UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

Prediction of alpha helices in proteins using Modified Logistic Regression Model

ORIENTADO: **Carmelina Figueiredo Vieira Leite**

ORIENTADOR: **Prof. Dr. Marcos Augusto dos Santos**

BELO HORIZONTE - MG

Agosto de 2016

Prediction of alpha helices in proteins using Modified Logistic Regression Model

Dissertação apresentada ao programa
de Pós-Graduação em Bioinformatica,
do Instituto de Ciências Biológicas,
da Universidade Federal de Minas Gerais
como requisito parcial para a obtenção do grau
de Mestre em Bioinformática

ORIENTANDO: **Carmelina Figueiredo Vieira Leite**

ORIENTADOR: **Prof. Dr. Marcos Augusto dos Santos**

To my parents, for their example and unconditional love.

To Rita de Matos, for her friendship and support.

João Gabriel Ramos Ribas, for his charity.

ACKNOWLEDGMENT

To my advisor, Professor Marcos, for the confidence and comprehension;

To the Financial Support, FAPEMIG e CNPQ;

To my relatives and friends, for the friendship and support;

To Joana Ferreira Ascensão, for the review and suggestions;

To PhD Lucianna Helene, for the availability;

To my colleagues of LBS, for the discussions and laughs.

Agradeço todas as dificuldades que enfrentei; não fosse por elas, eu não teria saído do lugar.

As facilidades nos impedem de caminhar. Mesmo as críticas nos auxiliam muito.

Chico Xavier

ABSTRACT

The advance in proteins secondary structure prediction produces directly impacts on health and biological processes knowledge. Despite the achievements and advances, the prediction of proteins structure remains a challenge. Considering this fact, we propose a *de novo* method for the prediction of alpha helix. Initially, we created a list of proteins with low identity between them, from the repository Protein Data Bank, using PISCES. Each protein was separated into fragments (of size 9) using the sliding window technique. From the obtained fragments, we classified them into the ones that were 100% a standard type alpha helix, the ones that were not a 100% of the same type of secondary structure. For each fragment, we used a sliding window of size 3 to characterize them. These had a value associated with the occurrence of the alpha helix structure. It was possible to predict the secondary structure group, alpha helix, of an unknown protein/query. To accomplish our goals, we used modified logistic regression and constructed two methods for prediction of these structures. Tests of accuracy and specificity applied to the methods gave results greater than 70%. Unfortunately, the sensitivity did not show good results. One of the methods revealed to be a very promising application for the secondary structure prediction problem, and to a possible usage in other purpose. All methods were implemented in MatLab R2015b (2015).

KEYWORDS: logistic regression, prediction, protein, structure.

RESUMO

O avanço na predição da estrutura secundária de proteínas produz diretamente impactos na saúde e no conhecimento de processos biológicos. Apesar das conquistas e avanços, a predição da estrutura de proteínas continua a ser um desafio. Neste trabalho, nós propomos um método *de novo* para a predição de alfa hélice. Primeiramente, criamos uma lista de proteínas com baixa identidade entre eles, a partir do Banco de dados Protein Data Bank, utilizando a ferramenta PISCES. Cada proteína foi separada em fragmentos de tamanho (9), utilizando a técnica de janela deslizante. Os fragmentos obtidos foram classificados em aqueles que são 100% alfa hélice do tipo padrão e aquelas que não têm 100% deste tipo de estrutura secundária. Para cada fragmento, utilizamos uma janela deslizante de tamanho 3 para caracterizar cada um. Estes tripletos têm um valor associado com a ocorrência da estrutura α hélice. Com isso, é possível prever a estrutura secundária de uma proteína desconhecida. Para isso, usamos regressão logística modificada e construídos dois métodos de predição. Testes de precisão, especificidade deram origem a resultados superiores a 70%. Infelizmente, a sensibilidade não teve um bom resultado. Um dos métodos criados revelou-se promissor, tanto para este problema quanto para os outros problemas. Todos os métodos foram implementados em Matlab R2015b (2015).

PALAVRAS-CHAVES: regressão logística, predição, protein, structure.

LIST OF FIGURES

LIST OF TABLES

SUMMARY

1 Introduction

1.1 General considerations

We have all heard of proteins in different areas, from childhood. Proteins are the most abundant biological macromolecules, occurring in all cells. Protein functions are varied, and many are the basis of complex physiological processes such as oxygen transport, immune function and muscle contraction (Lehninger *et al.*, 2005; Elliott e Elliott, 2009).

The discovered of proteins and amino acids has decades, however, about a few years later it was confirmed that the relationship between then. It was Emil Fischer and Franz Hofmeister, in 1902, who demonstrated that proteins are polypeptides (Jayanthi, 2010).

The term *protein is of Greek origin, proteios, meaning first, in origin.* The Dutch chemist Gerrit Mulder (1802-1880) studying some organic compounds suggested that there was an important compound and they were present in many organisms. He also suggested that this compound would be synthesized in plants, and they were passed to the animal kingdom, through feeding. Mulder wrote to Jacob Berzelius (1779-1848) in 1838, reporting the identification of a central substance in various organisms. Jacob Berzelius was a famous Swiss chemist, best known for his contribution to the discovery of several chemical elements (Ramos, 2004). Berzelius responded by suggesting the name "protein" for the compound because it was the predominant of animal nutrition. Berzelius also helped with his fame, to disclose the concept of protein at the time.

The first amino acid to be isolated was asparagine in 1806, by Vauquelin and Robiquet (Vauquelin e Robiquet, 1806). The last one to be found was threonine in 1938. All the amino acids have trivial or common names, in some cases derived from the source from which they were first isolated. Asparagine was isolated in asparagus, glutamate in wheat gluten and tyrosine from cheese (from Greek *tyros*, "cheese"). Glycine also has the name from Greek origin, because of the sweet taste that it has (Greek *glykos*, "sweet") (Lehninger *et al.*, 2005).

Pauling, Corey, and Branson (1951) defined the structure of alpha helix and β sheet for proteins; as a refinement of some previous works of Astbury and Bell (1941), Huggins (1943) and Bragg Kendrew and Perutz (1950). Linderstrom-Lang (1952) introduced the terms primary, secondary and tertiary to describe protein structure.

Around 1955, Frederick Sanger completed the insulin sequence, working in component residues and development of sequencing techniques of this hormone. Various procedures were used to analyze protein primary structure, proving that all proteins have specific structures (Sanger, 1988; Hodgman, 2000; Lehninger *et al.*, 2005).

In 1954, Anfinsen and his contributors studied the relation between the chemical structure and the catalytic activity of an enzyme. The studies were based on the folding of protein chains (Anfinsen *et al.*, 1954; Anfinsen, 1972).

The amino acid discovery is also important in various areas, for example, in astronomy. In August 2009, NASA scientists discovered the presence of glycine, a fundamental building block of life, in samples of comet Wild 2 returned by NASA's Stardust spacecraft. This glycine had an extraterrestrial carbon isotope signature, indicating that it originated on the comet. The dust captured by the probe, composed of primordial material was redirected back to Earth. A small portion of this sample was what brought these new results. These results corroborate the theory of cosmic panspermia that argues that some compounds on earth were formed in space and were delivered to Earth long ago by meteorite and comet impacts during his youth around the sun (http://www.nasa.gov/mission_pages/stardust/news/stardust_amino_acid.html).

Nutritionally, for young adults, skeletal muscle accounts for approximately 45% of total body weight. The recommended dietary allowance (RDA) for adults is 0.8 grams of protein per kilogram of weight *per day* (Chernoff, 2004; Brasil, 2009).

1.2 Proteins

Proteins are dehydrated polymers, with one or more chains, which consist of a vast number of amino acids residues linked together (Lehninger *et al.*, 2005; Elliott e Elliott, 2009). Every amino acid, except proline, has a carboxyl group and an amino group bonded to the same carbon atom, called α carbon (Figure 1). The link between two amino acids, called peptide bond (-CO-NH-), is a specific type of covalent bond (Figure 2).

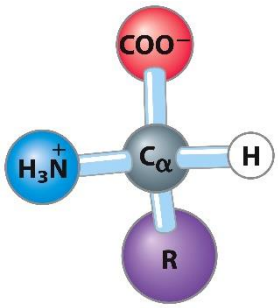

Figure 1 – Schematic representation of an amino acid (Lehninger, Nelson e Cox, 2005).

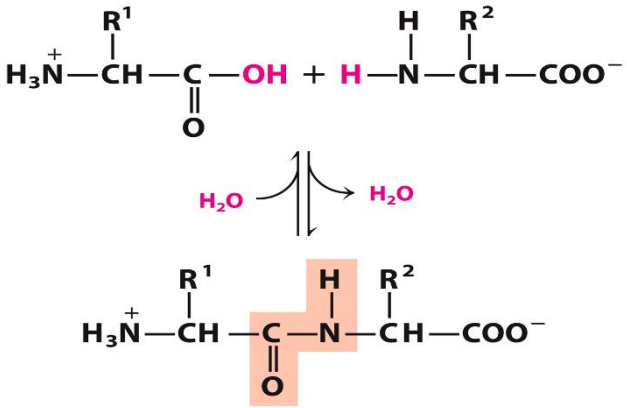

Figure 2 – Reversible formation of a peptide bond by condensation (Lehninger, Nelson e Cox, 2005).

To the α carbon is connected to an R group, which differ from each other in their chains. The R group varies in structure, size, electric charge and solubility in water. All the 20 standard amino acids (a term employed to distinguish them from less common amino acids that are residues modified after the protein has been synthesized) can be organized into five groups, according to their polarity (Lehninger *et al.*, 2005). They are divided in nonpolar, aliphatic; aromatic; polar, uncharged; positively charged and negatively charged (Figure 3).

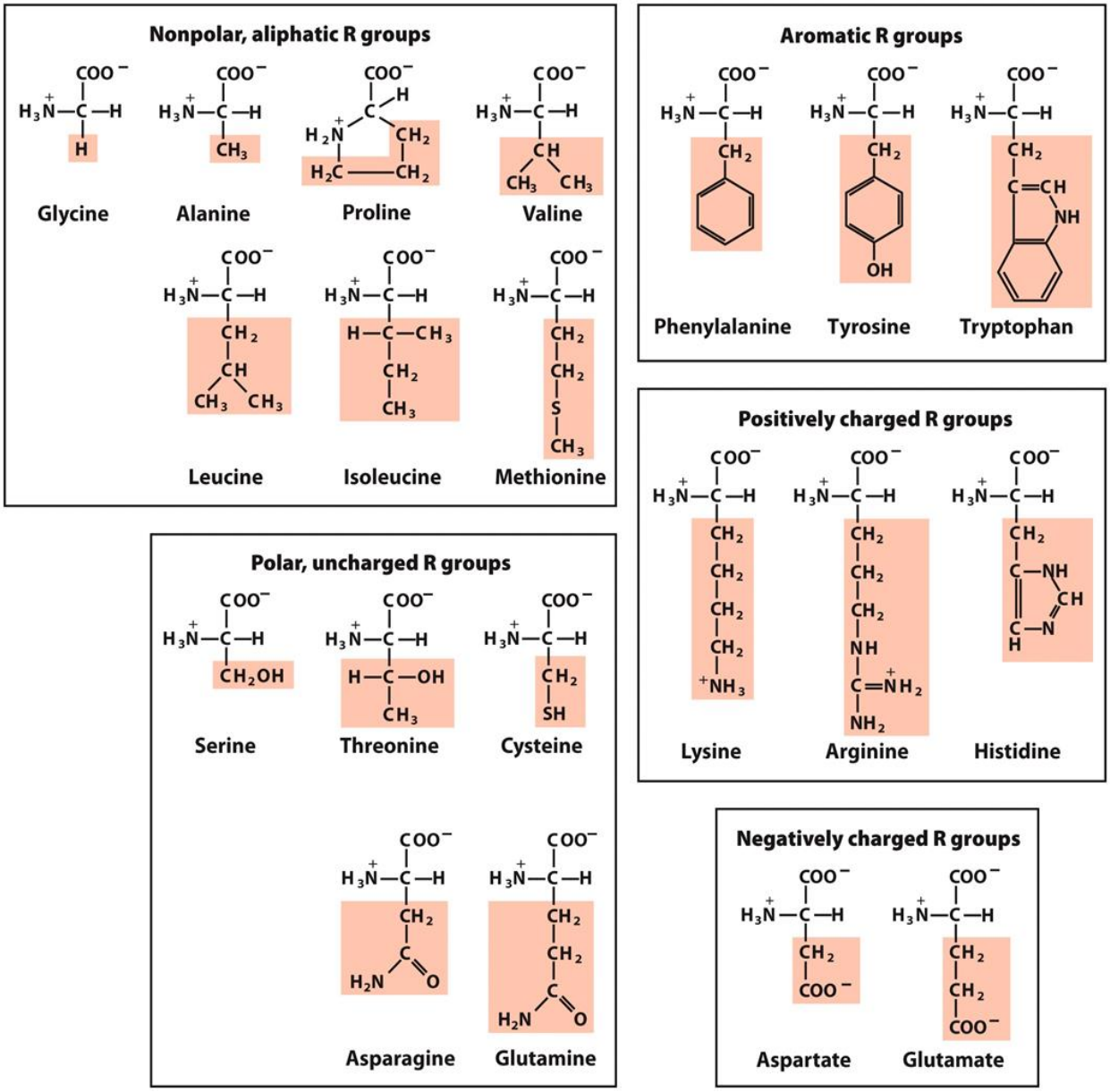

Figure 3 – The 20 standard amino acids of proteins. The unshaded portions are those common to all the amino acids; the portions shaded in pink are the R group (Lehninger, Nelson e Cox, 2005).

The six atoms of the peptide group lie in a single plane, with the oxygen atom of the carboxyl group and the hydrogen atom of the amide nitrogen *trans* to each other. The peptide bonds are unable to rotate freely because of their partial double-bond character. Rotation is permitted about the N-C_α and the C_α-C bonds. By convention, the angle bond for the first (N-C_α) is named ϕ (phi) and ψ (psi) for the C_α–C (Figure 4).

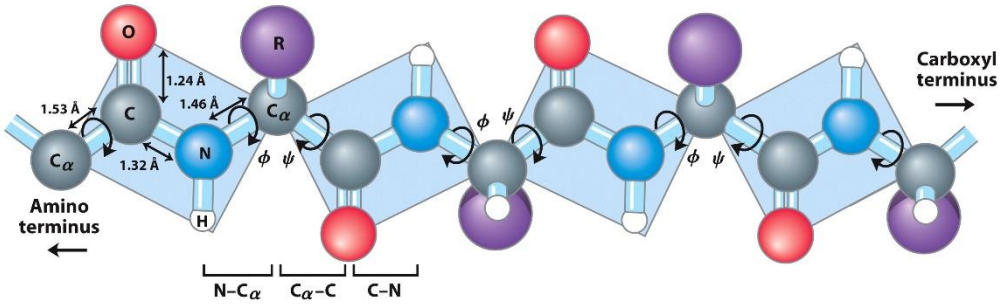

Figure 4 – The planar peptide group. The three bonds separate sequential α carbons in a polypeptide chain. The N-C_α and C_α-C bonds can rotate, with bond angles designated ϕ and ψ, respectively (Lehninger, Nelson e Cox, 2005).

1.3 Secondary structure of proteins

The sequence of amino acids linked together covalently into a polypeptide backbone is called primary structure. The primary structure by itself does not say anything about the three-dimensional space (Elliott e Elliott, 2009).

The polypeptide backbone is arranged in a conformation known as the secondary structure, and this form organizes in a tertiary structure. The molecule formed by the primary, secondary and tertiary structures may be the final functional protein or can be a protein monomer or a subunit. This association is called the quaternary structure (Elliott e Elliott, 2009). However, several types of noncovalent bonds are critical in maintaining the three-dimensional structures of large molecules such as proteins and nucleic acids. Covalent and noncovalent bonds are responsible for the structure.

The simplest arrangement of the polypeptide chain can assume with its rigid peptide bonds a helical structure, which is called the alpha helix (Figure 5). In this structure, the polypeptide backbone is highly wrapped around an imaginary axis drawn longitudinally through the middle of the helix. The R groups of the amino acid residues project to outside from the helical backbone and the carboxyl groups point in the direction of the axis of the helix. These groups are linked to the amino groups by hydrogen bond, generating a maximum bond strength and making the helix a very stable structure. The cross section of an alpha helix shows a virtual cylinder with all the R groups projecting to the outside (Lehninger *et al.*, 2005; Elliott e Elliott, 2009). The left-handed one is more stable than the right-handed helix, and the number of amino acids per turn is 3.6 units that can be proven by X-ray (Figure 5).

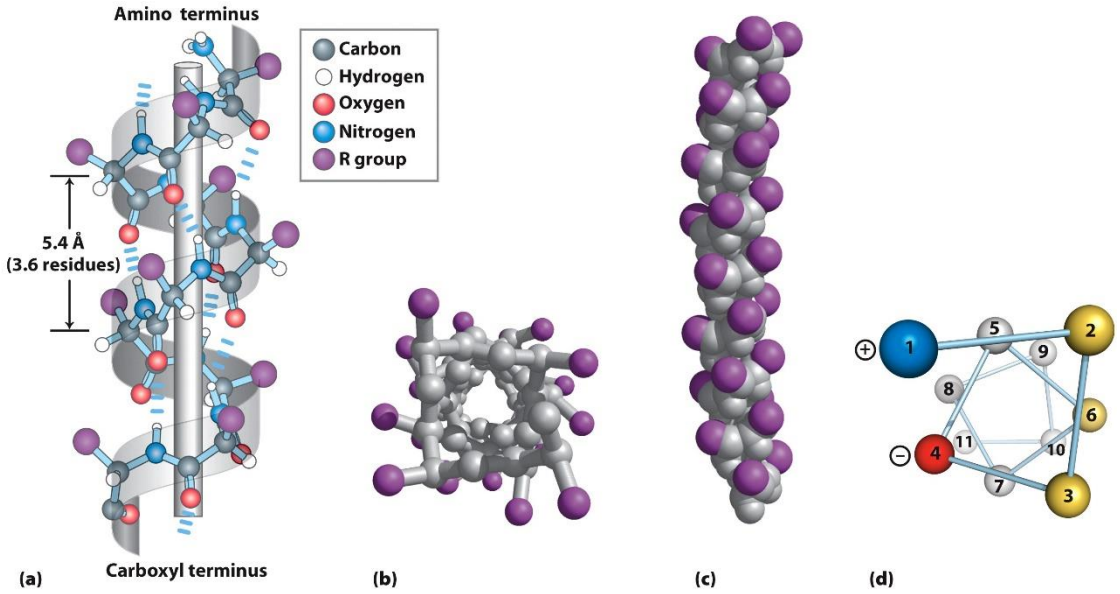

Figure 5 – The four models of the α helix, in different aspects of its structure. (a) The planes of the rigid peptide bonds are parallel to a long axis of the helix. (b) The α helix as viewed from one end, looking down the longitudinal axis, with the R groups. (c) The model shows the atoms in the center of the α helix are very close. (d) The α helix as viewed from one end, looking down the longitudinal axis, with the numeration of carbon alpha (Lehninger, Nelson e Cox, 2005).

Another secondary structure is the β sheet (Figure 6). The polypeptide chain extends into a zigzag pattern. In this conformation, the carboxyl and amino groups make a hydrogen bond to those of neighborhood chains. Several chains can form a polypeptide sheet, and it is gauffered because successive α-carbon atoms of the amino acids residues lie slightly above and below the plane of the β sheet alternately (Lehninger *et al.*, 2005; Elliott e Elliott, 2009). The close polypeptide chains bonded together can run in the same direction (parallel) or opposite directions (antiparallel) (Figure 6).

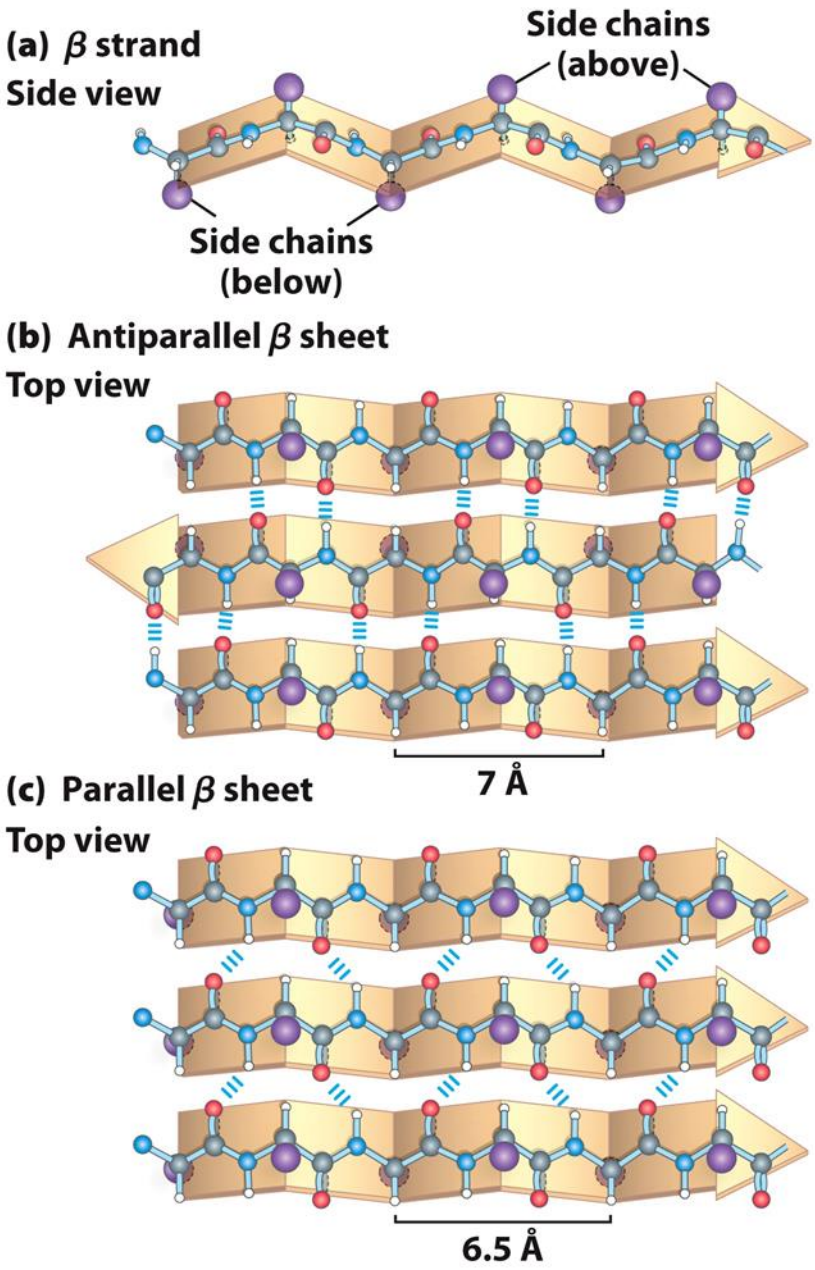

Figure 6 – The β conformation of polypeptide chain. (a) side view of β sheet. **(b)** antiparallel β sheet. **(c)** parallel β sheet (Lehninger, Nelson e Cox, 2005).

Proteins with alpha helices and β sheets section are connected by an unstructured polypeptide sometimes known as random coils or connecting loops (Elliott e Elliott, 2009).

Some amino acids are better accommodated than other in the different types of secondary structures. Some biases are the presence of proline and glycine residues in β conformation and their privation in an alpha helix (Figure 7) (Lehninger *et al.*, 2005).

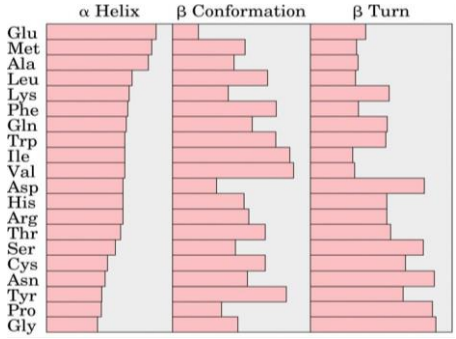

Figure 7 – Relative probabilities that a given amino acid will occur in the three types of secondary structure (Lehninger, Nelson e Cox, 2005).

The bond angles ϕ and ψ can have any value between -180º and +180º, but many values are prohibited by steric interference between atoms in the polypeptide backbone and amino acids side chains. Allowed values for ϕ and ψ are graphically represented in a Ramachandran plot (Figure 8). In this, alpha helix and β conformation fall within a restricted range of sterically allowed structure. Most values of these angles of known protein structures fall into the expected regions, near the predicted values (Lehninger *et al.*, 2005).

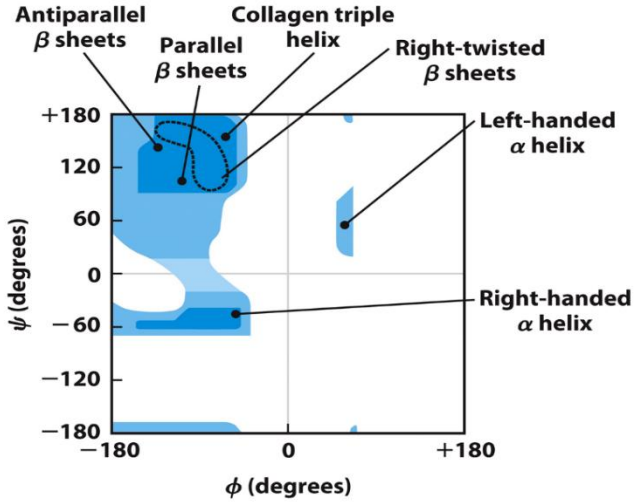

Figure 8 – Ramachandran plots for a variety of structures (Lehninger, Nelson e Cox, 2005).

Kabsch and Sander (1983) defined a minimal helix as the two consecutive n-turns, in others words, a single hydrogen bond of type (*i*, *i*+*n*) between the residue *i* if there is an H-bond from CO (*i*) to NH (*i*+*n*).

The alpha helix characterized by an (*i*, *i*+*4*) pattern, the 3_{10} and the π helix by repeating (*i*, *i*+*3*) and (*i*, *i*+*5*) hydrogens bonds, respectively (Kabsch e Sander, 1983) (Fodje e Al-Karadaghi, 2002).

1.4 Proteins secondary structure prediction

Direct prediction of protein structure from sequence is a challenging problem, and it is a major step towards elucidating its three-dimensional structure, as well as its function. Proteins have covalent and noncovalent forces that help to form the architecture of the three-dimensional structure. Those structures for most proteins are determined by their one-dimensional sequences of amino acid residues. How to accurately predict three-dimensional structures from one-dimensional sequences has been an unsolved problem for the last half century. The problem lies in the challenge of developing an efficient technique to search for an astronomically large conformational space and a highly accurate energy function to rank and guide the conformational search, both of which are not yet available (Heffern*an et al*., 2015).

Today, the methods and tools to predict the secondary structure of proteins are enormous, but the accuracy for large proteins remains a challenge.

The Critical Assessment of Structure Prediction (CASP) is an organization that leads community-wide experiments to measure the state-of-the-art in the modeling of protein structures from amino acid sequences. The objective is to test these methods via the process of blind prediction. CASP happens every two years, since 1994. During these years, much online and off-line software and databases of models were created and are available for all. The protein prediction accuracy was largely improved through a combination of refined methods (Moul*t et al*., 2014).

There are two computational approaches to protein three-dimensional structural modeling and prediction: methods independent of the mold structures (also called template free methods) and that include *ab initio* and *de novo* prediction; and methods based on template structures (also called template based) which include threading and comparative modeling (Xiong, 2006; Verli, 2014).

The *ab initio* approach is a simulation based method and predicts structures based on physicochemical principles governing protein folding without the use of structural templates. The *de novo* methods are those which use some structural information such as protein fragments, secondary structure prediction, and statistical potentials, from non-homologous protein to the target sequence (Xiong, 2006; Verli, 2014).

The principle behind comparative modeling is that if two proteins share enough high sequence similarity, they are likely to have very similar three-dimensional structures. If one of the protein sequences has a known structure, then the structure can be copied to the unknown protein with a high degree of confidence. Homology modeling produces an all-atom model based on alignment with template proteins. The homology-based methods do not depend only on the statistics of residues of a single sequence but on common secondary structural patterns conserved among multiple homologous sequences (Xiong, 2006; Verli, 2014).

Fragment-based approaches are a widely used *de novo* method in the prediction of secondary structures of proteins. The technique used for selecting the fragment is made by the similarity with the target sequence. The database needs to have a high number of fragments to increase the probability of choosing a correct fragment. A major problem for all fragment-based *de novo* approaches occurs when the fragment library for a given target does not contain good fragments for a particular region. In that case, low accuracy models will be generated regardless of the precision of the potentials being used and, in any situation, of the amount of computation time invested in the modeling routine (De Oliveira *et al.*, 2015).

The Rosetta method uses small protein sequences with known structure. The compact design of the generated protein is assembled by a random combination of these fragments, using the Monte Carlo simulated annealing search. The structures are built using only nine residue fragments (9mers) (Simon*s et al.*, 1997). Rosetta method uses the torsion space of the skeleton of the protein, angles ϕ, ψ and ω. The standard measure of this similarity is backbone - carbon root-mean-squared deviation or CαRMSD (Holmes e Tsai, 2004; Rohl *et al.*, 2004; De Oliveira *et al.*, 2015).

Another method, it is the FRAGFOLD which is based on the assembly of super-secondary structural fragments taken from highly resolved protein structures using a simulated annealing algorithm (Jones e Mcguffin, 2003). The FRazor method (L*i et al*., 2008) introduced other structural information items, such as secondary structure, solvent accessibility, and contact capacity. The HHFRAG (Kalev e Habeck, 2011) differs by selecting fragments by generative models of local protein structure such as hidden Markov models. This approach does not have restriction caused by the size and diversity of a structure database, and it is possible to assess the probability of each fragment.

Another algorithm, QUARK, for *ab initio* protein structure prediction, selects features by a neural network. The full-length structure models are assembled from fragments using replica-exchange Monte Carlo simulations, which are guided by a composite knowledge-based force field (Xu e Zhang, 2012).

Tools to predict the secondary structure alpha helix in proteins are numerous today (Koehler Leman *et al.*, 2015). The majority use the support vector machine and neural networks (Heffernan et al., 2015). It is also possible to use other proteins features to predict the secondary structure. An integrative tool can be very handy to comprehend the prediction and to understand the sequence analysis better. We summarized a few tools and methods available to predict secondary structure (for further details, see Appendix A). A summary of some tools which are usually used for prediction of membrane proteins is in Koehler Leman, Ulmschneider and Gray (Koehler Leman *et al.*, 2015).

1.5 Logistic Regression

The regression method is helpful to any data analysis concerned with describing the relationship between a response variable and one or more explanatory variables. The main focus of logistic regression analysis is a classification of individuals in different groups (Cokluk, 2010).

The term logistic regression analysis comes from logit transformation, which is applied to the dependent variable. This case, at the same time, causes certain differences both in estimation and interpretation. Simple and multiple linear regression analysis are used to evaluate the mathematical correlation between dependent variables and independent variable(s) (Cokluk, 2010).

A standard regression equation consists of true values of a few independent variables and weights produced by the model to predict the value of the dependent variable. The dependent variable is the predicted variable, while the independent variables are constants or categorical (Cokluk, 2010).

In the logistic regression, the estimated value ranges from 0 to 1. Therefore, the logistic regression shows the possibility of appropriate consequences for each subject. The analysis produces a regression equation, which enables us to make an accurate estimation for the possibility

that an individual falls into one of the categories. (Cokluk, 2010) The outcome variable in logistic regression is *binary* or *dichotomous* (Hosmer e Lemeshow, 2000).

Logistic regression is similar to both multiple regression and discriminant analysis. The main difference between the multiple linear regression analysis and logistic analysis is that the value of the dependent variable is estimated in the first analysis, while the possibility of occurrence of one of the values that the dependent variable might have, is estimated in logistic regression analysis. Moreover, the discriminate analysis aims to explain and to predict group membership using a group of independent variables. Logistic regression analysis, unlike discriminant analysis and multiple regression analysis, does not require to meet assumptions concerning the distribution of independent variables, like the normal distribution of independent variables, linearity and equality of variance-covariance matrix do not have to be met (Cokluk, 2010).

The quantity is the mean value of the outcome variable, given the value of the independent value. This quantity is called the conditional mean and will be expressed as:

$$E(Y \mid x) \tag{1}$$

where *Y* denotes the outcome variable and *x* is the independent variable. The quantity *E(Y | x)* is the expected value of Y, given the value x. In multiple regression, it assumes that this mean is expressed as an equation in x, such as:

$$E(Y|x) = \alpha_{n+1} + \sum_i \alpha_i x_i, \tag{2}$$

where x can be extent between $-\infty\ to + \infty$. With dichotomous data, the conditional mean must be: $0 \leq E(Y|x) \leq 1$.

Inspired by the diversity of methods, the present dissertation proposes a combination of the logistic regression method and sliding window technique for prediction of the standard type alpha helix. In our search for related papers in the literature, we could not find the conjugate use of this methodology for the purpose in question. In this dissertation, we propose a new method that contributes to the way of predicting alpha helix of proteins, with performance measures greater than 70%. The following flow chart (Figure 9) represents a simplified background of the problem in structural bioinformatics area and which, consequently, originated the motivation of carrying out this work.

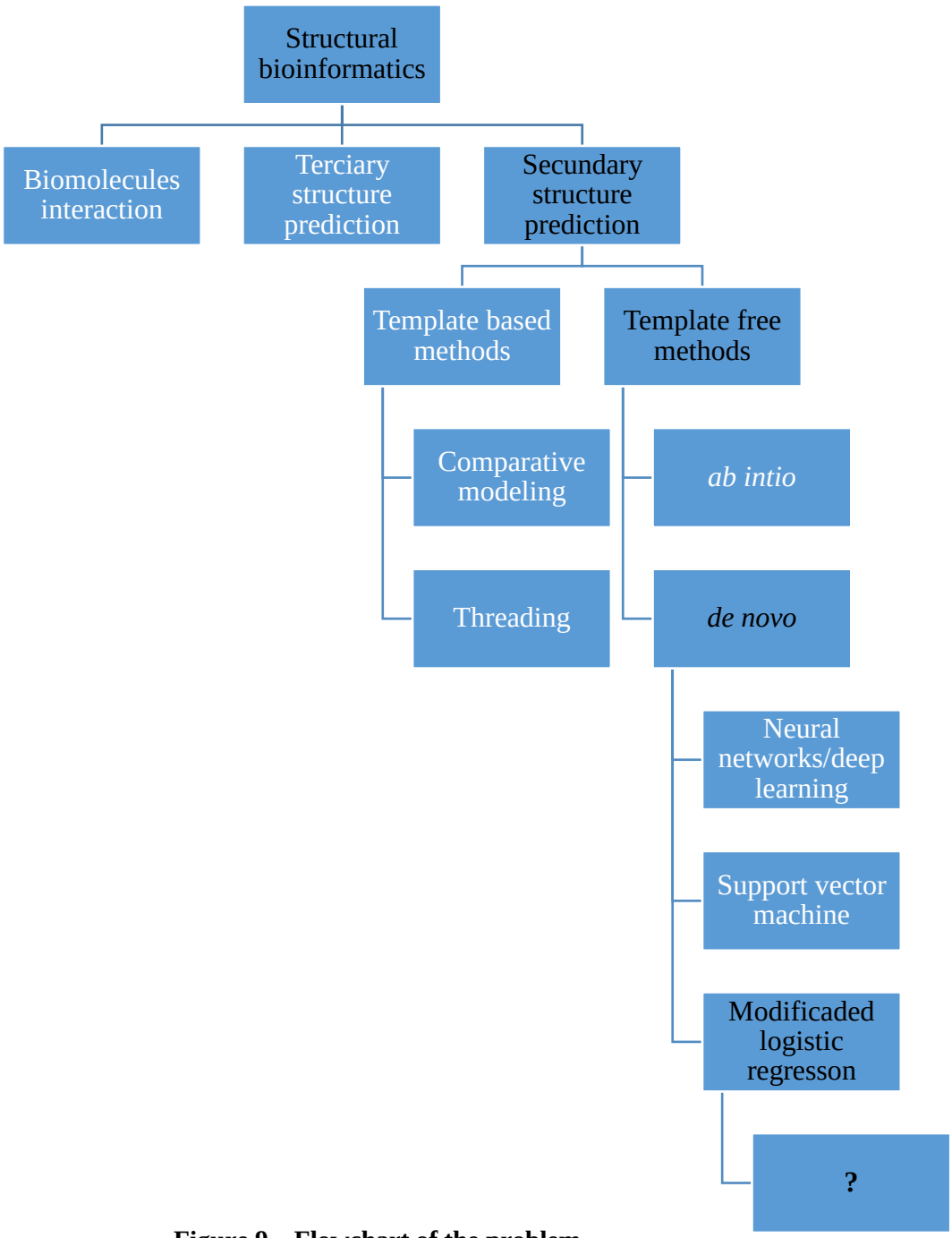

Figure 9 – Flowchart of the problem

2 Objectives

2.1. General Objective

2.1.1. To predict alpha helices in fragments with 9 residues, using modified logistic regression method.

2.2. Specifics Objectives

2.2.1. To search the PDB database for proteins with low identity to each other;

2.2.2. To use the sliding window technique to generate fragments of sequence with nine residues;

2.2.3. To apply the modified logistic regression method for the construction of alpha helices prediction model;

2.2.4. To use accuracy, specificity, and sensitivity metrics to evaluate the proposed model;

2.2.5. To use the MatLab environmental engineering to build and run the model of alpha helix prediction.

3 Methods

3.1 Database preparation

The model was built using available protein structures from the Protein Data Bank (PDB) (Berma*n et al*., 2000). The secondary structure class was obtained from the secondary structure assignment as provided by the author of a PDB entry. To achieve a meaningful representation, a non-redundant set of proteins from the PDB, we employed the PISCES tool (Wang e Dunbrack, 2003). The following features were used: sequence identity percentage less than or equal to 5; Resolution of 0.2 ~ 2.0; R-factor of 0.2; Sequence length of 40 ~ 10000; excluded Non X-ray entries and CA-only entries. The selection was made by entries, and the length was the shortest allowed by the program. Those parameters were chosen to guarantee a better resolution and diversity.

With the PISCES software outcomes, we created a subset of proteins with low percentage sequence identity, high resolution, and desirable length cutoff. The sequences were provided by the Research Collaboratory for Structural Bioinformatics (RCSB), and the R-value data were also obtained from the Uniformity Project files. Some of the missing values were achieved from the PDB-FINDER database. To estimate sequence identity at longer evolutionary distances, PISCES uses the PSI-BLAST to calculate these identities, which were used locally to build a position-specific scoring matrix (PSSM) or profile from homologous sequences in National Center for Biotechnology Information's non-redundant protein sequence database. Three iterations were performed for each *query*, with an E-value cutoff of 0.0001 for inclusion in the profile (Wang e Dunbrack, 2003).

3.2 The matrix

In MatLab R2015b software, we wrote a script to construct the matrix that was used by the modified logistic regression method. All fragments, size nine, were generated by the sliding window technique through the sequence of the first chain and were placed in columns. They were considered individuals. The rows were all representations of the combinations of triplets residues (20 ^ 3 = 8000). The number three in the sliding window to create tripeptides with all possible combinations was chosen to generate a reasonable and workable number of residues combinations

in our computers. We also considerate the work of our research group, which demonstrated that this size of sliding window generates better results (Couto *et al.*, 2007).

We also perform the reverse sliding window to decrease the sparsity of the matrix. According to Deerwester et al. (1990), the best performance of any IR (irrational) system are problems with a reasonable size using a rich and high-dimensional representation. That is why we incremented with the reverse sliding windows.

To apply the modified logistic regression model, we have to concatenate one line to the matrix, with the value of probability: the 0 value if the fragment was not 100% alpha helix default, and 1 if the fragment was 100%. MatLab considers only the secondary structure annotation done by 'Author Secondary Structure'.

3.3 Singular Value Decomposition

Presently, every area in society have an enormous amount of data stored in their databases, and it is a challenge to extract useful information. A technique for information retrieval, using linear algebra techniques, is the singular value decomposition (SVD). When the SVD is utilized in a matrix, it allows the matrix to be represented by a set of derived matrices, which can have different representations of data without loss in semantic meaning. In other words, it is possible to compute an approximate basis for this space using representatives subspaces (Berry *et al.*, 1995; Élden, 2006) (Berry *et al.*, 1995).

A matrix using the SVD can be represented as:

$$A = U\Sigma V^T, \quad (3)$$

where **A** is a matrix of real numbers or complex numbers composed of *m* rows by *n* columns. The **U** is an orthonormal *m* x *m* matrix and the eigenvectors of AA^T; the Σ is an *m x n* matrix, known as the diagonal matrix, with real and non-negative numbers and contain the singular values of *A*. The matrix V^T is known as a conjugate transpose, an *n* x *n* unit matrix with real or complex numbers. As the diagonal values of Σ are ordered in descending order, Σ is a direct function of matrix A and characterizes the singular values of this matrix, ordering them from the most significant to the least

significant values. Considering a subset of singular values of size k<n, we can obtain A_k, that is an approximate of matrix A:

$$A_k = U_k \Sigma_k V_k^T \quad (4)$$

Thus, data approximation depends on how many singular values are used, with this k-dimensional (Élden, 2006; Kumar *et al.*, 2011; Santos *et al.*, 2011).

The possibility of extracting information based on less data is part of the reason for this technique's success, as it allows data analysis, with an execution time that does not increase exponentially with increasing matrix size. A data set represented by a smaller number of singular values than the original full-size dataset has a tendency to group data items that would not be grouped if we used the original one. This strategy could explain why clusters derived from SVD can expose non-trivial relationships among the original data set items. This derived representation, which captures associations, is used for retrieval (Berry *et al.*, 1995).

The meaning representation in the reduced space representation is economical, in the sense that N original index features have been replaced by the k < N best surrogates by which they can be approximated. It is essential for the method that the derived k-dimensional factor space does not reconstruct the original term space perfectly. Our aim with this technique is to be able to represent features and fragments, in a way that escapes the unreliability, ambiguity, and redundancy of features. It is also important because it allowed working arrays on our computers. The beauty of an SVD, however, is that it allows a simple strategy for optimal approximate fit using smaller matrices. How to choose the appropriate number of dimensions is an open research issue (Deerwester *et al.*, 1990). Everitt and Dunn (1991) proposed an alternative approach where singular values whose relative variance is less than 0.7/n, where n is the number of proteins in the document-term matrix, must be ignored. If the singular values in S, are ordered by size, the first and largest k may be kept and the remaining smaller ones set to zero (Deerwester *et al.*, 1990).

It is important to note that for a square, symmetric matrix X, singular value decomposition is equivalent to diagonalization, or solution of the eigenvalue problem (Wall *et al.*, 2003).

3.4 Clustering

After setting up the matrix with the logistic regression values, we generated clusters using the k-medoid algorithm. This method minimizes the average dissimilarity of all objects of the data set the nearest centroid (Kaufman *et al.*, 1987). The k-medoid method is based on average dissimilarities instead of the sum of squares of dissimilarity of the objects to the representative objects they are assigned to (Kaufman e Rousseeuw, 1987).

Kaufman and Rousseeuw (1990) said that the main objective to find k clusters at medoid method is to show a high degree of similarity between them whereas objects belonging to different clusters are as diverse as possible. All the k objects should represent the various aspects of the structure of the data. The k must be chosen for the location in such a way that the sum of distances from all the objects of the data set to the nearest of these is as small as possible. The spot of the center is interpreted as the selection of the representative object. The centroid does not have to be one of the objects in the original data set, and it cannot be defined when the data is a set of dissimilarities not based on interval scaled measurement values (Kaufman e Rousseeuw, 1987).

The medoids method uses the PAM (Partition Around Medoids) algorithm, which divides the data set into k clusters, where the integer k needs to be specified by the user. The algorithm proceeds through two phases. In the first phase, a representative set of k objects is found. The first object selected has the shortest distance to all other objects, which it is the center. An addition k-1 objects are selected one at a time in such a manner that distance decrease as much as possible (Struyf *et al.*, 1997).

In the second phase, possible alternatives to the k objects selected in phase one, are considered iteratively. At each step, the algorithm considers all pairs of objects and make the swap (if any) which decreases the objective function the most. These iterations continue until convergence (Struyf *et al.*, 1997). Normally, the number of clusters is not defined in a data set and fluctuates for each case (Kaufman, L. e Rousseeuw, P. J., 1990).

Two of the most difficult tasks in cluster analysis are deciding on the appropriate number of clusters and deciding how to tell a bad cluster from a good one. Kaufman and Rousseeuw (1990) define a set of values called silhouettes that provide key information about both tasks. One way to

selecting a value of k is using the silhouette coefficient (Kaufman e Rousseeuw, 1987; Rousseeuw, 1987) (Rousseeuw, 1987).

The goal of k-means is to minimize a sum of squared Euclidean distances, implicitly assuming that each cluster has a spherical normal distribution. The k-medoids is robust because it minimizes a sum of unsquared dissimilarities. Moreover, PAM does not need initial guesses for the cluster centers, contrary to k-means (Struyf *et al.*, 1997). We did not use k-means because it is sensitive to the selection of the initial partition and may converge to a local minimum or create an empty group (Macqueen, 1967).

3.5 Modified Logistic Regression Model

For the logistic regression the maximum likelihood is used, a general method of estimation that conducts to the sum of least squares. This method produces values for the unknown parameters, which maximize the probability of obtaining the observed set of data. First, it has to construct the function, the likelihood function, which expresses the probability of the observed data as a function of the unknown parameters (Hosmer e Lemeshow, 2000).

In the logistic regression, the quantity $P = E(Y|x)$ (1) is used to represent the conditional mean of Y given x when the logistic regression model we use is:

$$P(x) = \frac{e^{\sum_i \alpha_i x_i + \alpha_{n+1}}}{1 - e^{\sum_i \alpha_i x_i + \alpha_{n+1}}} \quad (5)$$

The data obtained from PDB database is represented by matrix A, with m rows and n columns, with rows representing fragments and columns representing triplets. The value of each position $x_{m,n}$ represents the triplets of a fragment. We will omit the indication of row *m* in the elements of vector x. That is $x = \{x_1, x_2, \ldots, x_m\}$ every time row *m* to which x refers to is evident in the context. Associated with each row *m* is P*i*(x) = 0/1 that informs the secondary structure of the fragment (100% alpha helix/ 100% non-alpha helix).

The logistic regression consists of finding a vector $\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_n)$ to fit the set of the equation (4). We observed that when $e^{\sum_i \alpha_i x_i + \alpha_{n+1}}$ drops to zero, Pi (x) also goes to zero. On the

other hand, if $e^{\sum_i \alpha_i x_i + \alpha_{n+1}}$ tends to infinity, Pi(x) approximates one. Viewing Pi (x) as the probability, the odds Ci(x) are given by:

$$C_i(\mathrm{x}) = \frac{P(x)}{1-P(x)} = \frac{\frac{e^{\sum_i \alpha_i x_i + \alpha_{n+1}}}{1 - e^{\sum_i \alpha_i x_i + \alpha_{n+1}}}}{1 - \frac{e^{\sum_i \alpha_i x_i + \alpha_{n+1}}}{1 - e^{\sum_i \alpha_i x_i + \alpha_{n+1}}}} = e^{\sum_i \alpha_i x_i + \alpha_{n+1}} \quad (6)$$

To implement the method, we use $\hat{C}i$ (x) ≈ Ci (x) = (0.99999 / (1-0.99999)) instead of Ci (x) when the odds are related to Pi (x) = 1. When Pi (x) = 0, we consider $\hat{C}i$ (x) ≈ Ci (x)= (0.00001 / (1-0.00001)).

Doing the *logit transformation* in equation (6), as:

$$C_i(\mathrm{x}) = Odds(\mathrm{x}) = \ln\left[\frac{P(x)}{1-P(x)}\right] = \ln\left[e^{\sum_i \alpha_i x_i + \alpha_{n+1}}\right] = \sum_i \alpha_i x_i + \alpha_{n+1}, \quad (7)$$

a linear algebraic model is created to determine α:

$$\mathrm{bi} = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_n x_n \quad (8)$$

This *logit transformation* can have the desirable properties of the linear regression model, like the parameters that may be continuous and may range from $-\infty\ to + \infty$ (Hosmer e Lemeshow, 2000).

For i = 1, 2, …, m, let $\bar{e} = (1, \ldots, 1)^T$ be a vector of m ones and $b = [b1, b2, \ldots, bm]^T$. The system of linear equations (8) may be represented by:

$$B\alpha = b, \text{ with } B = [\bar{e}\ A] \quad (9)$$

The system (7) has an infinite number of solutions, since n + 1 >> m. It is usual to circumvent this difficulty by pruning the model and keeping only a small subset of n fragments. This procedure resembles the features selection in data mining. We propose the usage of a stabilizing term in the logistic regression model found in the works of Linnik (1961) and Golub (1965) and later by Abreu *et al.* (2008) and Menard (2010). It allows the assignment of values to α parameters by minimizing the square sum of the residuals (equation 9) summed to the squares of α, thus letting the system to have a unique solution. Therefore, to assign a solution to (equation 7), we are considering that, minimize f (α):

$$f(\alpha) = \alpha^T\alpha + (B\alpha - b)^T * (B\alpha - b), \quad (10)$$

As f (α) is a convex function, the argument α* that minimizes (equation 10) is given by the derivative of f (α) in α and making it equals zero. This results in the following system of linear equations:

$$(I + B^TB)\ \alpha = B^Tb, \quad (11)$$

where I is an identity matrix of dimension *n*. One should note that the identity matrix does not allow the rank to become deficient. The optimal solution α* of (10) is obtained by the solution of (9) and is unique. So, given a *query* q = [q1, q2, …, qm] with the fragment with nine residues levels of expression of n triplets, the probability of q to be 100% alpha helix or 100% non-alpha helix is given by:

$$P(q) = g(q) / (1 + g(q)), \quad (12)$$

where $g(q) = \exp([1\ q^T]\ \alpha)$.

Suppose that the sample of *n* independent observations of the pair $(x_i, y_i), i = 1,2, \dots, n,$ where y_i is the value of a dichotomous outcome variable and x_i is the independent variable for the i^{th} subject. In the linear regression, the method used often for estimating unknown parameters, β_0 and β_1, is least squares, which minimize the sum of square derivations of the observed values of *Y* from the predicted values based upon the model. Unfortunately, in the method of least squares with a dichotomous outcome, the estimators have no longer the same properties (Hosmer e Lemeshow, 2000).

3.6 Project and calculate the odds of an unknown *query*

Firstly, it was necessary to project the query utilizing the *eigenvalue* of U, of the singular value decomposition (equation 2), being k=30:

$$U(:,1:30)^T * U(:,1:30) * \tilde{q} = U(:,1:30) * q \Leftrightarrow U(:,1:30)^T * q \quad (13)$$

The equation 13, a linear equation, which minimizes the sum of squared residuals, gives us the query in the dimension and the pattern of the matrix.

Now that we had a query with 30 dimensions, we calculated the distance between it and the medoids of the matrix A1. The medoid, which has the shortest distance, was selected and multiplied by the query. With that, the size was the same (8000x1) and we multiplied by the alpha value and calculated the value of probability. We used the equation 13 in this step.

After the matrix was constructed, it was possible to assemble several models. Initially, we thought four methods for building models: small and near clusters method (conventional and reverse sliding window technique), medoids method, direct distance method and residue analysis method. These methods were designed not only for the protein secondary structure problem but, also, with the perspective of data mining problems. The amount of information and scalability of a model is a relationship that we wanted to build. The direct distance method is similar to the search engines, like Google (Cilibrasi e Vitanyi, 2007). It would use the hierarchical clustering and a distance parameter to define the secondary structure of the query. The search is done through tree using a parameter distance. The residue analysis method would analyze all the residues of the fragments generated, in particular, the 5th. According to Pauling (1951), the 1st residue binds to the 4th, in an alpha-helix. Therefore, it would be interesting to study this residue, these positions.

The other two methods will be detailed below. They were chosen to be tested because they are different and have the capability of being scalable to a diversity of applications.

3.6.1 Small and near clusters method

Briefly, we decomposed the matrix and grouped it hierarchically (dendrogram). In the first, the aim was to promote the approach of analogous individuals and minimize the noise. While the second step (clustering with the dendrogram), allowed creating groups with similar individuals in the same cluster.

The degree of granulometry will be low, and the differences between the alpha-helix individuals and not 100% alpha helix will be more significant.

The dendrograms for both matrices were made in Matlab. For the A0 matrix (with fragment 100% non-alpha helix) the following parameters were chosen: Partitioning algorithm – k-medoids; the number of nodes per level - 5; maximum number of individuals per cluster - 5000. For A1

matrix (with fragments 100% alpha helix) have the following parameters: Partitioning algorithm – k-medoids; the number of nodes per level - 5; maximum number of individuals per cluster - 1000. These parameters are summarized in Table 1.

Table 1 - Parameters of the dendrograms.

PARAMETERS / **MATRIX**	**PAM algorithm**	**distance**	**Nº max/cluster**	**Nodes**
A0	k-medoids	cosine	5000	5
A1	k-medoids	cosine	1000	5

Source: Author

To calculate which cluster that the *query* belongs, we verified which medoid in the last level of A1 was closer, by calculating the cosine. We tested all distance types available in MatLab (euclidean, squared euclidean, cityblock, hamming, jaccard, and cosine) and the most appropriate to our problem was cosine distance.

Given the medoid, we selected the cluster it belonged. For the A0 matrix, we calculated the nearest medoid in the last level, related to the selected medoids of A1, and selected the cluster that includes. (Figure 10).

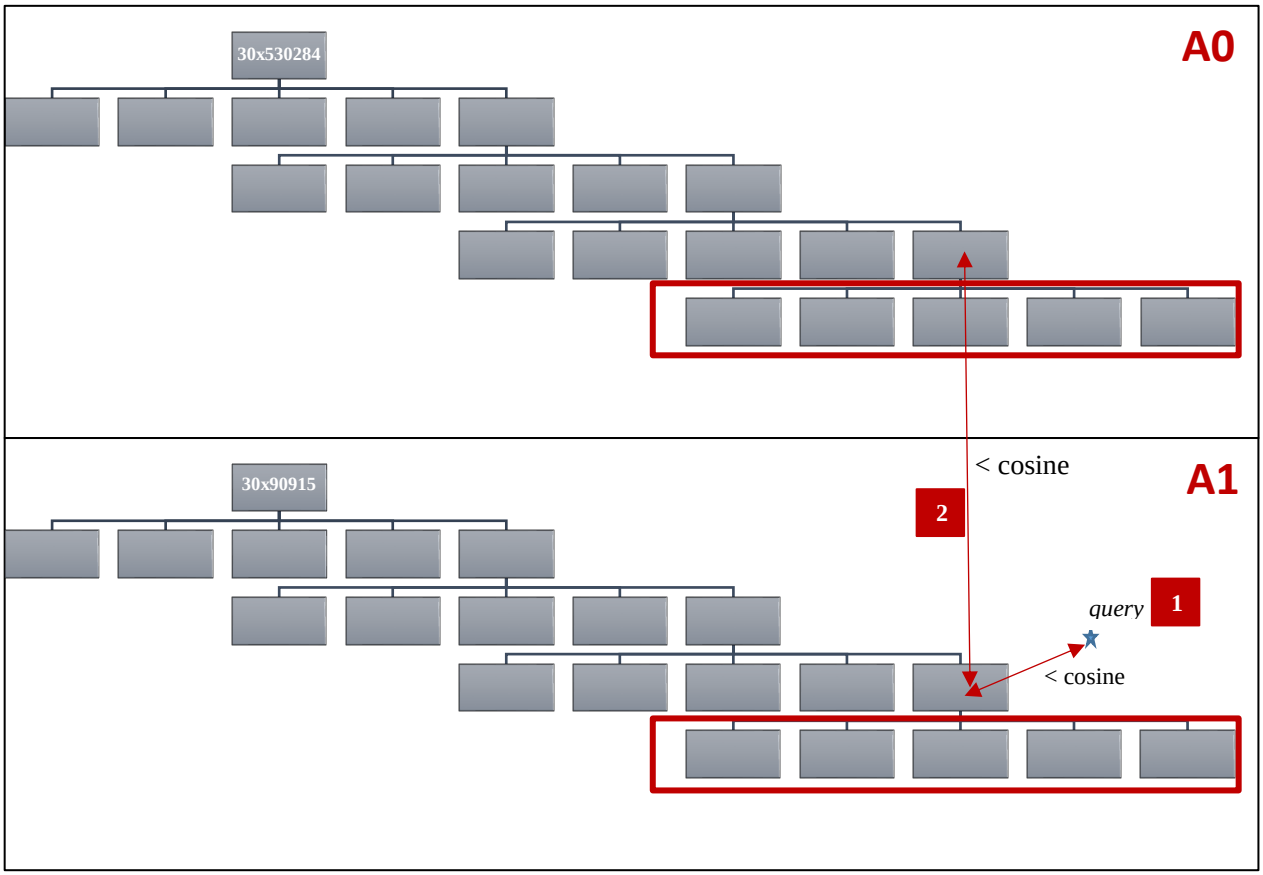

Figure 10 - Didactic scheme of the small and near clusters method: 1 - project the query in A1 and select the medoid closest and the cluster that it belongs; 2 - select the closest medoid of A0 and the cluster that it belongs.

Then, we applied the modified logistic regression method. From the established method, we calculated the probability of the fragment (query) being or not a 100% alpha-helix (default).

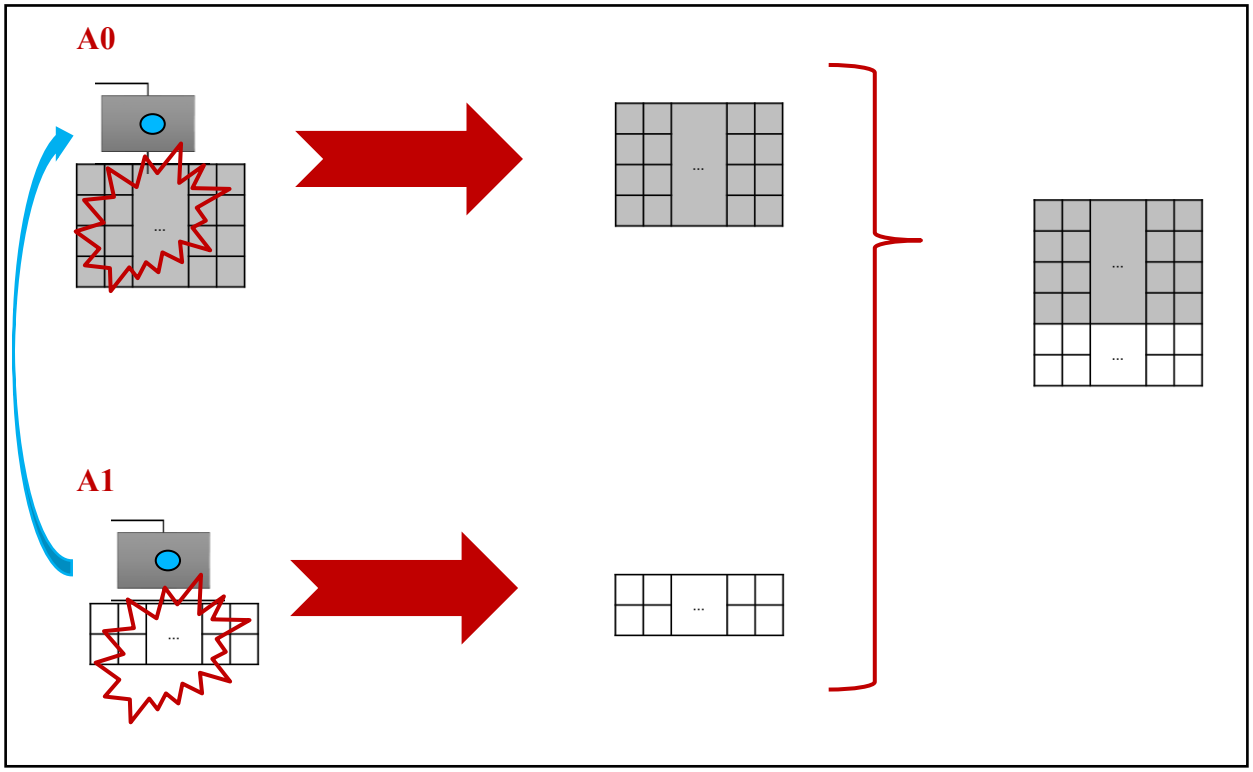

Figure 11 – **Didactic scheme of the Small and near clusters method.** The blue circle is a medoid. After selecting the cluster A1, we selected the nearest cluster in A0. With the two matrices concatenated, it is possible to construct the model.

3.6.2 Medoids method

After clustering both A0 and A1 matrix using the k-medoids algorithm, we obtained a medoids list. Since a medoid is the element that represents each cluster, we extracted these individuals and built a model only with them. The goal was to concatenate all representatives and build a model (Figure 12).

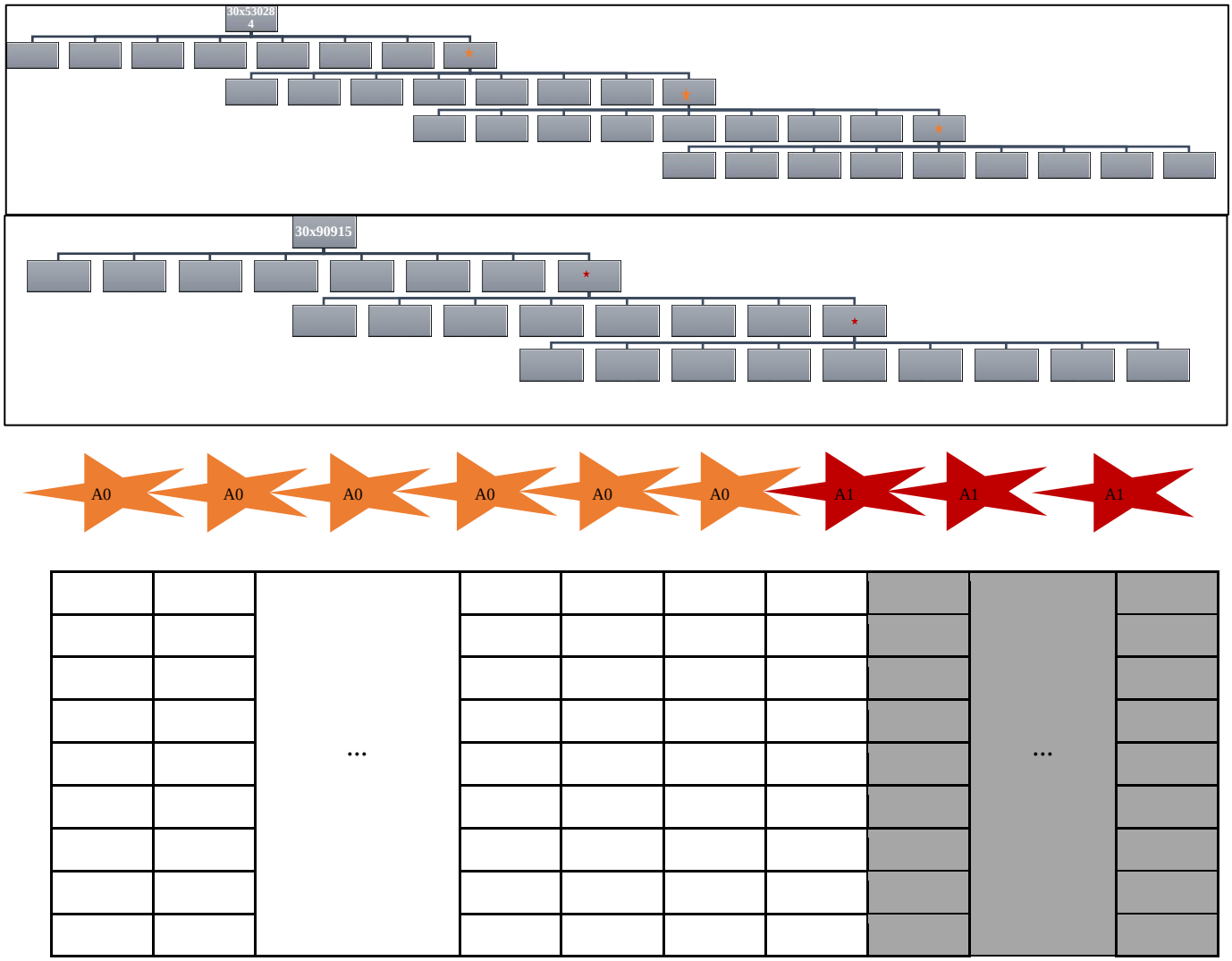

Figure 12 – Didactic scheme of the medoids method.

3.7 Cross-validation and performance Measures

Cross-validation is a popular strategy for algorithm selection, to calculate the risk of new models. The main idea was to split data, once or several times, for estimating the risk of each algorithm: part of the data (the test sample) was used for training each algorithm, and the remaining part (the validation sample) was used for estimating the risk of the algorithm. Then, we calculated the sensitivity and specificity to estimate the risk (Arlot e Celisse, 2010). For each of the K experiments, we used K-1 folds for training and the remaining one for testing.

The advantage of K-Fold Cross-validation is that all the examples in the dataset are alternatively used for both training and testing. It only assumes that the data are identically distributed, and test/validation samples are independent. The design of this strategy is shown in figure 13.

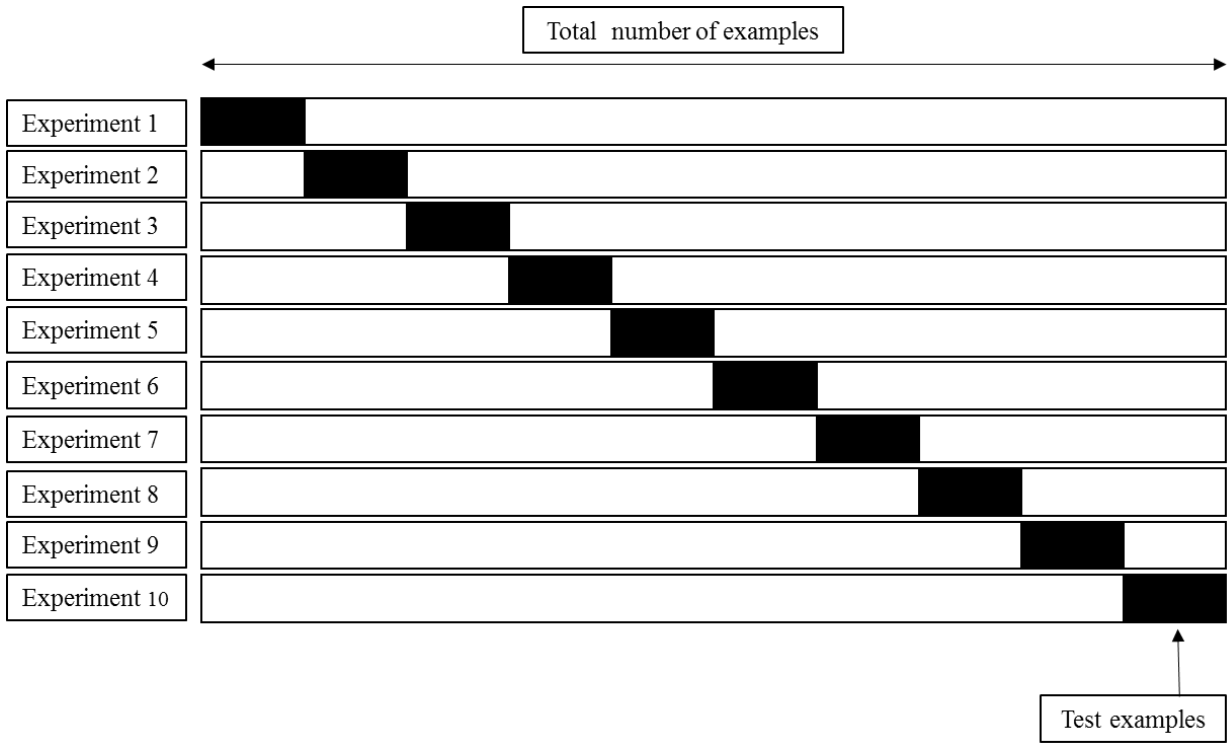

Figure 13 - Didactic scheme of the method K-fold cross validation.

3.7.1 Sensitivity (also known as the true positive rate) and specificity (also known as the true negative rate)

To evaluate the effectiveness of our proposed methodology, we calculated the sensitivity (equation 13), specificity (equation 14) and accuracy (equation 15) (Kurgan e Homaeian, 2006; Dehzangi *et al.*, 2014).

The sensitivity measures the proportion of correct predictions of proteins compared to the whole number of samples, which are classified as correct and is calculated as follows:

$$Sensitivity = (TP/(TP + FN)) \times 100, \quad (13)$$

where TP is the number of correct predictions (true positive), while FN is the number of incorrect predictions (false negative) (Kurgan e Homaeian, 2006; Dehzangi *et al.*, 2014). In our study, true positive meant that it was truly an alpha helix.

The specificity measures the proportion of the number of correct rejected samples compared to the whole number of rejected samples and is calculated as follows:

$$Specificity = (TN /(TN + FP)) \times 100, \quad (14)$$

where TN is the number of correctly rejected (true negative) samples while FP is the number of incorrectly accepted samples (false positive). In our study, true negative meant that, in effect, it was not an alpha helix. These two parameters are associated with the prediction error, which is 100% sensitive and specific, consequently as a perfect predictor (Kurgan e Homaeian, 2006; Dehzangi *et al.*, 2014).

The accuracy is defined as the ratio between the number of correct predictions and *n*, which is the total number of predictions (proteins) (Kurgan e Homaeian, 2006):

$$accuracy = ((TP+TN)/n) \times 100, \quad (15)$$

4 Results and Discussion

4.1 Database preparation

With our strategy, we obtained 3098 entries in July 2015, which were imported into MatLab software for analysis. Inspired by the Rosetta method, the length of the sliding window was nine. When the window slid through the sequence, it generated fragments with nine residues. A fragment was classified as a default alpha helix structure only if 100% of it was in this conformation.

4.2 The matrix

It was generated 621176 fragments. Where 90,915 were associated with the value b0 = 1, and 530,261 were related to b0 = 0. There were no fragments associated with two values of b0. We can consider a sample of PDB due to our chosen strategy and, consequently, this number of fragments generated.

In theory, the number of combinations is 5.12×10^{12} (20 ^ 9), whereas the number of the nature of fragments is much lower due to affinities and preferably contacts.

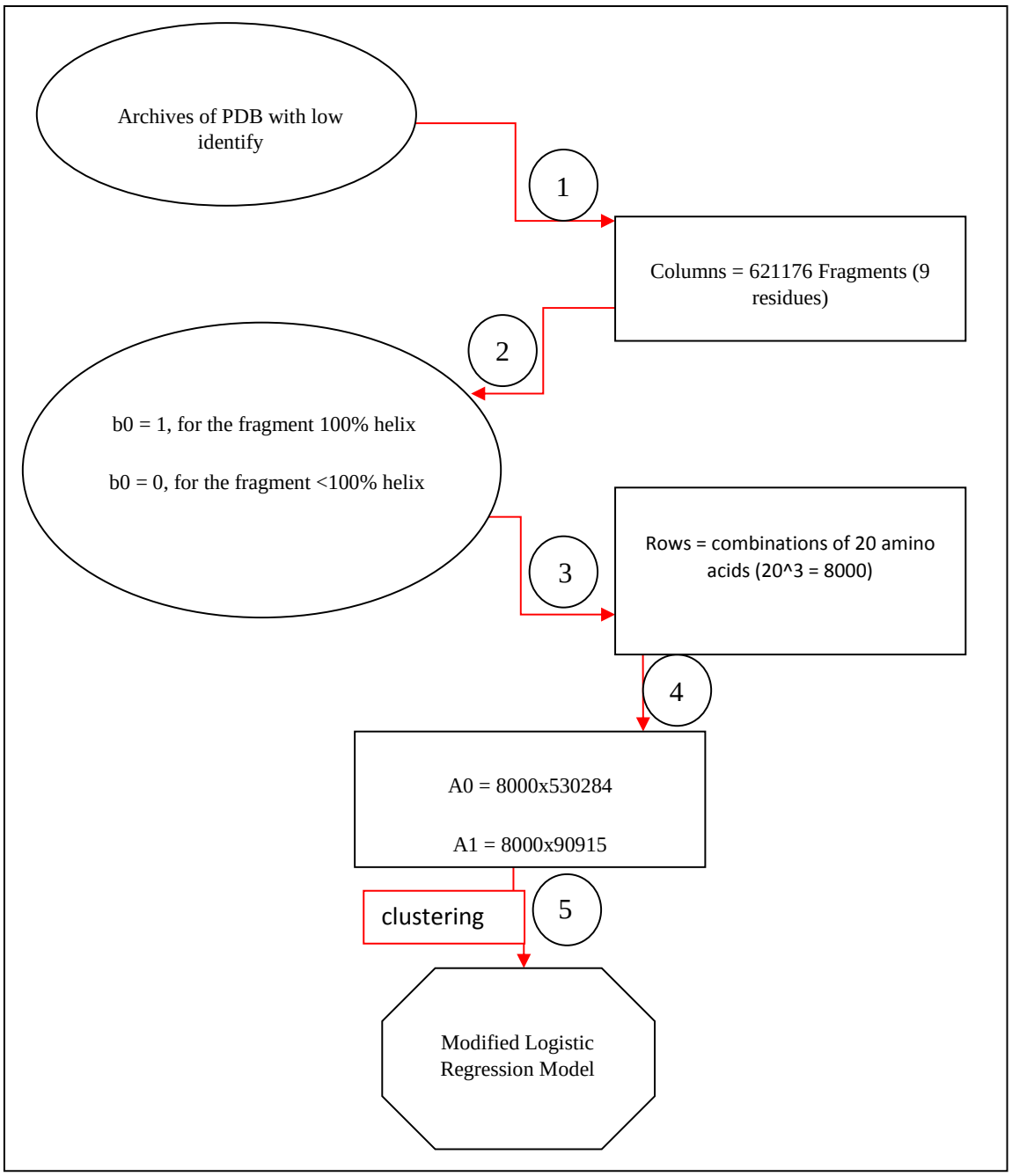

Figure 14 – Flowchart of the creation of the matrix.

4.3 Singular Value Decomposition

To reduce the dimensions designed to the 3D plot we utilized the script done by Marcolino *et al*. (2010), which utilizes SVD recursively.

We plotted the relative S in a decreasing magnitude order and chose 30 singular values, where there were no longer significant variations. The product of the resulting matrices was a matrix X that was only approximately equal to X, and was of rank k. It can be shown that the new matrix X is the matrix of rank k, which is closest to X. The presented graphics are only for the conventional windows. The behavior with the reverse window was similar.

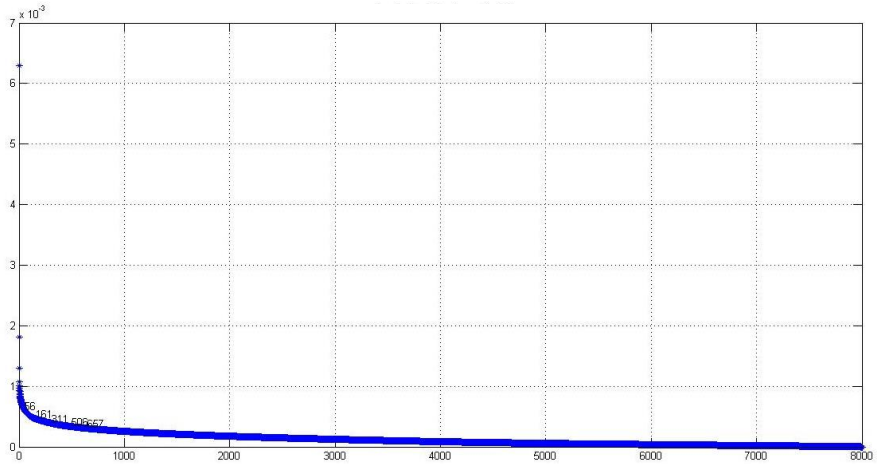

Figure 15– Plot of S relative of matrix A0.

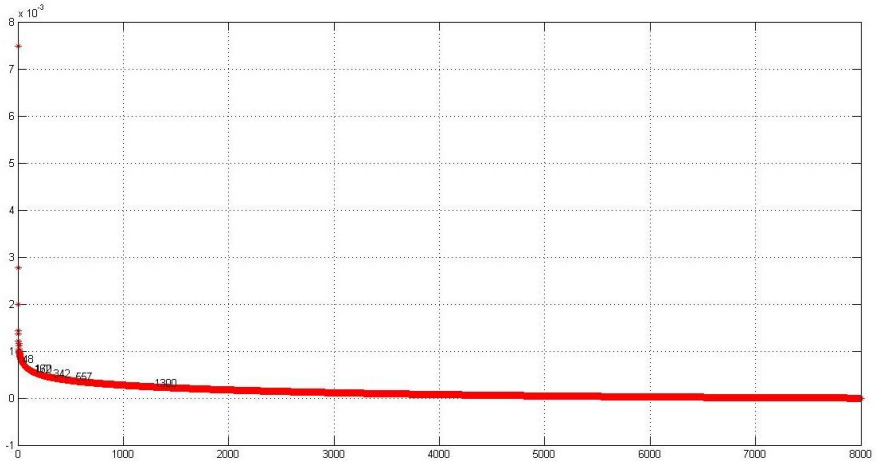

Figure 16 – Plot of S relative of matrix A1.

4.4 Clustering

After considering the parameters described in Table 1, we obtained the following results medoids/clusters described in Table 2.

Table 2 – Number of medoids.

TOTAL MATRIX	Medoids	medoids/last level
A0 (conventional sliding window)	395	313
A1 (conventional sliding window)	330	265
A0 (reverse sliding window)	370	297
A1 (reverse sliding window)	335	269

Source: Author

Clusters were well grouped (Figure 17 and 18). We plotted the first model with conventional and reverse sliding window, as an example. These techniques originated similar results.

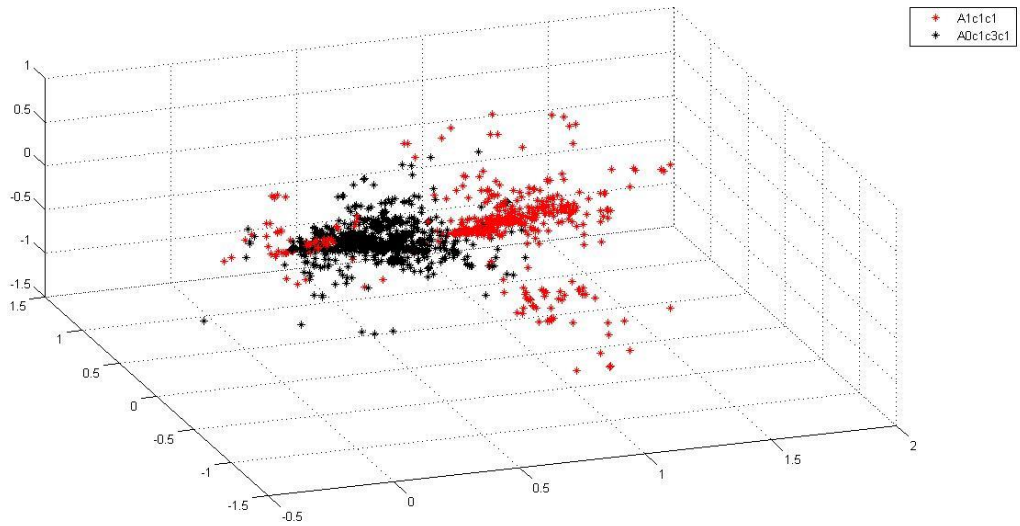

Figure 17– Plot 3D of the 1st model, with conventional sliding window.

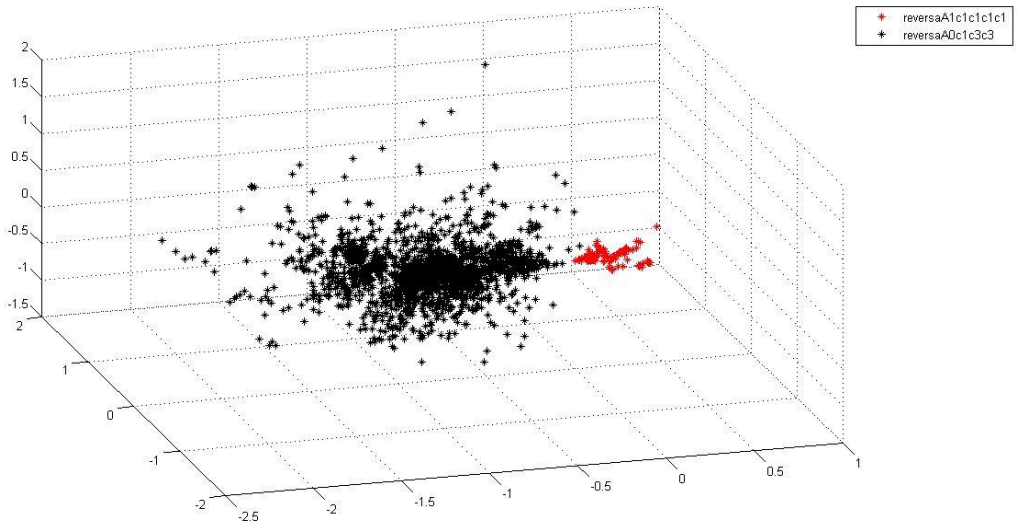

Figure 18– Plot 3D of the 1st model, with reverse sliding window.

4.5 Modified Logistic Regression Model

4.5.1 Small and near clusters method

Firstly, we used the equation for calculating the value of alpha (equation 11) of all features (triplets) for all models generated. Therefore, we did it 265 times for conventional sliding window technique and 269 for the reverse one.

As an example, we plotted the first model of the two techniques (Figures 19 and 20). As we can see, in both plots, alpha values are well distributed by 8000 triplets, varying only in magnitude.

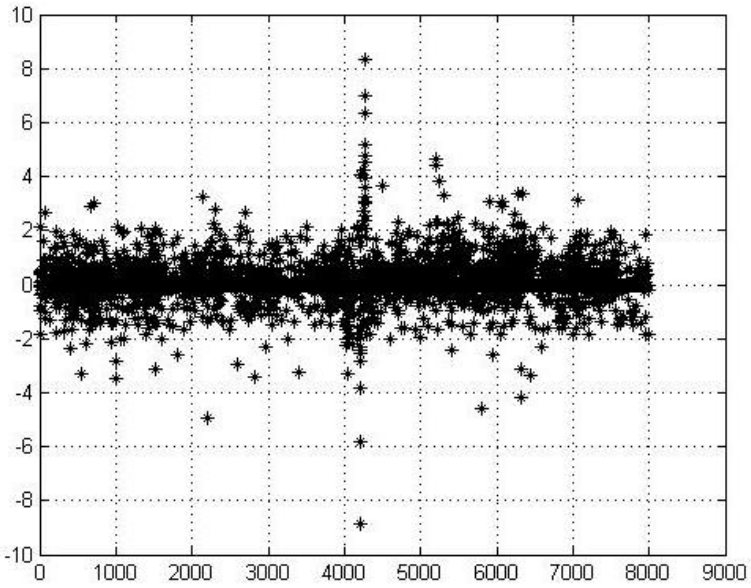

Figure 19– Alpha values of the 1st model, with conventional sliding window.

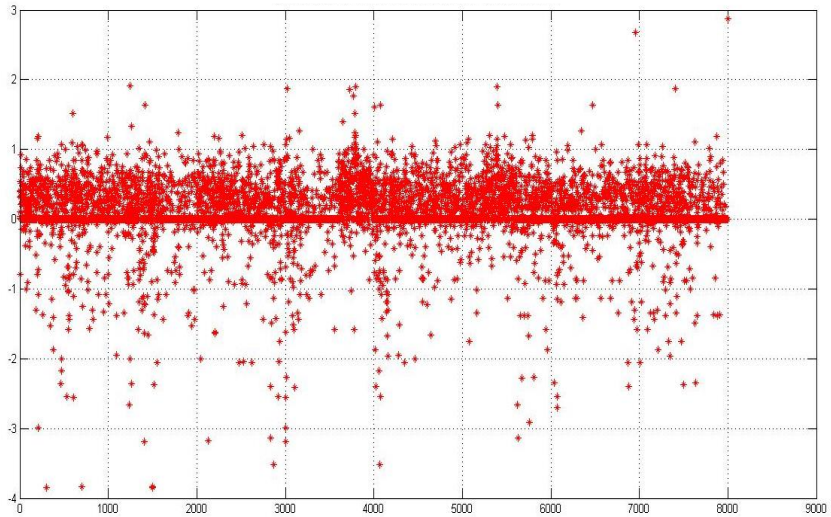

Figure 20– Alpha values of the 1st model, with reverse sliding window.

The graphs of the value of P (equation 12), using the two techniques, were well defined and separated (Figures 21 and 22).

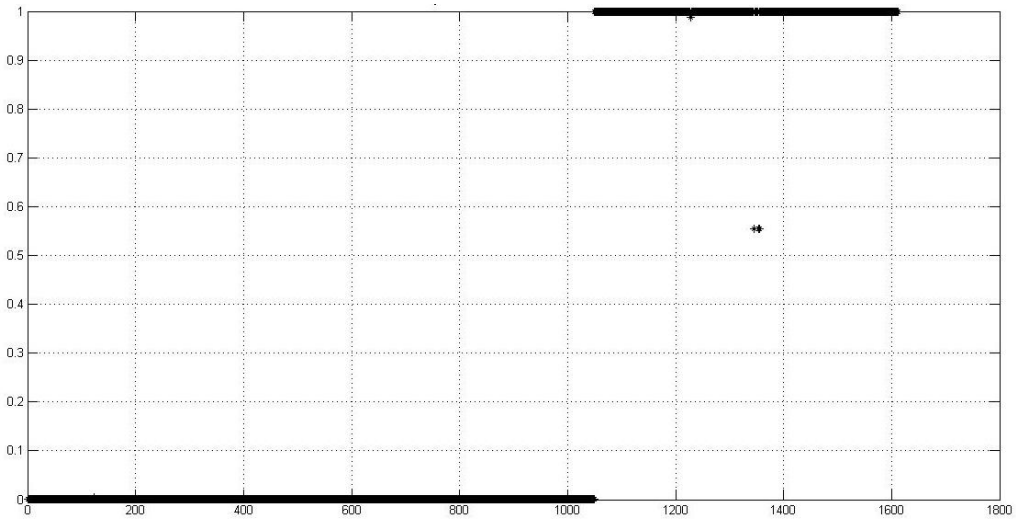

Figure 21– P values of the 1st model, with conventional sliding window.

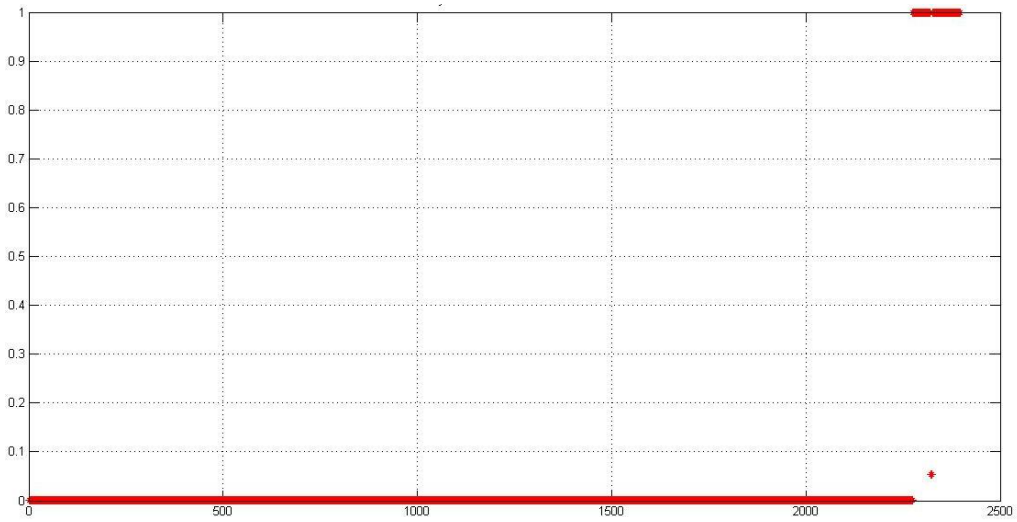

Figure 22– P values of the 1st model, with reverse sliding window.

The value of the norm of the model evaluated must be close to zero (equation 11). If it is not, we multiplied by a value that favored the remainder instead of the sum of the squares (Figures 23 and 24).

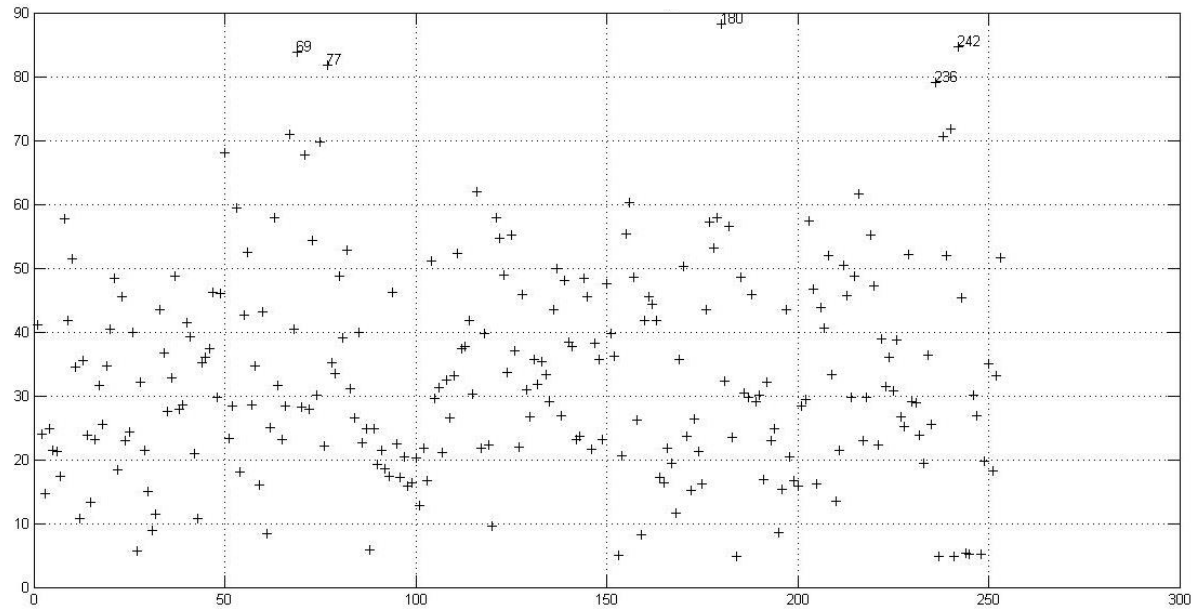

Figure 23– Plot of norm with conventional sliding window.

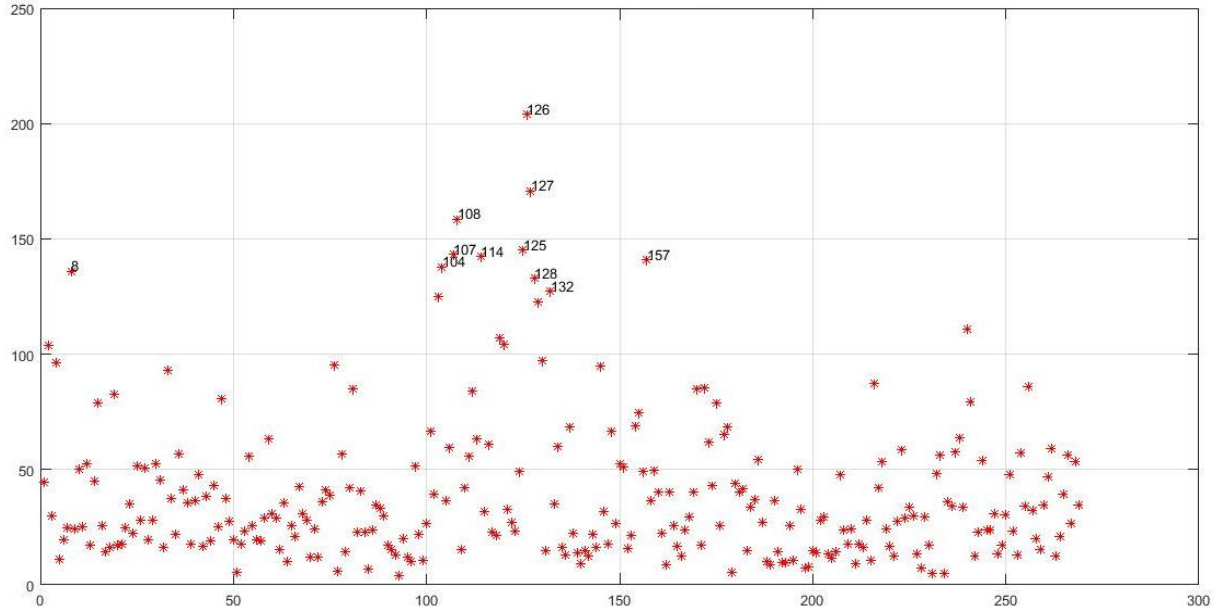

Figure 24– Plot of norm with reverse sliding window.

However, it was not necessary to multiply by any value because the worst norm (the highest) generated an acceptable model in both techniques (Figures 25 and 26).

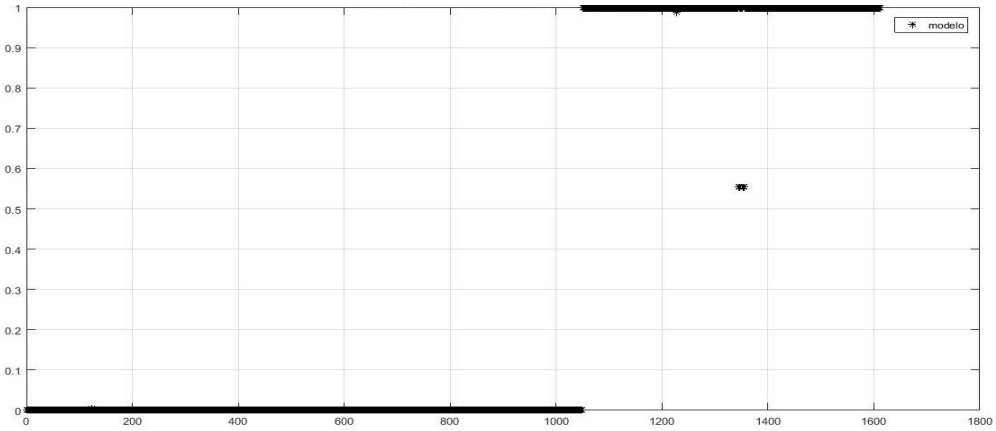

Figure 25– Plot of a model with a high norm, using conventional sliding window.

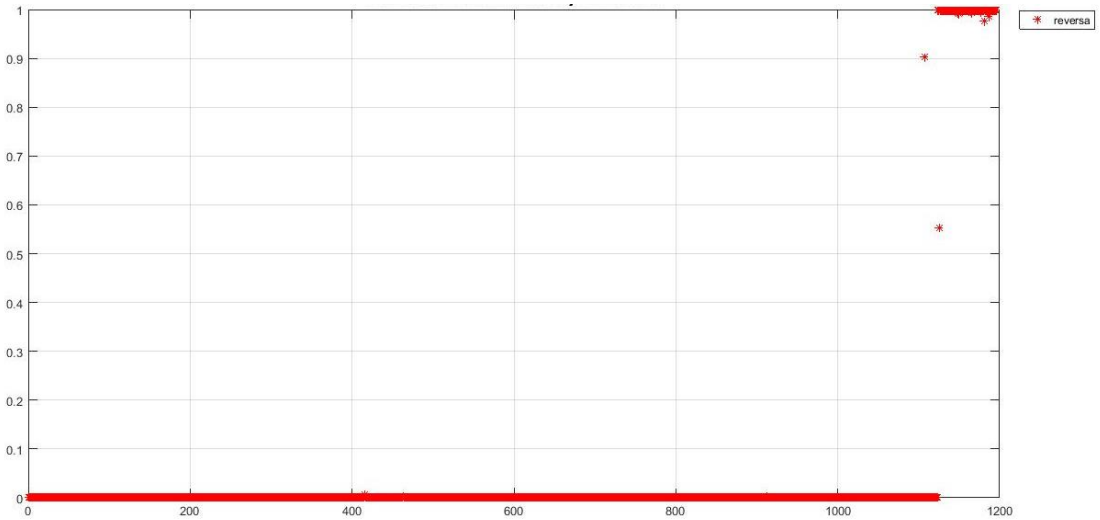

Figure 26– Plot of a model with a high norm, using reverse sliding window.

4.5.2 Medoids method

The matrices A0 and A1 had, with a conventional sliding windows technique, 395 and 330 medoids, respectively. The purpose was to extract the representatives of each cluster and concatenate them. Separately, the same procedure was done in the other matrices, which used the reverse sliding window technique. In this last case, we had 335 and 370 medoids, in A1 and A0, respectively. Once again, the purpose was to extract the representatives of each cluster and concatenate them. However, the built model was not good since there was no adequate separation

of individuals. Probably, we can infer that making a sample of a sampling was not the correct approach.

4.6 Cross-validation and performance measures

4.6.1 Small and near clusters method

As the only valid method was the small cluster and near cluster method, the cross-validation was done only for this method with the conventional window. The behavior with the reverse window was similar.

We chose the method K-fold cross-validation, where k is equal to 10% of the population. Therefore, we conducted 10 experiments with the selected cluster. Then, we calculated the sensitivity and specificity of each experiment and the respective average. The results of the first cluster A1 (one cluster of the 265's) are shown in the next figure (Figure 27). The cut-off was 0.5.

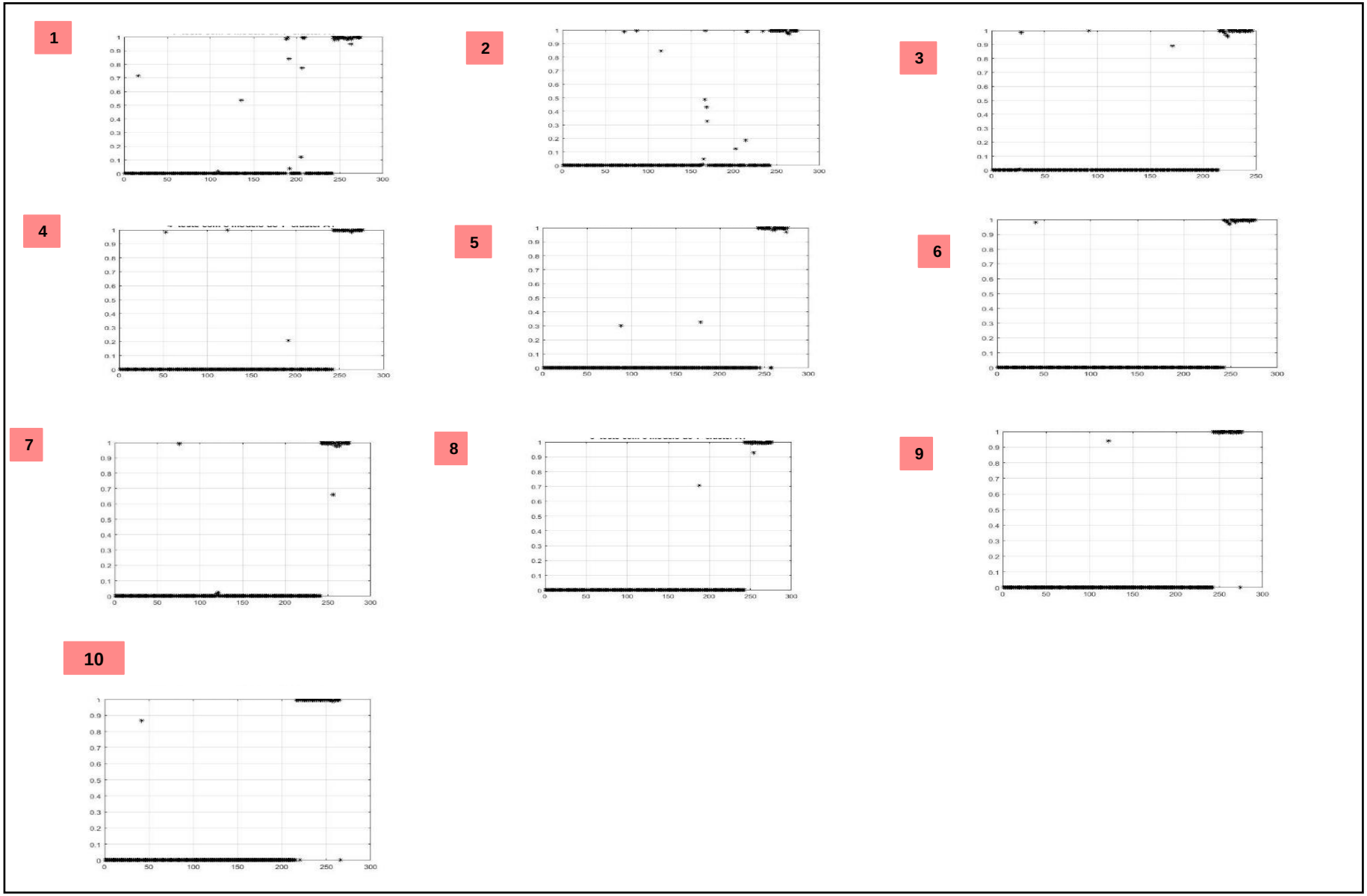

Figure 27 – Ten folds of cross-validation, using a model with conventional sliding window.

As the results of cross-validation were satisfactory, it was possible to proceed to the test with the unknown query (Table 3).

Table 3 - Results of performance measures of a model with conventional sliding window.

Test	Accuracy	Sensibility	Specificity
1	0,963504	1	0,960159
2	0,974453	1	0,972
3	0,98374	0,967742	0,986047
4	0,98913	0,970588	0,991736
5	0,985401	0,891892	1
6	0,992701	0,969697	0,995851
7	0,992674	0,969697	0,995833
8	0,992701	0,969697	0,995851
9	0,989051	0,941176	0,995833
10	0,992701	0,983051	0,995349
$\bar{x}$	0,985625	0,964809	0,988617

Source: Author

4.7 Comparing with another tool

Although we can predict only one type of secondary structure, it is necessary to compare the results obtained with another method. The QUARK tool, which uses fragments with a size of 20 residues, was the chosen tool for comparing. QUARK had a good performance in CASP (http://predictioncenter.org).

The comparison is still very limited, and therefore we only used the sensitivity metric. A fragment classified as negative, by our method, can still contain residues in alpha-helix. The PDB file (Berman *et al.*, 2000) randomly chosen was ID: 3etj to be tested with the QUARK tool and our method. The resources of PDBsum (Laskowski *et al.*, 1997) and PDB (Berman *et al.*, 2000) were used to generate elucidatory images about the secondary structure of 3etj (Figure 28 and 29).

Table of helices

No.	Start	End	Type	No. resid	Length	Unit rise	Residues per turn	Pitch	Deviation from ideal	Sequence
*1.	Gln11	Gly20	H	10	15.48	1.49	3.58	5.34	2.0	QLGRMLRQAG
*2.	Glu21	Leu23	G	3	-	-	-	-	-	EPL
*3.	Pro36	Ala38	G	3	-	-	-	-	-	PAA
*4.	Phe41	Gln43	G	3	-	-	-	-	-	FQQ
5.	Ala57	Ala63	H	7	10.94	1.51	3.79	5.71	7.5	ALTRQLA
6.	Phe74	Ala78	H	5	7.85	1.42	3.91	5.57	23.4	FPIIA
7.	Arg80	Lys89	H	10	15.27	1.46	3.59	5.25	4.7	RLTQKQLFDK
8.	Arg103	Glu105	G	3	-	-	-	-	-	RSE
9.	Trp106	Leu113	H	8	11.47	1.40	3.78	5.29	13.7	WPAVFDRL
10.	Ala136	Gln141	G	6	10.86	1.90	3.03	5.75	30.5	ANETEQ
11.	Ala144	Cys146	G	3	-	-	-	-	-	AEC
12.	Ala200	Leu217	H	18	26.66	1.46	3.65	5.34	12.6	AQQQARAEEMLSAIMQEL
13.	Asn245	Trp249	G	5	9.76	1.99	3.63	7.23	24.8	NSGHW
14.	Thr250	Gly253	H	4	4.16	0.75	4.32	3.25	45.0	TQNG
15.	Gln258	Ile266	H	9	13.92	1.50	3.64	5.46	8.7	QFELHLRAI
16.	Tyr292	Lys296	G	5	9.64	1.82	3.20	5.84	16.7	YDWLK
17.	Thr325	Leu335	H	11	17.19	1.50	3.61	5.42	8.1	TSRLTATLEAL
18.	Ile336	Leu338	G	3	-	-	-	-	-	IPL
19.	Pro341	Tyr343	G	3	-	-	-	-	-	PEY
20.	Ala344	Lys353	H	10	16.14	1.54	3.61	5.58	7.1	ASGVIWAQSK

Figure 28 - Image generated by PDBsum with the PDB ID 3etj (Laskowski *et al.*, 1997).

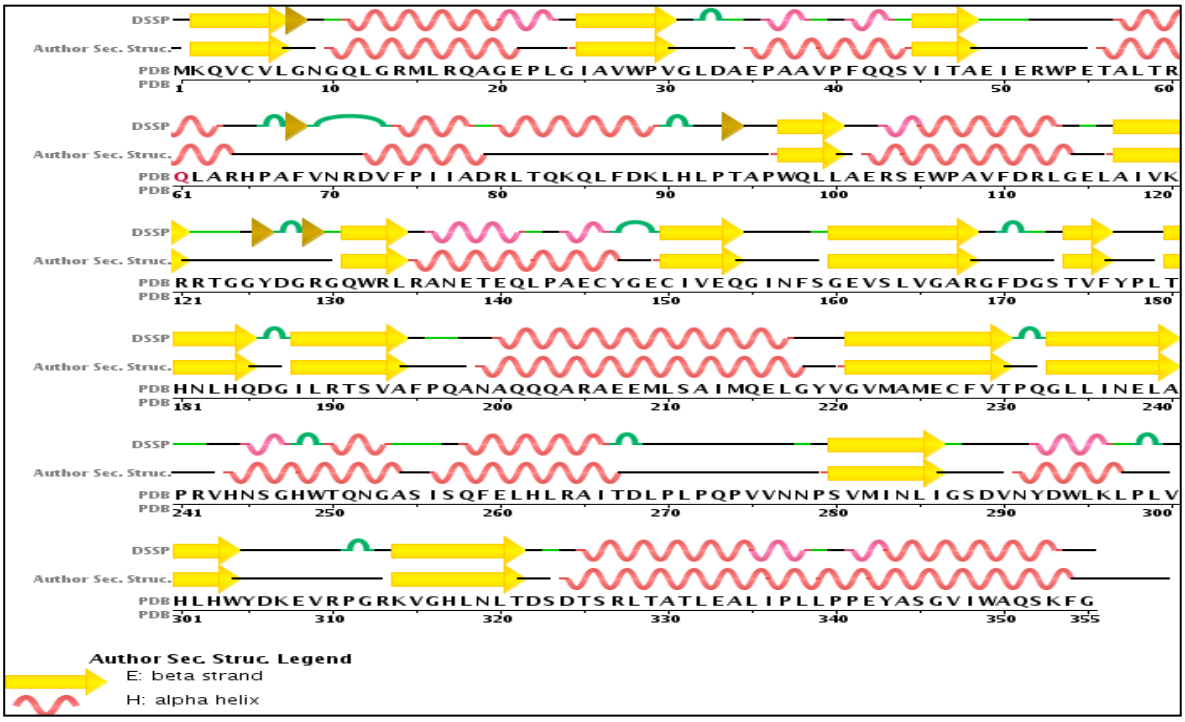

Figure 29 – Image of secondary structure of PDB ID 3etj (Berman *et al.*, 2000).

The QUARK tool only accepts size fragments with 20 residues. Thus, we inserted the sequence containing the alpha-helix, as highlighted in Table 4. The overall sensitivity was 67,61%.

Table 4 - Results of the comparison with QUARK tool. Blue highlight - alpha-helix; Yellow highlight - remaining fragment.

Fragment with 20 residues containing the alpha helix (PDB ID: 3etj)	C-coil;H-helix;E-sheet;T-beta turn				number of correctly predicted residues
HELICE 1: QVCVLGNGQLG RMLRQAGEP	1 Q E 2 V E 3 C E 4 V E 5 L E	6 G C 7 N C 8 G C 9 Q H 10 L H	11 G H 12 R H 13 M H 14 L H 15 R H	16 Q H 17 A C 18 G C 19 E C 20 P C	8/10
HELICE 7: NRDVFPIIADRLT QKQLFDK	1 N C 2 R C 3 D C 4 V H 5 F H	6 P H 7 I H 8 I H 9 A H 10 D H	11 R H 12 L H 13 T H 14 Q H 15 K H	16 Q H 17 L H 18 F H 19 D C 20 K C	8/10
HELICE12: ANAQQQARAEE MLSAIMQEL	1 A C 2 N C 3 A H 4 Q H 5 Q H	6 Q H 7 A H 8 R H 9 A H 10 E H	11 E H 12 M H 13 L H 14 S H 15 A H	16 I H 17 M H 18 Q H 19 E H 20 L C	17/18
HELICE15: QFELHLRAITDL PLPQPVVN	1 C 2 F C 3 E E 4 L E 5 H E	6 L E 7 R E 8 A E 9 I H 10 T C	11 D C 12 L C 13 P C 14 L C 15 P C	16 Q C 17 P C 18 V C 19 V C 20 N C	1/9
HELICE 17: GHLNLTDSDTSR LTATLEAL	1 G C 2 H C 3 L C 4 N C 5 L C	6 T C 7 D C 8 S C 9 D C 10 T H	11 S H 12 R H 13 L H 14 T H 15 A H	16 T H 17 L H 18 E H 19 A H 20 L C	10/11
HELICE20: ALIPLLP PEY ASGVIWAQSK	1 A C 2 L C 3 I C 4 P C 5 L C	6 L C 7 P C 8 P H 9 E H 10 Y H	11 A H 12 S C 13 G C 14 V E 15 I E	16 W E 17 A E 18 Q E 19 S C 20 K C	4/13
TOTAL					48/71 (67,61%)

Source: Author

4.8 Project and calculate the odds of an unknown *query*

4.8.1 Small and near clusters method

After constructing all models by Small and near clusters method, we tested it with unknown queries i.e., to calculate the odds of biomolecules randomly selected from PDB. In the first step of the tests, we chose only enzymes (one of each enzyme family). This step worked as a preliminary test to evaluate the performance of the method. For each tested biomolecule, we verified if it was not present in the sample that was created of the databank initially. The same procedure was done with the conventional sliding window and the reverse one. The result value P is shown in Figure 30. The behavior of the reverse window was similar.

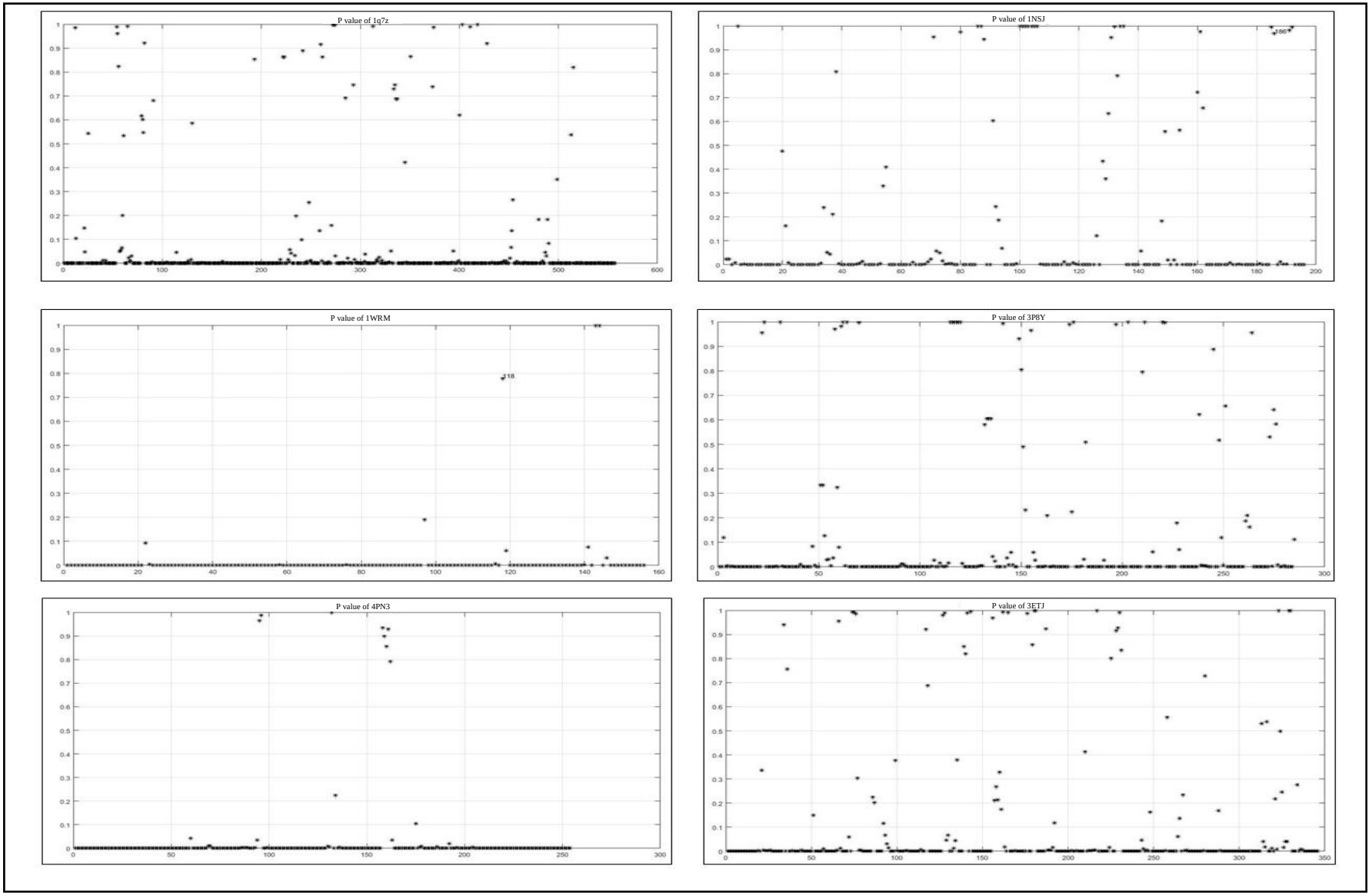

Figure 30 - Results of P values of the six tested queries, using conventional sliding window.

Considering a reference cut-off of 0.70, the results of sensibility (equation 13), specificity (equation 14), and accuracy (equation 15), is described in Table 5.

Table 5 - Results of the tested queries.

ID PDB	Enzyme Classification	Sensibility	Specificity	Accuracy
1Q7Z	Transferase	6/41 = 0.20	277/305 = 0.91	283/346 = 0.82
1NSJ	Isomerase	4/14 = 0.30	151/182 = 0.83	155/196 = 0.79
1WRM	Hidrolase	2/37 = 0.10	118/119 = 0.99	120/156 = 0.77
3P8Y	Ligase	8/52 = 0.15	198/233 = 0.85	206/285 = 0.72
4PN3	Oxidoreductase	0/73 = 0	173/181 = 0.96	173/254 = 0.68
3ETJ	Lyase	7/41 = 0.17	252/305 = 0.83	259/346 = 0.75
$\bar{x}$		0.15	0.90	0.76

Source: Author

As can be seen, the specificity had good values, i.e., the method could identify fragments created 100% non alpha-helix. However, sensibility was low. This behavior was consistent among the tested proteins. Therefore, we chose to stop here. We believed that it was an indication that something was wrong and to test more biomolecules would not provide more useful information.

We made several tests on the clusters, such as number variation and maximum size. We also inverted the starting point to A0 matrix and combined the distance metrics. However, the behavior of the tested biomolecules did not vary in both techniques of sliding windows.

In our efforts, we found the possible mistake when we checked how many triplets coincided between the queries and the model. The number of triplets that matched should have been 7 but were nevertheless 2-3. This classification would be difficult because it is necessary seven overlapping triplets associated with the alpha value to calculate the odds of the fragment (equation 12). Thus, we understand the results of sensitivity and specificity we got.

5 Final Considerations

This work had many and different stages. Working with the PDB database proved to be a more complex task than one could predict due to the quantity of information and the failure of the same.

The hierarchical clustering using the k and medoid algorithm could satisfactorily separate database.

Validation of small and near method obtained good results in the cross-validation. This method was the method chosen by us, not only by the results close to satisfactory but also for its potential. With it, we can classify the fragments 100% non alpha-helix.

Logistic regression provided a useful method for classification by modeling the probability of relationship of a class based on linear combinations of exploratory variables. One particular problem was multicollinearity: the estimated equations had no unique solution. The modified logistic regression model proposed in this work is a solution to this problem, with no need for previous feature selection nor matrix dimensionality reduction. Based on the work of Linnik (1961) and Golub (1965) we modified the classical logistic regression model to include a stabilizing term (equations 11) that allowed for the assignment of values to alpha parameters by minimizing the square sum of the residuals ($B\alpha - b$) summed to the squares of alpha.

Although our outcomes were not as expected, it was a great learning experience to acquire knowledge of the proposed topic. At the end of this work, we have a method (small and near method cluster) with strong evidence that might be a potential method of predicting secondary structure. However, this method has a versatile and scalable properties to be applied to other problems.

6 Perspectives

With our hypothesis unrepresentative, we could use all PDB data bank to create another database or to extract just one family of biomolecules. It is also necessary, to make other models using fragments with different percentages of alpha-helix. As well as, models for beta sheet and coils.

After implementing those improvements described above, we can publish this work in the form of an article. With the resources that MatLab offers, it will also be possible to provide an extension to the scientific community for predicting secondary structure.

Another perspective, it is to use the small method and near method in other data mining problems, considering the scalable features of the model.

7 References

ABREU, M. N. et al. Ordinal logistic regression models: application in quality of life studies. **Cad Saude Publica,** v. 24 Suppl 4, p. s581-91, 2008. ISSN 0102-311x.

ANFINSEN, C. B. **Studies on the principles that govern the folding of protein chains** 1972.

ANFINSEN, C. B. et al. STUDIES ON THE GROSS STRUCTURE, CROSS-LINKAGES, AND TERMINAL SEQUENCES IN RIBONUCLEASE. **Journal of Biological Chemistry,** v. 207, n. 1, p. 201-210, 1954.

ARLOT, S.; CELISSE, A. A survey of cross-validation procedures for model selection. **Statistics surveys,** v. 4, p. 40-79, 2010. ISSN 1935-7516.

BERMAN, H. M. et al. The Protein Data Bank. **Nucleic Acids Research,** v. 28, n. 1, p. 235-242, 2000.

BERRY, M. W.; DUMAIS, S. T.; OBRIEN, G. W. Using Linear Algebra for Intelligent Information Retrieval. **SIAM Review,** v. 37, 1995.

BRAGG, L.; KENDREW, J. C.; PERUTZ, M. F. Polypeptide Chain Configurations in Crystalline Proteins. **Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences,** v. 203, n. 1074, p. 321-357, 1950-10-10 00:00:00 1950.

BRASIL, I. FUNçõES PLENAMENTE RECONHECIDAS DE NUTRIENTES. 2009.

CHERNOFF, R. Protein and older adults. **Journal of the American College of Nutrition,** v. 23, n. sup6, p. 627S-630S, 2004. ISSN 0731-5724.

CILIBRASI, R. L.; VITANYI, P. M. B. The google similarity distance. **IEEE Transactions on knowledge and data engineering,** v. 19, n. 3, p. 370-383, 2007. ISSN 1041-4347.

COKLUK, O. Logistic Regression: Concept and Application. **Educational Sciences: Theory and Practice,** v. 10, n. 3, p. 1397-1407, 2010. ISSN 1303-0485.

COUTO, B.; LADEIRA, A. P.; SANTOS, M. A. Application of latent semantic indexing to evaluate the similarity of sets of sequences without multiple alignments character-by-character. **Genet Mol Res,** v. 6, n. 4, p. 983-999, 2007.

DE OLIVEIRA, S. H. P.; SHI, J.; DEANE, C. M. Building a Better Fragment Library for De Novo Protein Structure Prediction. **PLoS ONE**, San Francisco, CA USA**,** v. 10, n. 4, p. e0123998.

DEERWESTER, S. et al. Indexing by latent semantic analysis. **Journal of the American society for information science,** v. 41, n. 6, p. 391, 1990. ISSN 0002-8231.

DEHZANGI, A. et al. Proposing a highly accurate protein structural class predictor using segmentation-based features. **BMC Genomics,** v. 15 Suppl 1, p. S2, 2014. ISSN 1471-2164.

ELLIOTT, W. H.; ELLIOTT, D. C. **Biochemistry and molecular biology**. 4th ed. Oxford: Oxford University Press, 2009. ISBN 9780199226719 (pbk.)

EVERITT, B. S. D.; EVERITT, G. B. S.; DUNN, G. **Applied multivariate data analysis**. 1991

FODJE, M. N.; AL-KARADAGHI, S. Occurrence, conformational features and amino acid propensities for the π-helix. **Protein Engineering,** v. 15, n. 5, p. 353-358, 2002.

GOLUB, G. H. Numerical methods for solving linear least squares problems. **Numer. Math.,** v. 7, n. 3, p. 206-216, 1965. ISSN 0029-599X.

HEFFERNAN, R. et al. Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. **Scientific Reports,** v. 5, p. 11476, 06/22/online 2015.

HODGMAN, T. C. A historical perspective on gene/protein functional assignment. **Bioinformatics,** v. 16, n. 1, p. 10-15, 2000. ISSN 1367-4803.

HOLMES, J. B.; TSAI, J. Some fundamental aspects of building protein structures from fragment libraries. **Protein Science : A Publication of the Protein Society,** v. 13, n. 6, p. 1636-1650.

HOSMER, D.; LEMESHOW, W. S. **Applied logistic regression**. Nova York: Wiley, 2000.

HUGGINS, M. L. The Structure of Fibrous Proteins. **Chemical Reviews,** v. 32, n. 2, p. 195-218, 1943/04/01 1943. ISSN 0009-2665.

JAYANTHI, P. G. **Molecular Biology**. MJP Publishers, 2010. ISBN 9788180940576.

JONES, D. T.; MCGUFFIN, L. J. Assembling novel protein folds from super-secondary structural fragments. **Proteins,** v. 53 Suppl 6, p. 480-5, 2003. ISSN 0887-3585.

KABSCH, W.; SANDER, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. **Biopolymers,** v. 22, n. 12, p. 2577-2637, 1983. ISSN 1097-0282.

KALEV, I.; HABECK, M. HHfrag: HMM-based fragment detection using HHpred. **Bioinformatics,** v. 27, n. 22, p. 3110-6, Nov 15 2011. ISSN 1367-4803.

KAUFMAN, L.; HOPKE, P. K.; ROUSSEEUW, P. J. Using a parallel computer system for statistical resampling methods. **Computational Statistics Quarterly,** v. 2, p. 129-141, 1987.

KAUFMAN, L.; ROUSSEEUW, P. Clustering by means of medoids. **Statistical Data Analysis Based on the L1-Norm and Related Methods**, p. North-Holland, 1987.

_____. **Finding Groups in Data An Introduction to Cluster Analysis**. 1990.

KAUFMAN, L.; ROUSSEEUW, P. J. Partitioning Around Medoids (Program PAM). In: (Ed.). **Finding Groups in Data**: John Wiley & Sons, Inc., 1990. p.68-125. ISBN 9780470316801.

KOEHLER LEMAN, J.; ULMSCHNEIDER, M. B.; GRAY, J. J. Computational modeling of membrane proteins. **Proteins,** v. 83, n. 1, p. 1-24, Jan 2015. ISSN 0887-3585.

KUMAR, N.; NASSER, M.; SARKER, S. C. A new singular value decomposition based robust graphical clustering technique and its application in climatic data. **Journal of Geography and Geology,** v. 3, n. 1, p. p227, 2011. ISSN 1916-9787.

KURGAN, L. A.; HOMAEIAN, L. Prediction of structural classes for protein sequences and domains—Impact of prediction algorithms, sequence representation and homology, and test procedures on accuracy. **Pattern Recognition,** v. 39, n. 12, p. 2323-2343, 12// 2006. ISSN 0031-3203.

LASKOWSKI, R. A. et al. PDBsum: a web-based database of summaries and analyses of all PDB structures. **Trends in Biochemical Sciences,** v. 22, n. 12, p. 488-490, 1997/12/01 1997. ISSN 0968-0004.

LEHNINGER, A. L.; NELSON, D. L.; COX, M. M. **Lehninger principles of biochemistry**. 4th ed. New York, N.Y. ; Basingstoke: W.H. Freeman, 2005. ISBN 9780716743392.

LI, S. C. et al. Designing succinct structural alphabets. **Bioinformatics,** v. 24, n. 13, p. i182-9, Jul 1 2008. ISSN 1367-4803.

LINDERSTROM-LANG, K. U. **Proteins and Enzymes**. Stanford, California: Stanford University Press, 1952. 115.

LINNIK, Y. V. **Methods of Least Squares and Principles of the Theory of Observations**. Oxford-London-New York-Paris: Pergamon Press, 1961.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, 1967, Berkeley, Calif. University of California Press, 1967. p.281-297.

MARCOLINO, L.; COUTO, B.; DOS SANTOS, M. Genome Visualization in Space. **Advances in Bioinformatics**, p. 225-232, 2010.

MENARD, S. **Logistic regression: From introductory to advanced concepts and applications**. Sage Publications, 2010. ISBN 1483351424.

MOULT, J. et al. Critical assessment of methods of protein structure prediction (CASP) — round x. **Proteins: Structure, Function, and Bioinformatics,** v. 82, n. Supplement S2, p. 1-6, 2014.

PAULING, L.; COREY, R. B.; BRANSON, H. R. The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. **Proc Natl Acad Sci U S A,** v. 37, n. 4, p. 205-11, Apr 1951. ISSN 0027-8424.

RAMOS, C. H. I. História-Proteinas II. **CBME Informação,** v. 3, p. 3, 2004.

ROHL, C. A. et al. Protein structure prediction using Rosetta. **Methods Enzymol,** v. 383, p. 66-93, 2004. ISSN 0076-6879.

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics,** v. 20, p. 53-65, 1987.

SANGER, F. Sequences, sequences, and sequences. **Annual review of biochemistry,** v. 57, n. 1, p. 1-29, 1988. ISSN 0066-4154.

SANTOS, A. R. et al. A singular value decomposition approach for improved taxonomic classification of biological sequences. **BMC Genomics,** v. 12, n. 4, p. 1-15, 2011. ISSN 1471-2164.

SIMONS, K. T. et al. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. **J Mol Biol,** v. 268, n. 1, p. 209-25, Apr 25 1997. ISSN 0022-2836.

STRUYF, A.; HUBERT, M.; ROUSSEEUW, P. Clustering in an object-oriented environment. **Journal of Statistical Software,** v. 1, n. 4, p. 1-30, 1997.

VAUQUELIN, L.-N.; ROBIQUET, P. J. The discovery of a new plant principle in Asparagus sativus. **Ann. Chim.(Paris),** v. 57, n. 2, p. 1, 1806.

VERLI, H. Bioinformática: Da Biologia à Flexibilidade Molecular. **São Paulo: SBBq**, 2014.

WALL, M. E.; RECHTSTEINER, A.; ROCHA, L. M. Singular value decomposition and principal component analysis. In: (Ed.). **A practical approach to microarray data analysis**: Springer, 2003. p.91-109. ISBN 1402072600.

WANG, G.; DUNBRACK, R. L. PISCES: a protein sequence culling server. **Bioinformatics,** v. 19, n. 12, p. 1589-91, Aug 2003. ISSN 1367-4803.

XIONG, J. **Essential bioinformatics**. Cambridge University Press, 2006. ISBN 113945062X.

XU, D.; ZHANG, Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. **Proteins: Structure, Function, and Bioinformatics,** v. 80, n. 7, p. 1715-1735, 2012. ISSN 1097-0134.

ÉLDEN, L. Numerical linear algebra in data mining. **Acta Numerica,** v. 15, 2006.

8 Appendix A

Table - Tools and method to predict the secondary structure of protein.

Name	Author (citation)	Main features
TOOLS		
C8-SCORPION	(Yaseen e Li, 2014)	The authors constructed the templates for 8-state secondary structure, from structural information of chains with certain sequence similarity. The structural templates are then incorporated as features with sequence and evolutionary information to train two-stage neural networks. In the case of structural templates absence, heuristic structural information is incorporated instead.
NEUROSVM	(Ghanty *et al.*, 2013)	It is a hybrid system consisting of neural networks and support vector machines for classification of secondary structures, using position-specific probability-based features and position-independent probability-based features. This method uses the position-specific scoring matrices (PSSM) derived from PSI-BLAST.
JPred4	(Cuf*f et al.*, 1998) (Drozdetski*y et al.*, 2015)	This server is a combination of six secondary structure prediction algorithms that exploit evolutionary information from multiple sequences.
FLOPRED	(Saraswathi *et al.*, 2012)	It uses neural network-based extreme learning machine and advanced particle swarm optimization, using the information from CATH (Protein Structure Classification) database.

sS2	(Sormanni *et al.*, 2015)	It is based on the NMR chemical shifts that provide quantitative information about the probability distributions of secondary-structure elements in disordered states.
BCL::ScoreProtein	(Woetzel *et al.*, 2012)	It uses the topology, considering that the majority of well-structured domains for the assembly of the secondary structure elements, in three-dimensional space defines the domain topology. It defines an amino acid pair potential, an amino acid environment potential, a secondary structure element packing potential, a β-strand pairing potential, a loop length potential, a radius of gyration potential, a contact order potential, and a secondary structure formation potential. This scoring function is specialized to evaluate the loop-less protein topology as defined by the secondary structure elements.
BCL::Fold	(Karakas *et al.*, 2012)	The algorithm uses the Monte Carlo Metropolis simulated annealing folding simulation. It optimizes a knowledge-based potential of BCL::Score. Discontinuation of the protein chain favors sampling of non-local contacts and the thereby creation of complex protein topologies.
DISSPred	(Kountouris e Hirst, 2009)	DISSPred predicts both the secondary structure and the backbone dihedral angles independently and combine the results.
SVM-PB-Pred	(Suresh e Parthasarathy, 2014)	It is the support vector machine which is used to predict the protein block, with the input sequence profile (Position-Specific Scoring Matrix) and secondary structures from different methods.
RKS_PPSC	(Yang *et al.*, 2010)	The RKS_PPSC predict protein structural classes particularly for low-homology amino acid sequences, based on features extracted from the predicted secondary

		structures of proteins rather than directly from their amino acid sequences
Phyre2	(Kelley *et al.*, 2015)	It uses remote homology detection methods to build 3D models, predict ligand binding sites and analyze the effect of amino acid variants (single nucleotide polymorphism).
METHODS:		
	(Wang *et al.*, 2015)	This method proposes a prediction for low-similarity datasets using reduced PSSM and position-based secondary structural features.
	(Zhang *et al.*, 2014)	This a method to predict the structural class of proteins, especially for low-similarity sequences, combining PSIPRED, feature selection, and support vector machine model.
	(Kong e Zhang, 2014)	This method uses 27 features that characterize general contents and spatial arrangements, using the support vector machine to implement the prediction.
INTEGRATIVE TOOLS:		
MULTICOM toolbox	(Cheng *et al.*, 2012)	This has a set of protein structure and structural feature prediction tools, including secondary structure prediction, solvent accessibility prediction, disorder region prediction, protein domain boundary prediction, protein contact map prediction, protein disulfide bond prediction, protein beta-sheet structure prediction, protein fold recognition, multiple template combination, template-based structure modeling, protein model quality assessment, and protein mutation analysis.
SCRATCH	(Cheng *et al.*, 2005)	This software includes predictors for secondary structure, relative solvent accessibility, disordered regions,

		domains, disulfide bridges, single mutation stability, residue contacts versus average, individual residue contacts, and tertiary structure. The user simply provides an amino acid sequence and selects the desired predictions, then submits to the server.
PredictProtein	(Yachdav et al., 2014)	The PredictProtein aggregates a large number of tools to predict the secondary structure. It features prediction for secondary structure, solvent accessibility, transmembrane helices, globular regions, coiled-coil regions, structural switch regions, B-values, disorder regions, intra-residue contacts, protein-protein, and protein-DNA binding sites, sub-cellular localization, domain boundaries, beta-barrels, cysteine bonds, metal binding sites and disulfide bridges.
MESSA	(Cong and Grishin, 2012)	It predicts the secondary structure, local sequence features, domain architecture, and function for a given protein sequence.
RaptorX	(Wang *et al*., 2011)	RaptorX uses conditional neural fields to predict 8-class, taking as input PSSM generated by PSIBLAST, the physicochemical properties of amino acids and their statistical properties to predict secondary structure.
ItFix-SPEED	(Debartolo *et al*., 2010)	It is also possible coupling tools, which predicts secondary and tertiary structure.
I-TASSER	(Roy *et al*., 2011)	This server matches the predicted 3D models to the proteins in 3 independent libraries which consist of proteins of known enzyme classification number, gene ontology vocabulary, and ligand-binding sites.