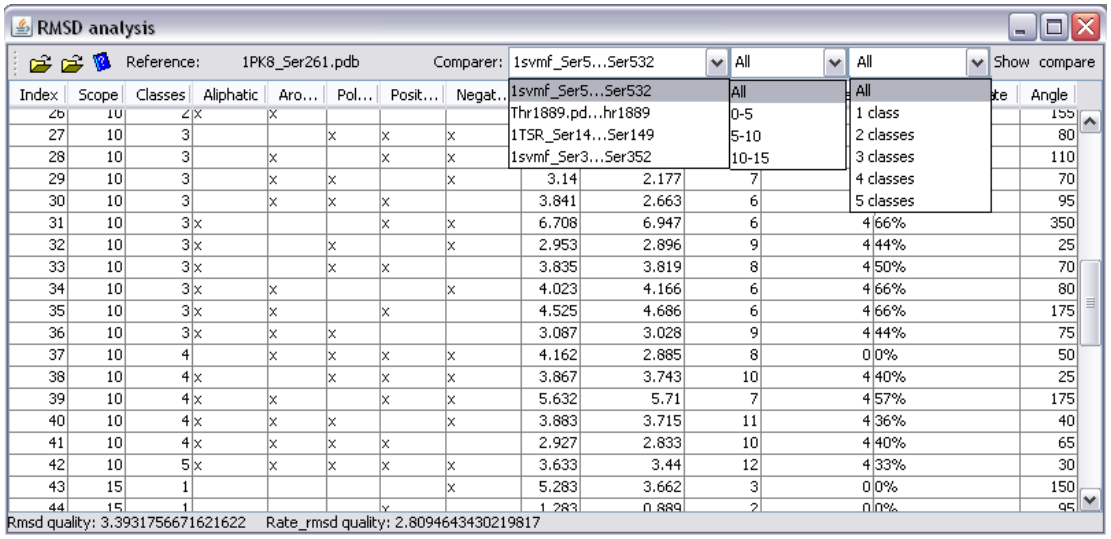

Figure 4.3 The interface of options in RMSD analysis frame

The user can view the comparison results coming from different comparative models by choosing model names from the drop down list. The user can also customize the table content by choosing different subsets of the data through choosing the limiting condition from drop down lists for scopes and classes.

The *compare* button in the top taskbar provides the view of a stereographic projection [18] of the selected comparison record in the sheet. The stereographic projection is a particular mapping (function) that projects a sphere onto a plane. The projection is defined on the entire sphere, except at one point — the projection point. Where it is defined, the mapping is smooth and bijective. It is also conformal, meaning that it preserves angles. On the other hand, it does not preserve area, especially near the projection point. Figure 4.4 shows how residues on a protein surface are projected into a flat surface. The Figure 4.5 shows the projection of the matched residues. The center of the graph is the glycosylation site, and each line in the graph connects a circle and a solid dot. The circle and the solid dot are the two matched residues. Circles represent the residues from the reference protein and solid dots are the matched residues in the protein being compared. As shown in Figure 4.6 and 4.7, the *show* button in the top taskbar provides the stereographic projection view of the reference protein and the comparative protein. The user can rescale, resize or rotate the image. The checkboxes at the bottom provide the options of viewing residues in different class subsets.

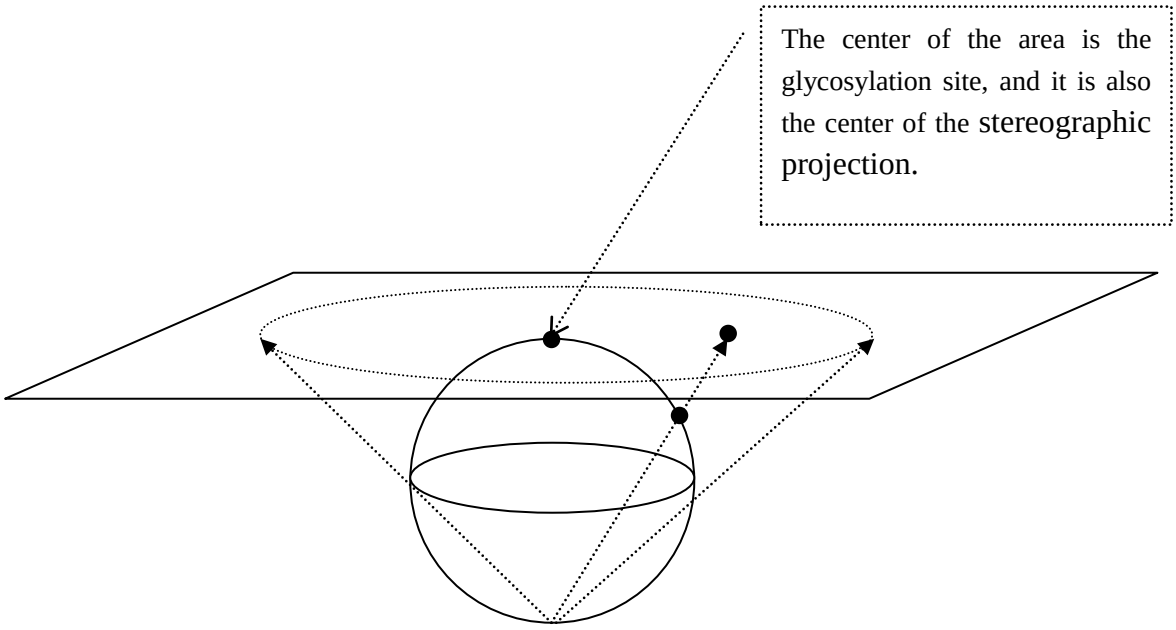

Figure 4.4 Stereographic projection

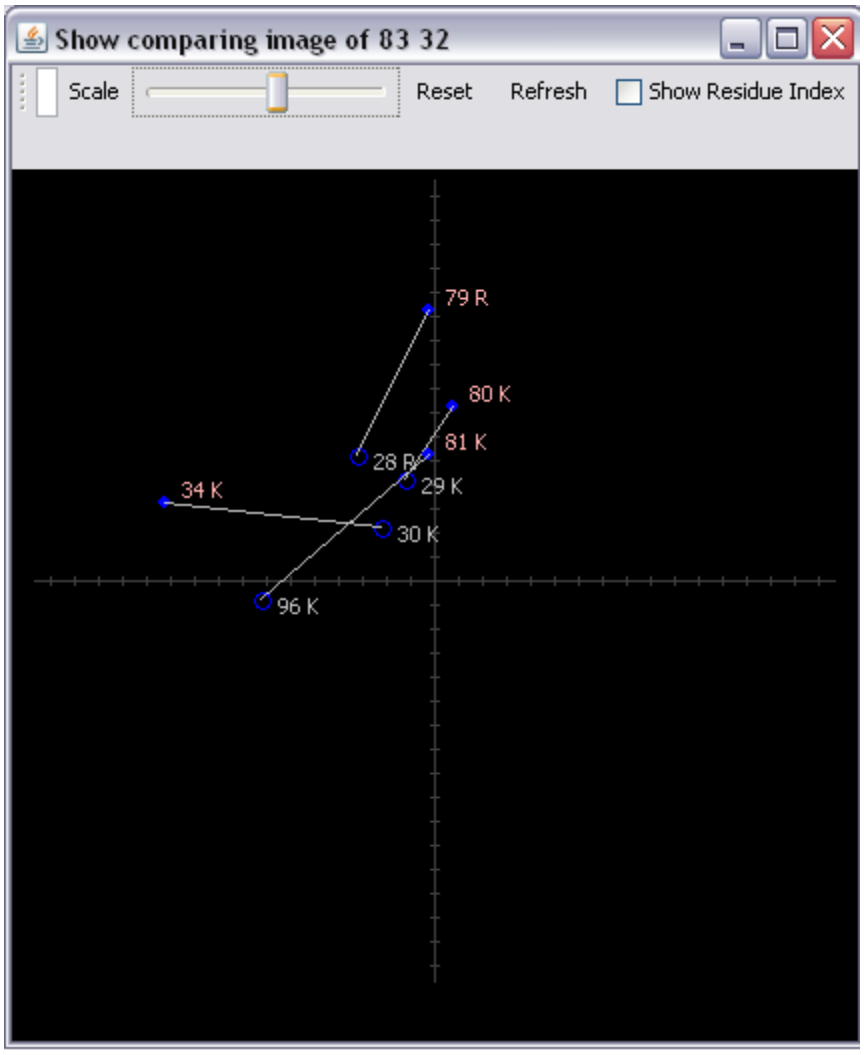

Figure 4.5 Projection of the matched residues

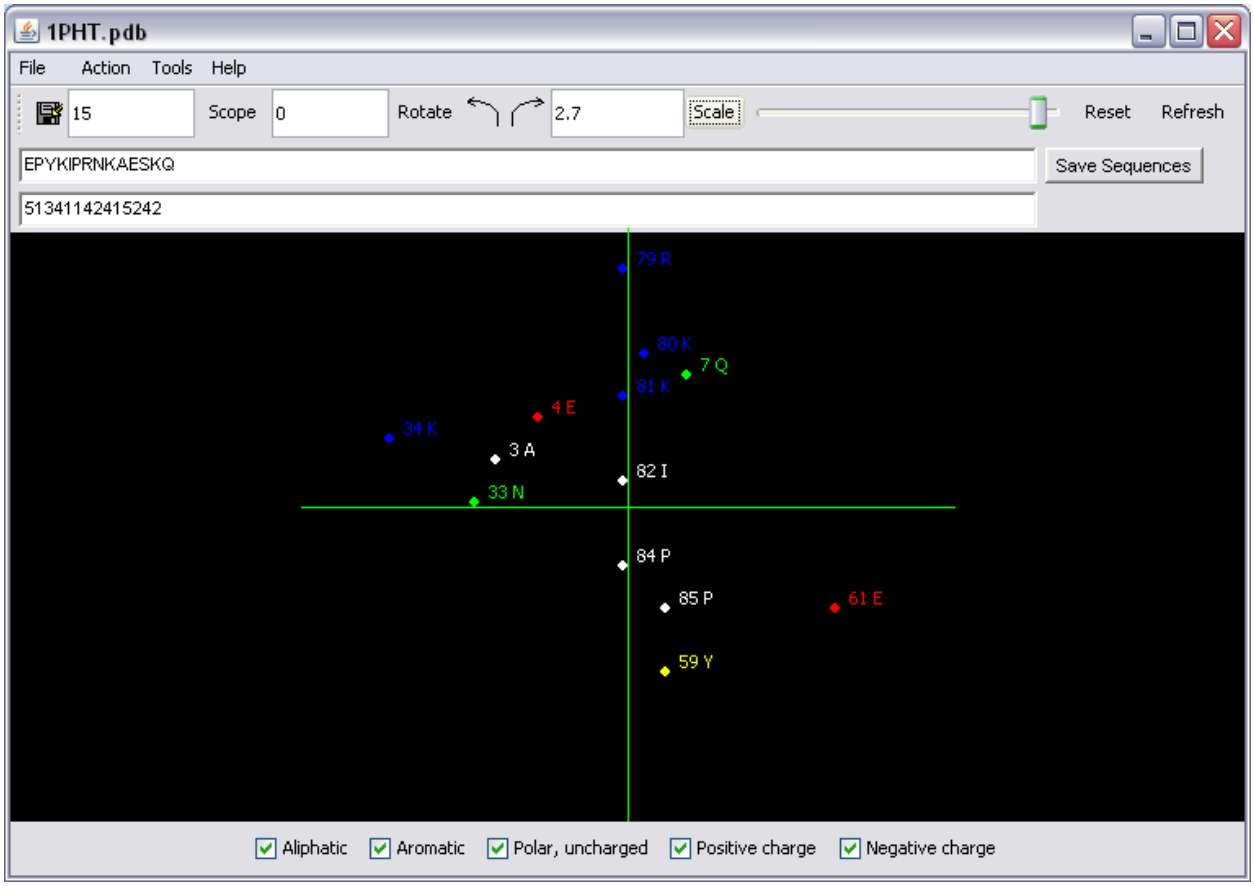

Figure 4.6 Projection of the selected residues in the reference protein

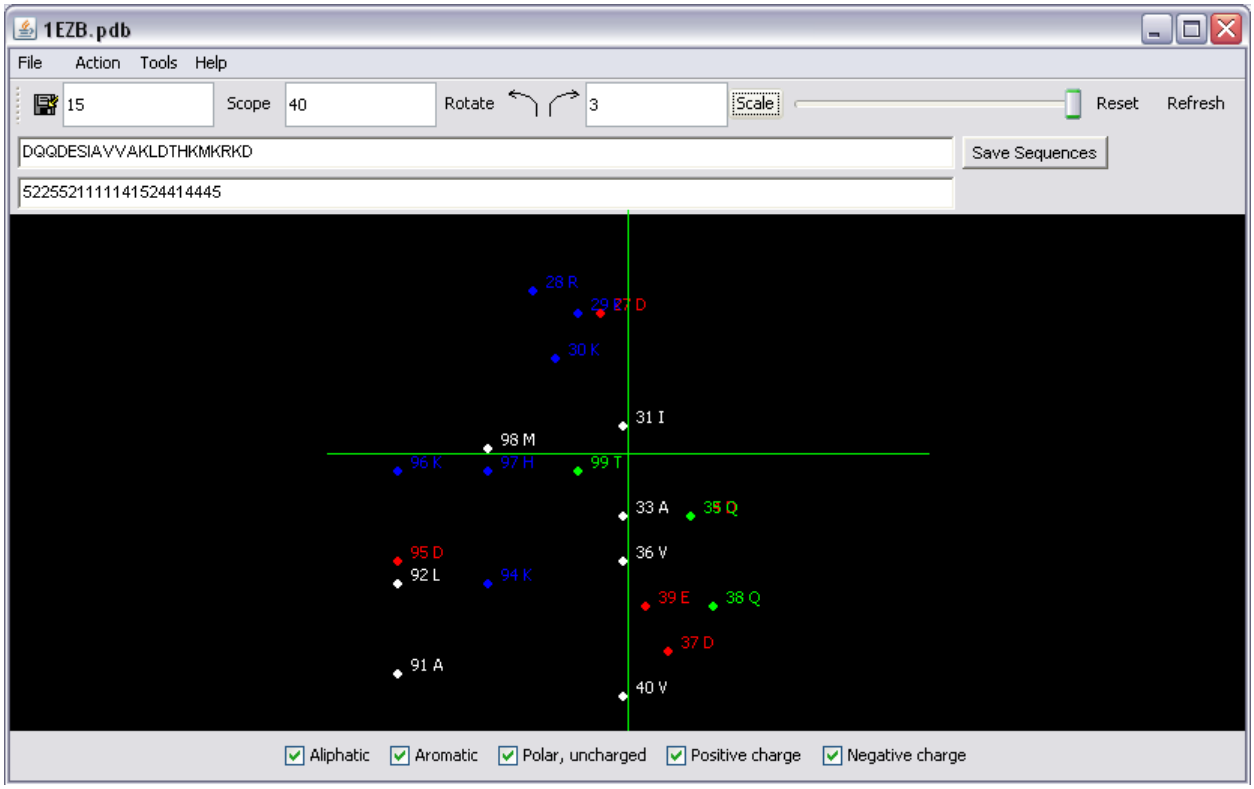

Figure 4.7 Projection of the selected residues in the comparative protein

CHAPTER 5

GSP EVALUATION

5.1 OVERVIEW

This chapter describes the measurement of the quality of the GSP program's performance. In the measurement, we use both simulated data and actual data to do tests.

5.2 EVALUATION STRATEGY

In order to evaluate the GSP program, two parameters are of utmost importance: sensitivity and specificity. The program should detect the differences between the target proteins at the highest degree (sensitivity). On the other hand, the program should ensure that the difference found by the program is positive (specificity).

To test the sensitivity, three test cases have been used in the program:

1. A comparison between a protein and a copy of it. In this case, two proteins in the comparison have been identical, and so the program should detect no differences between the two proteins;
2. A comparison between a protein and a modified copy. In this case, the coordinate of a surface residue in the copied protein is slightly changed. The

program should be able to detect the difference when the amount of the residue's coordinate changes;

3. A comparison between a protein and a modified copy. In this case, the type of a few residues is modified in copied proteins. The program should be able to detect the difference when the number of modified residues increases.

5.3 SENSITIVITY

The test has been performed to show that the program has good sensitivity in detecting the differences between proteins. In the comparison of two identical proteins, the program shows that all RMSD and r-RMSD values are zero. When one residue's coordinate is slightly modified in the copied protein, the RMSD and r-RMSD value in the comparison results are no longer zero. The RMSD and r-RMSD values show growth in step with the increase of the change in the residue's coordinate. The results are shown in Table 5.1 and Figure 5.1. By changing the coordinates of all residues with a step of 0.2, we can see the corresponding RMSD and r-RMSD values show a monotonic increment.

The changes in the residue type also lead to a difference in the comparison results. As shown in Table 5.2 and Figure 5.2, the RMSD and r-RMSD values in comparison results show a difference when the number of modified residue changes. The number of modified residues changed from 0to 7 (changed residue indice: 180, 268, 261, 262, 260, 22, 25) and the RMSD and r-RMSD values show a corresponding monotonic increment.

Table 5.1 Comparison result of coordinate changed residues

Change in Coordinate	Scope	Classes	RMSD	R-RMSD
0.4	15	5	0.376	0.414
0.6	15	5	0.526	0.557
0.8	15	5	0.873	0.959
1.2	15	5	1.974	2.13

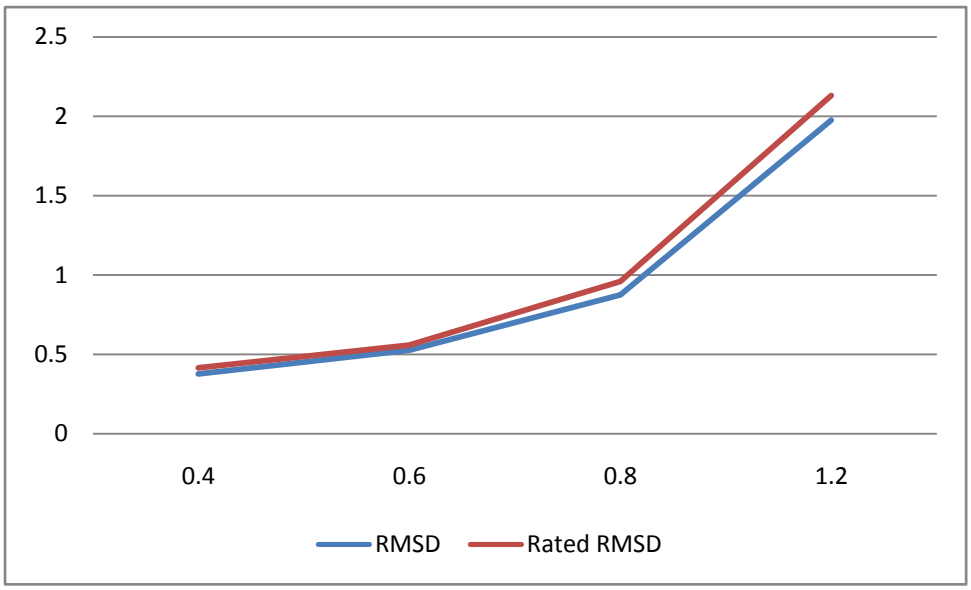

Figure 5.1 Alignment of target sequence

Table 5.2 Comparison result of class changed residues

Number of changed residues	Scope	Classes	RMSD	R-RMSD
1	15	5	0.388	0.423
2	15	5	0.388	0.425
3	15	5	0.408	0.442
4	15	5	0.426	0.457
5	15	15	0.426	0.457

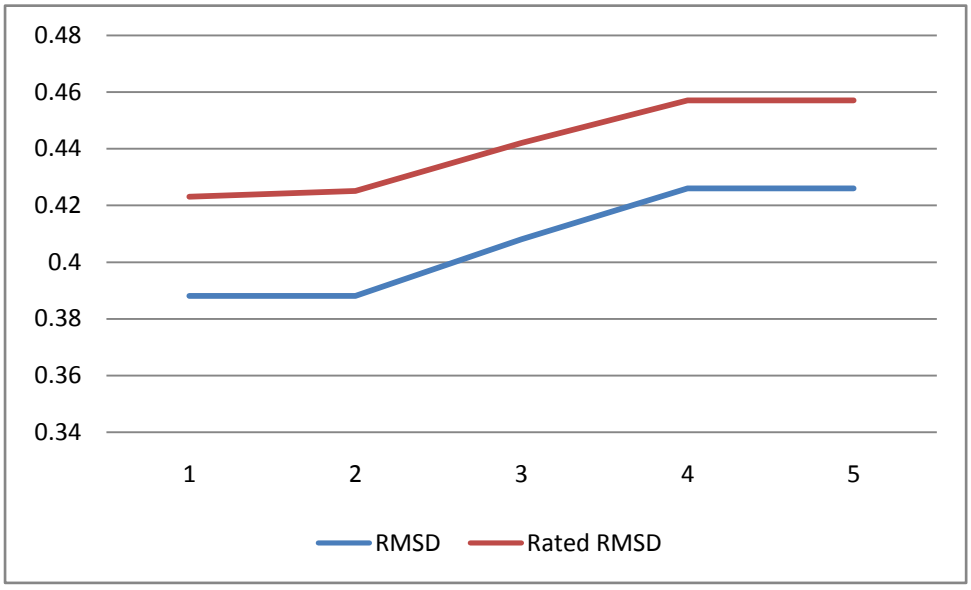

Figure 5.2 Alignment of target sequence

5.4 SPECIFICITY

To test the specificity of the program, two sets of tests have been conducted:

1. A comparison between a protein and a modified copy of it. In the copied protein, the coordinate of one or more surface residues have been slightly modified. In the comparison, this change should only affect the comparison results that involve the modified residues. It means that comparison results that do not involve the modified residue should be the same as the result from the identical proteins;
2. A comparison between a protein and a modified copy in which the type of one or more residues have been changed. In this case, this change should also only affect the comparison results that involve the modified residues. It means that comparison results from other unchanged residues should have no difference from the comparison results of identical proteins;

In the test results, when the copied protein includes the modified residues, some RMSD and r-RMSD values in comparison results are no longer zero. These positive RMSD and r-RMSD values result from comparisons that involve the modified residues. Comparisons that exclude these residues still produce the same result as from the identical proteins. For example, we pick *1PHT* as the test protein model and changed the type of residue 81 from *LYS* to *ALA* in a copy of *1PHT*. In the comparison result of *1PHT* and its modified copy, we can find that the RMSD and r-RMSD values of matched alignments are no longer zero when the alignments include the modified residue. At the same time, the RMSD and r-RMSD values of other alignments that

exclude the modified residue still keep to be zero. This means that the modified residues don't have an effect on the comparison results that exclude these residues. This result proves the specificity of the program.

5.5 EXPERIMENTS WITH ACTUAL EXPERIMENTAL DATA

We have used the actual experimental data of glycosylated proteins to evaluate the GSP program. Models of glycosylated proteins are not easy to find, so we have used phosphorylation models instead of the glycosylation models. Phosphorylation is similar to glycosylation and it usually occurs on serine, threonine, and tyrosine residues in eukaryotic proteins [10]. Both glycosylation and phosphorylation are found on serine and threonine side chains of nucleoplasmic proteins. All of the known GlcNAcylated proteins are also phosphoproteins and many proteins are modified by O-GlcNAc or Ophosphate on the same or at proximal sites [20]. So, we suppose that glycosylation and phosphorylation share a common catalytic mechanism.

Protein phosphorylation is the primary means of switching the activity of a cellular protein rapidly from one state to another [7]. It is considered as being a key event in many signal transduction pathways of biological [7]. Phosphorylation is the addition of a phosphate (PO4) group to a protein or other organic molecule. Reversible phosphorylation of proteins is an important regulatory mechanism that occurs in both prokaryotic and eukaryotic organisms. Enzymes called kinases (phosphorylation) and phosphatases (dephosphorylation) are involved in this process [10]. Protein kinases are converter enzymes that catalyze the ATP-dependent

phosphorylation of serine, threonine, and/or tyrosine hydroxyl groups in target proteins. Phosphorylation introduces a bulky group bearing two negative charges, causing conformational changes that alter the target protein's function [21]. Protein kinases represent a protein superfamily whose members are widely diverse in terms of size, subunit structure, and subcelluar localization. Protein Kinases are classified as Ser/Thr and /or Tyr-specific and are subclassified in terms of the allosteric activators they require and the and the consensus amino acid sequence within the target protein that is recognized by the kinase [21]. For example, as shown in Table 5.3, cAMP-dependent protein kinase (PKA) phosphorylates proteins having Ser or Thr residues within an R(R/K) X(S*/T*) target consensus sequence. That is PKA phosphorylates Ser or Thr residues that occur in an Arg-(Arg or Lys)-(any amino acid)-(Ser or Thr) sequence segment [21].

Table 5.3 Ser/Thr protein kinase [21]

Protein Kinase Class	Target Sequence	Activatiors
cAMP-dependent (PKA)	-R(R/K)X(S*/T*)-	cAMP
cAMP-dependent	-(R/K)KKX(S*/T*)-	cAMP
Phosphorylase kinase (PhK)	-KRKQIS*VRGL-	Phosyphorylation by PKA
MAP kinases	-PXX(S*/T*)P-	Phosyphorylation by MAPK kinase

* denotes glycosylation site

X denotes any amino acid

5.6 PREDICTION OF TARGET SEQUENCE

From *Protein Data Bank* [22], we have found proteins models that contain the consensus sequences shown in Table 5.3 [21]. By setting the first protein model *1PHT* as the reference, the glycosylation prediction program has been able to detect all possible matched sequences in all target proteins. Figure 5.4 shows one of the matched alignments in the two proteins. Figures 5.3 to 5.9 show the alignments found by the GSP program. We can see that in all six target protein models, there are sequences with the same consensus sequence RKKXS*/T*. For example, the program detects that residues 215, 216, 217 (*ARG*, *LYS*, *LYS*) in comparative protein model *1FY7* are matched to residues 79, 80, 81 (*ARG*, *LYS*, *LYS*) in the reference protein *1PHT*. At the same time, residue 219 (*SER*) in protein model *1FY7* is the predicted phosphorylation site and residue 83 (*SER*) is the predicted phosphorylation site in protein model *1PHT*. From the Table 5.3 we already know that the sequence RKKXS is the motif of the target protein of PKA. This means that the program found the target sequence RKKXS through searching for similar ontology in reference and the comparative proteins. To verify the applicability of the GSP program, we have picked 3 groups, a total of 20 proteins in the comparison test. The test results show that the program is compatible to other models. It is able to detect the matched target sequences in all comparative proteins in the test. Based on the already discussed similarity of glycosylation and phosphorylation, the GSP program should also be able to work in the same problems on the glycoprotein models.

Table 5.4 Proteins that have the common motif "RKKXS*/T*"

Target protein	Index of motif	Residue in motif
1PHT	79, 80, 81, 82, 83	ARG, LYS, LYS, ILE, SER
1BT4	269, 270, 271, 273, 273	ARG, LYS, LYS, ALA, SER
1FY7	215, 216, 217, 218, 219	ARG, LYS, LYS, CYS, THR
1BVB	8, 9, 10, 11, 12	ARG, LYS, LYS, CYS, SER
1EZA	28, 29, 30, 31, 32	ARG, LYS, LYS, LIE, SER
1EZB	28, 29, 30, 31, 32	ARG, LYS, LYS, LIE, SER
1EZC	28, 29, 30, 31, 32	ARG, LYS, LYS, LIE, SER
1EZD	28, 29, 30, 31, 32	ARG, LYS, LYS, LIE, SER

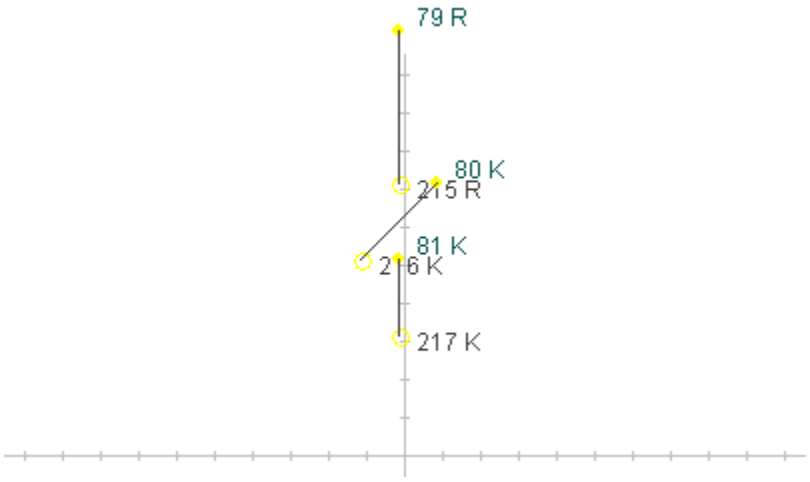

Figure 5.3 Alignment of matched motif in "1FY7**"**

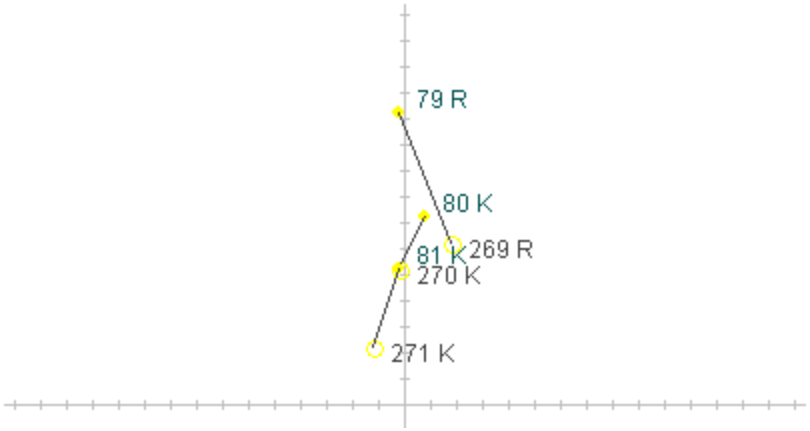

Figure 5.4 Alignment of matched motif in "1BT4**"**

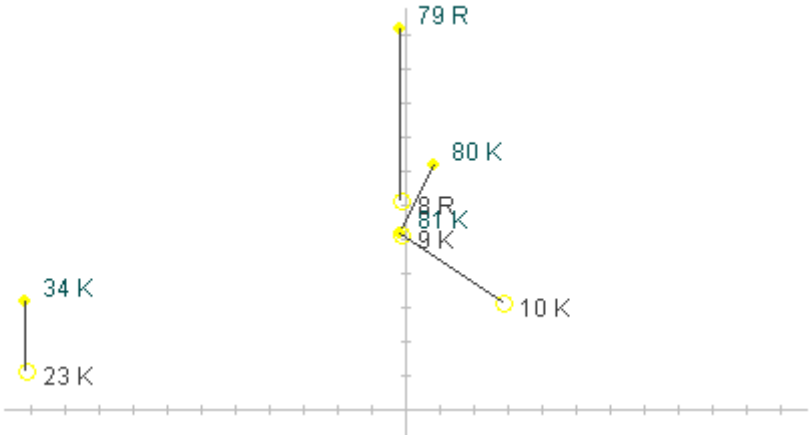

Figure 5.5 Alignment of matched motif in "1BVB**"**

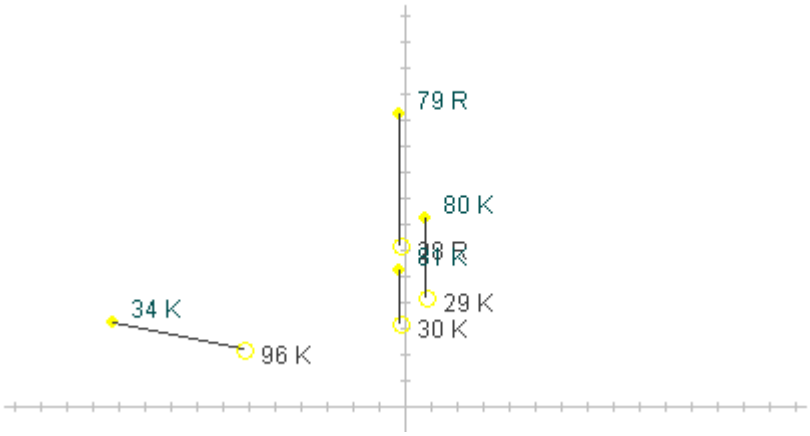

Figure 5.6 Alignment of matched motif in "1EZA**"**

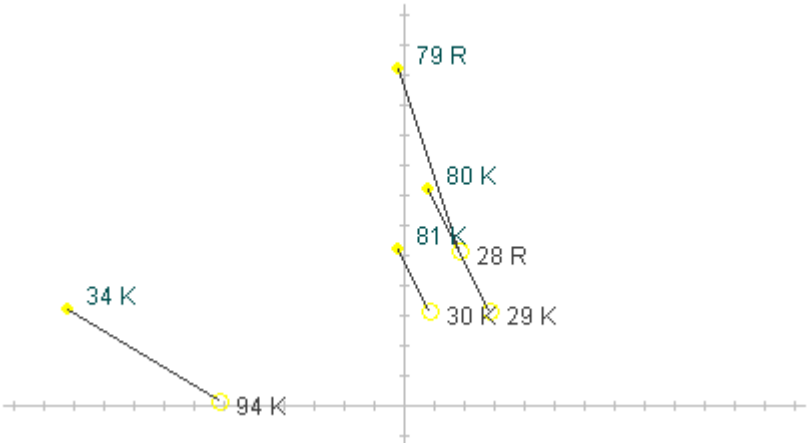

Figure 5.7 Alignment of matched motif in "1EZB**"**

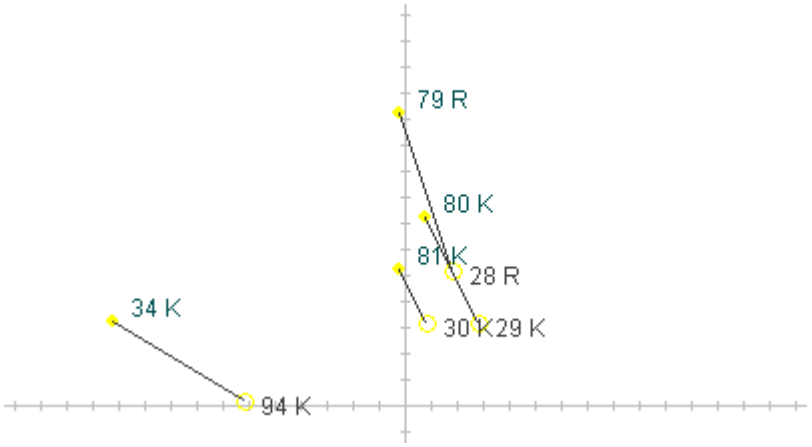

Figure 5.8 Alignment of matched motif in "1EZC**"**

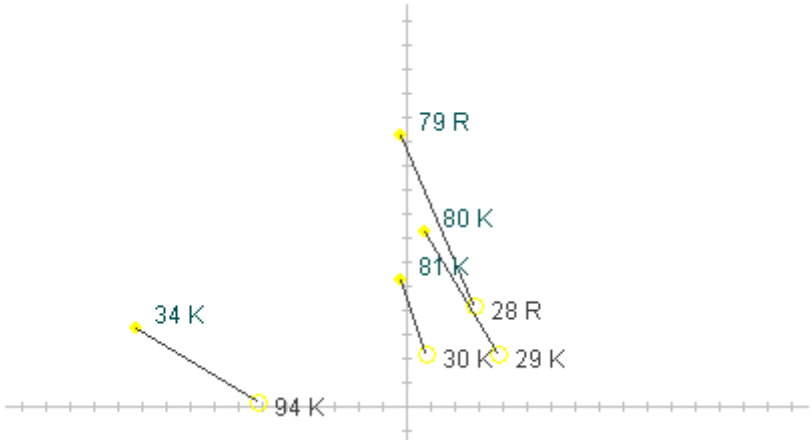

Figure 5.9 Alignment of matched motif in "1EZD**"**

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

The goal of the glycosylation prediction program is to identify the structural features in the vicinity of potential glycosylation sites, which are responsible for directing the enzyme to a particular position in the protein. In this experiment, we designed a GSP program which uses RMSD and r-RMSD as the criterions of measuring the similarity of the topological feature of different proteins. The GSP program is able to compare the topological differences of proteins based on PDB format models. The GSP program also provides the view of comparison results and the view of a stereographic projection of the selected comparison record. So, with the GSP program, the user is able to find and analyst the possible similarity in topological features of protein models. To verify the performance of GSP program, we did sensitivity and specificity tests with simulated data. We also tested the GSP program with actual experimental data. The test cases have proved that the GSP program is able to identify the difference as well as detect the similarity between proteins. The performance of the program is encouraging since it is able to identify the particular consensus sequence, which is responsible for phosphorylation sites in proteins.

Further improvements using this implementation of the RMSD are currently being considered. Firstly, currently the GSP program is still not able to do comparison on the user

specified vicinity. It might be helpful if the GSP program could perform the comparison in the vicinity in which the user is interested. Secondly, the program is expected to perform more accurate comparison. Thirdly, the program should improve the calculation speed by optimizing the RMSD algorithm. At the same time, a better understanding of the mechanism of glycosylation will be helpful for future improvements to stereoselectivity and regioselectivity. We anticipate that this program would work well if we can improve the GSP program in the above aspects.

REFERENCES

1. Lesk, A.M., *Introduction to Bioinformatics*. 2002: Oxford University Press, USA.
2. Myers, E., Altschul S.F., Gish W., Miller E.W., Lipman D.J., NCBI. *BLAST, Basic Local Alignment Search Tool*. [cited; Available from: http://en.wikipedia.org/wiki/BLAST.
3. W.R., P. *FASTA is a DNA and Protein sequence alignment software package first described (as FASTP) by David J. Lipman and William R. Pearson in 1985 in the article Rapid and sensitive protein similarity searches*. [cited; Available from: http://en.wikipedia.org/wiki/FASTA.
4. RCSB. *The RCSB PDB provides a variety of tools and resources for studying the structures of biological macromolecules and their relationships to sequence, function, and disease.* . [cited; Available from: http://www.rcsb.org/pdb/home/home.do.
5. NCBI. *The National Center for Biotechnology Information (NCBI) provides an integrated approach to the use of gene and protein sequence information, the scientific literature (MEDLINE), molecular structures, and related resources, in biomedicine.* [cited; Available from: National Center for Biotechnology Information.
6. Webster, D.M., *Protein Structure Prediction: Methods and Protocols*. Vol. 143.
7. Nikolaj Blom1, T.S.-P., Ramneek Gupta1, Steen Gammeltoft2 and a.S. Brunak1, *Prediction of post-translational glycosylation and phosphorylation of proteins from the*

amino acid sequence. Proteomics 2004. **4**.

8. AD, S., *Oxford Dictionary of Biochemistry and Molecular Biology*. 1997: Oxford University Press
9. David L. Nelson, M.M.C., *Lehninger Principles of Biochemistry*. 3rd edition ed.
10. Ajit Varki, R.C., Jeffrey Esko, Hudson Freeze, Gerald Hart, Jamey Marth, *Essentials of Glycobiology*. 1999, New York Cold Spring Harbor Laboratory Press Cold Spring Harbor.
11. Huang, Z., *Drug Discovery Research*: John Wiley and Sons.
12. H. Franklin Bunn, K.H.G., P. M. Gallop, *The Glycosylation of Hemoglobin: Relevance to Diabetes Mellitus.* Science, 1978. **200 7**: p. 21-27.
13. Lee PD, S.L., O'Day MR, Rognerud CL, Ou CN, *Comparisons of home blood glucose testing and glycated protein measurements.* Diabetes Res Clin Pract, 1992. **16(1)**: p. 53-62.
14. RE, B., *Nonenzymatically glycosylated proteins.* Adv Clin Chem, 1987. **26**: p. 1-78.
15. Wolff SP, J.Z., Hunt JV, *Protein glycation and oxidative stress in diabetes mellitus and ageing.* Free Radic Biol Med 1991. **10(5)**: p. 339-52.
16. K. Julenius, A.M., R. Gupta and S. Brunak. *Prediction, conservation analysis and structural characterization of mammalian mucin-type O-glycosylation sites*. 2005 [cited; Available from: http://www.cbs.dtu.dk/services/NetOGlyc/.
17. Gupta., R. *Prediction of glycosylation sites in proteomes: from post-translational modifications to protein function.* 2001 [cited; Available from:

http://www.cbs.dtu.dk/services/YinOYang/.

18. Snyder, J.P., *Flattening the Earth*. 1993: University of Chicago.

19. Sanner, M. *MOLECULAR SURFACES COMPUTATION*. [cited; Available from: http://www.scripps.edu/~sanner/html/msms_home.html.

20. Zihao Wang, M.G., and Gerald W. Hart, *Cross-talk between GlcNAcylation and phosphorylation: Site-specific phosphorylation dynamics in response to globally elevated O-GlcNAc.* PNAS, 2008. **105**.

21. Reginald H. Garrett, C.M.G., *Biochemistry*. 1999.

22. RCSB. [cited; Available from: http://www.rcsb.org/pdb/home/home.do.