Un Modelo Basado en Reglas de Asociación para la Extracción de Información en Datos de Secuencias de Proteínas

An Association Rule Based Model for Information Extraction from Protein Sequence Data

David Becerra, Ing., Giovanni Cantor, Ing., Luis F. Niño, PhD.,
Jonatan Gómez, PhD., Leonardo Bobadilla, PhD.
Laboratorio de Investigación en Sistemas Inteligentes, ALGOS-UN
Universidad Nacional de Colombia - Sede Bogotá
{dcbecerrar, gacantorm, lfninov, jgomezpe, jlbobadillam} @unal.edu.co

Recibido para revisión 28 de Noviembre de 2007, aceptado 14 de Febrero de 2008, versión final 28 de Febrero de 2008

Resumen—**En este trabajo se presenta una técnica de minería de datos para la extracción de patrones en secuencias de proteínas. Específicamente, el objetivo es explorar el uso de reglas de asociación como una base para construir exitosamente predictores de estructuras secundarias en una capa estructura - secuencia. En esta investigación no se toma en cuenta información biológica ni heurística, es decir, que solo la información otorgada por las reglas de asociación se utiliza como una base de construcción de un predictor de estructura secundaria.**
Este trabajo proporciona elementos de comprensión acerca de las características de predicción de estructuras secundarias para ser usadas en algoritmos de aprendizaje, es esperado que este trabajo sea útil para alcanzar mejoras substanciales en la precisión de la predicción de estructuras secundarias en trabajos futuros.

Palabras Clave—**Minería de Datos, Predicción de Estructuras Secundarias, Reglas de Asociación.**

Abstract—**In this paper, a data mining technique for protein sequence pattern extraction is developed. Specifically, the aim is to explore the use of association rules as a basis to build successful secondary structure predictors, in a sequence-structure layer.**
No heuristic or biological information is taken into account in the present study and only the information given by the association rules is used as a basis for building a secondary structure predictor. This work gives some insights about secondary structure prediction features to be used in learning algorithms; this is expected to be useful to achieve substantial improvements of accuracy in protein secondary structure prediction.

Keywords—**Data Mining, Secondary Structure Prediction, Association Rules.**

I. INTRODUCTION

Scientists have studied the complex process that determines the structure, properties and function of proteins for decades; however, such processes and mechanisms about protein folding and the prediction of secondary structures still remain unknown. Predicting a protein secondary structure consists of the classification of the amino acids in a sequence as either helices (H) or sheets (E) or coils (C).

Secondary structure prediction could be studied as a machine learning problem by performing either classification or pattern recognition; classification is then based on the features of a protein sequence.

Secondary structure prediction methods can be categorized in four different generations [1]. The first generation was based on propensities of single residues, i.e., it was based on single amino acid propensities for finding a specific amino acid in a specific structural element; the methods developed by Chou and Pasman[2] and the method GOR developed by Granier et el[3] were among the most significant. Second generation methods were based on propensities of segments as opposed to isolated amino acids. Representative approaches are the work by Rooman and Wodak [4], Strelets [5] and the method called PREDATOR [6]. In the third generation, information from homologues sequences to the query sequence and state of the

art machine learning methods were used. Among the representative approaches are Zvelebil[7], PHD[8] and PSI-PRED[9]. In fourth generation approaches, a matching between secondary and tertiary protein structure was used; in other words, information about 3D protein conformation was added to secondary structure predictive methods as in the work by Meiler and Baker [10].

In spite of the progress achieved by secondary structure prediction approaches, they have reached around 77% average prediction accuracy per residue in unknown protein sequences [11].

Since the 70's, several approaches to solve the secondary structure prediction problem based on machine learning techniques have been proposed; support vector machines and neural networks have been successfully applied, obtaining similar results in terms of prediction accuracy.

Despite of good prediction showed by machine learning methods, the results given by some of them, especially those based on neural networks, are difficult to interpret. Therefore, some probabilistic models, which are easier to interpret, have been developed [12].

Figure 1 depicts a general model for secondary structure prediction based on sequence. The first step consists of potential conserved and interesting patters taken from a sequence data base. The next step consists of the extraction and definition of the patterns found in the first step and developing the model itself. Finally an optimized predictor is reported as an approach to solve the secondary structure prediction problem.

Accordingly, this paper focused on the development and implementation of a data mining technique for the extraction of protein sequence patterns. Specifically, the aim is the development of a data mining technique for association rule extraction (see figure 1). The focus is on the use of association rules as a method for extraction of secondary structure information from protein sequence. Therefore, instead of developing a secondary structure predictor, we explore the use of association rules as a basis to build good predictors. Thus, a framework to understand and study the association rules as a first step to build an accurate secondary predictor model is presented. Such questions have taken on increased practical significance with the realization that a lot of currently approaches to the secondary structure problem are associated with association rules or frequent items as a sequence to structure layer in the process.

Moreover some authors [1] believe that substantial improvement on the accuracy of secondary structure prediction methods can only be possible if better representations for the secondary structure features are found, instead of continuing applying different machine learning algorithms on the same set of profile-based features, which has shown to yield similar results in the past.

The rest of this paper is organized as follows. First, a short biological background necessary to understand the secondary structure prediction problem is presented. Then, the training dataset is described. Subsequently the proposed data mining approach is explained. Thus, the proposed data preprocessing, amino acid frequent pattern recognition and prediction models are described. Next the experimental framework and its results are discussed. Finally, some conclusions from this work are devised.

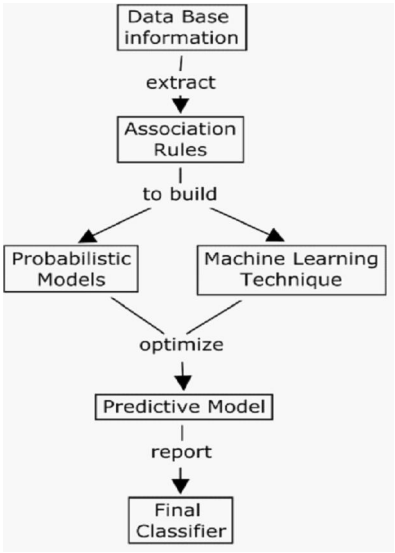

Fig. 1. General model of a structure -sequence layer of a predictor based on association rules.

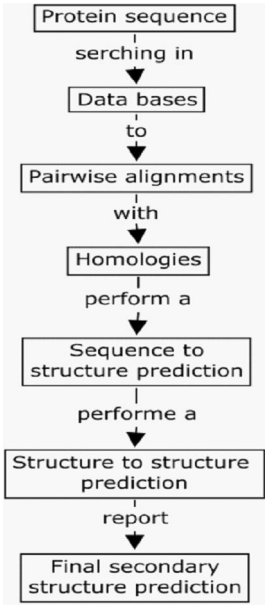

Fig. 2. General pipeline of a fourth generation secondary structure prediction approach.

II. BACKGROUND

Proteins are formed from one or more amino acid sequences in a folding process in which a three-dimensional structure is

obtained. This three-dimensional structure is highly important because it helps to determine its function of the protein. In order to understand the structure and formation of proteins, it is convenient to consider four structural levels. Primary structure consists in the order of the amino acids in the sequence. Secondary structure contains regular components such as α-helices, β -sheets and β -turns, where these types of structures contribute to the stabilization of protein folding. Tertiary structure where the elements of secondary structure are folded forming an almost solid compact structure that is stabilized by weak interactions. Quaternary structure consists of several polypeptides chains with tertiary structure that are joined by weak connections - non-covalent - to form a protein complex [16].

III. ASSOCIATION RULES IN PROTEIN SECONDARY STRUCTURE PREDICTION

The proposed approach is based on the application of a data mining procedure on a sequence data set to discover amino acid patterns in association rules that characterize protein secondary structure. Those patterns will be the first information source to build a machine learning technique to predict protein secondary structure. The analysis of this model will give some insights about the use of association rules as a technique to build secondary structure predictors. Additionally, a framework to extract information from a biological dataset based on association rules is proposed.

It is important to notice that no heuristic or biological information is taken into account and only the information give by the association is used as a basis for building a secondary structure predictor.

Analyzing figures 1 and 2, it is observed that the developed predictor does not perform completely the sequence to structure layer.

A DATA PREPROCESSING

Main biological databases have reliable and curated information that has been carefully found. The necessary guidelines and requirements to construct a data set for the training and verification of the methods of secondary structure prediction could be found in the literature [7, 8, 9, 13]. Based on these guidelines, some consensus criteria among all of them could be established; for example, it could be stated that protein sequence identity above 25%, structural homologues and transmembrane proteins should be avoided. On the other hand, well resolved crystal structures with a resolution better than 2.5 Å should be favored; in addition, the data set should be a representative subset of the known fold space.

In this work, the protein data set CB513 proposed by Cuff and Barton in 1999[13] and the data set SCOP-SFR developed by Birzele et al in December 2003 [14] are used. These protein data sets fit the requirements enounced as consensus criteria in the paragraph above; they also have the following characteristics. CB513 data set contains 513 protein chains with 84119 amino acids. The helix content is 34.5%, 22.7% sheets and 42.8% coils. The SCOP-SFR data set contains 940 protein chains with a total of 157813 residues distributed in 36.79% helices, 22.78% sheets and 40.42% coils.

Given that the amino acids of some common conformation as the β sheets can be quite distant from each other in the linear sequence, the creation of a window of interaction was necessary to analyze the interaction of amino acids in a sequence interval. The window size refers to the specific number of amino acids in the patterns that will be studied. In this work, a 20 amino acid window was generated; all the amino acids in the protein sequences of the data set were scanned using this window. In figure 3 an example showing the process followed to build a transaction table is depicted. In this example, for each amino acid of the sequence VLSEGEWQ five different transactions are created using a window of size four.

A binarization process on the data set was performed, given the fact that some algorithm that finds association patterns requires the data to be binary [17]. Then, a transaction table was built, where each amino acid *g* of each sequence VLSEGEWQ and a window of size four, creating five different sets.

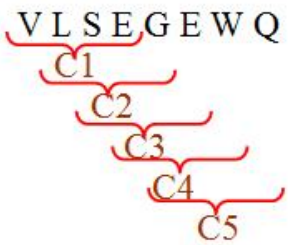

Fig. 3. Scanning of sequence VLSEGEWQ using a window of size 4.

B. FINDING FREQUENT PATTERNS

The existence of consecutive patterns in amino acid sequences could be useful in identifying important characteristics in function and structure; such features could be based on chemical or evolutionary properties.

Although some secondary structure prediction methods such as the ones developed in [18] and [19] identified frequent patterns in protein data sets to be associated with specific structural states. In this work, similarly to the work developed in [1] and [14], an algorithm to discover association rules called A priori is used [20] to search consecutive items and amino acid patterns in the data sets CB513 and SCOP-SFR. Therefore, in the proposed approach, an implementation of the A priori algorithm based on prefix trees was used to organize the counters in the item sets [20].

The main aim to perform an A priori search in the data set is to find a set of frequent words or N-grams that represent consecutive amino acids patterns of variable, with the objective

of applying this codified information in the development of a predictive model of secondary structure prediction.

There are two main challenges to face in the classic implementation of the A priori algorithm. The first one is defining a frequency and support scoring; the second one is to preserv the order and sequence of the frequent patterns found.

There are two different approaches to define the frequency of a pattern in a data set: the occurrence of a pattern in the data set and the number of sequences in which the pattern is found. Given that the goal here is similar to the work in [14], namely, using the frequent patterns to structurally classify a region or residue around a pattern, it is convenient to count the occurrence of a pattern as an independent event without taking in account its successive occurrence in the same sequence [1]. Then, the frequency could be defined as follows:

$$freq(p, D) = \sum_{s \in D} number\ of\ ocurrences\ of\ p\ in\ s. \quad (1)$$

where D represents the protein data set, p represents the pattern and s the amino acid sequence.

In order to guarantee the application of equation (1), it is necessary to use the window concept given in section 3.A. Accordingly, every sequence of amino acids is divided into windows of size 20. A pattern found in one of these windows will be present in the next i windows, where i is the position of the window where the first element of the pattern is found (see figure 4). Then, it is guaranteed that a pattern that could be found more than once in a sequence will be counted as an independent event in each occurrence.

Pos	1	2	3	4	5	6	7	8
Seq 1	V	L	S	E	G	E	W	Q
Seq 2	L	S	E	G	E	W	Q	V
Seq 3	S	E	G	E	W	Q	V	I
Seq 4	E	G	E	W	Q	V	I	A
Seq 5	G	E	W	Q	V	I	A	M
Seq 6	E	W	Q	V	I	A	M	F

Fig. 4. Consecutive amino acid sequences using a window of size 8.

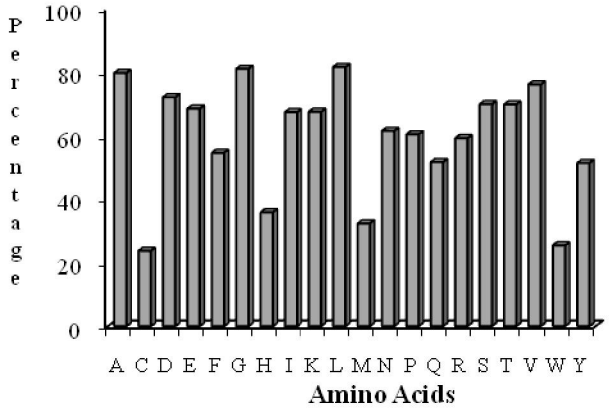

Fig. 5. Percentage of sequences in which each essential amino acid is present after a windows of size 20 is applied in the data set CB-513.

Based on the application of the A priori algorithm over the data set, it was possible to perform an exploratory analysis of the data; for example, in figure 6, a diagram showing the percentage of sequences in which each essential amino acid is present after a window of size 20 is applied in the data set CB-513 is shown. It is clear that the Leucine (L) is present in 81.7% of the sequences out of 74372 sequences derived from the data preprocessing with a window of size 20. On the other hand, the least frequent amino acid is Cysteine, present in 23.7% of the sequences.

C. ASSOCIATION RULES

Several machine learning models for protein secondary structure prediction have been proposed; particularly, neural networks [22,23], support vector machines [24,25,26] and hidden Markov models [27] have been successfully applied.

The aim of this research is, in principle, a bit different from typical prediction methods; in the next section, a simple association rule approach to classify residue chains in their secondary structures is performed.

In the application of association rules over the preprocessed data set, the A priori algorithm implemented in [20] was used. It was applied over 74372 pre-processed protein sequences of size 20 in CB513 and 139953 with similar features on SCOP-SFR

The transactional file generated in the application of the association rules follows the pattern shown in figure 6, in which 'Sequence' reports a sequence of the preprocessed data set, 'Class' reports the secondary conformation at which each amino acid of the sequence belongs, and 'Transaction' reports the file structure used as input to the a priori algorithm.

Position	1	2	3	4
Sequence	V	L	S	E
Class	α	α	β	β

Transactions: V α VL αα VLS ααβ VLSE ααββ

Fig. 6. Transactional sequence for association rule application

From figure 6, the generation of a transactional file requires the generation of new items. In the specific case of figure 6, the transaction has 8 items for a sequence of 4 amino acids (VLSE). In general, it is possible to say that using the presented pattern, the maximum number of items in a set of preprocessed sequences, with a size N window, could be defined by the generation of not repeated items formally expressed by equation (2).

$$I = 2a * ((\sum_{j=1}^{N} p_i) - 2Na) \quad (2)$$

where I is the maximum number of items in a transactional file, a is the window size, p_i is the size of protein i and N is the number of proteins in the data set. Equation 2 determines the maximum number of items in a transactional file with the features described

in this work. Even in a realistic experiment the maximum number will not be achieved, because one of the 20 amino acids will be repeated in a data set. This equation is very important because it proves the feasibility of the proposed approach, showing that the number of items in a transactional file will not be so high to make the computations unfeasible.

The association rules we are interested in have the following structure: the left side of the implication represents an amino acid sequence, and the right side represents the classes each amino acid of the sequence belongs to.

$$A_1 A_2 A_3 \ldots\ldots A_i \ldots\ldots A_{n-1} A_n \rightarrow C_1 C_2 C_3 \ldots\ldots C_i \ldots\ldots C_{n-1} C_n \quad (3)$$

In the experimental framework, a significant amount of experiments were run to test the performance of the developed models (section D). As a particular case, by running the algorithm with a minimum support equal to 0.1% and a confidence of 50%, 287 association rules, that satisfied the structural requirements previously defined, were generated.

D. PREDICTION MODELS

A prediction model based on the obtained association rules was developed. The use of association rules as a basis for secondary structure prediction is presented in the next section. A predictor model based on association rules is described in section C.2 and the details to develop a neural network model are presented in section C.3.

The proposed model was implemented in order to automate the training and prediction phase of the model.

D.1 An Evaluation Patterns

The evaluation pattern is based on the hypothesis that the classification of an amino acid in its secondary structure should depend on the interactions of such amino acid with other amino acids in the protein chain. Then, it is important to generate a model that takes into account the patterns found from the association rules.

Figures 7 shows all the possible combinations with a maximum size of 3, in which amino acid E situated in position 4 is involved in an interaction with other neighbor amino acids.

An analysis model was generated, in which, for every amino acid in a sequence, all possible combinations of the interactions between this residue and the other amino acids were evaluated. The number of combinations depends on the window size and the highest size of any association rule generated by the A priori algorithm.

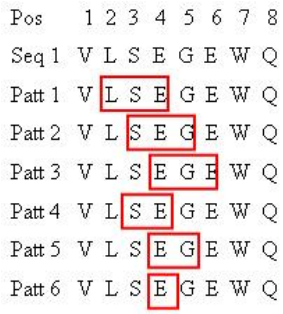

Fig. 7. Evaluation patterns for amino acid E situated in position 4

D.2 Association Rule Model

An association rule-based model to secondary structure prediction based on two main features was developed. The first feature of the model in the training phase is redundancy elimination and the second one is an effective indexing of ARs. These features are characterized by the use of a hash table. The key will be the association rule antecedent, and the value will be the association rule consequent concatenated with the confidence ζ of the implication, see figure 8.

In the verification phase the proposed model was used as a predictor based on the following four steps i) Reception of a query sequence ii) Pattern evaluation iii) Verifying the existence of the patterns in the hash table iv) Decision making using a voting system v) Secondary structure prediction, where steps from ii) to v) constitute an iterative process over all the amino acids of the query sequence.

Step ii) returns a set of patterns corresponding to all the combinations, where the studied amino acid has an interaction given a size window. Then the existence of each one of these patterns is studied in the hash table; if it exists, a voting system will be used to accumulate the contribution of each pattern given the associate confidence. Finally, the contributions of each class are accumulated and the class with the highest contribution is reported.

KEY	VALUE
$A_{1,1}A_{1,2}A_{1,3}\ldots\ldots A_{1,j}\ldots\ldots A_{1,n-1}A_{1,n}$	$C_{1,1}C_{1,2}\ldots C_{1,j}\ldots C_{1,n-1}C_{1,n}+\zeta$
$A_{2,1}A_{2,2}A_{2,3}\ldots\ldots A_{2,j}\ldots\ldots A_{2,k-1}A_{2,k}$	$C_{2,1}C_{2,2}\ldots C_{2,j}\ldots C_{2,k-1}C_{2,k}+\zeta$
.....	.....
$A_{i,1}A_{i,2}A_{i,3}\ldots\ldots A_{i,j}\ldots\ldots A_{i,t-1}A_{i,t}$	$C_{i,1}C_{i,2}\ldots C_{i,j}\ldots C_{i,t-1}C_{i,t}+\zeta$
.....	.....
$A_{p-1,1}A_{p-1,2}\ldots\ldots A_{p-1,j}\ldots\ldots A_{p-1,n}$	$C_{p-1,1}C_{p-1,2}\ldots C_{p-1,j}\ldots C_{p-1,n}+\zeta$
$A_{p,1}A_{p,2}A_{p,3}\ldots\ldots A_{p,j}\ldots\ldots A_{p,n-1}A_{p,n}$	$C_{p,1}C_{p,2}\ldots C_{p,j}\ldots C_{p,n-1}C_{p,n}+\zeta$

Fig. 8. Indexing hash table

D.3 Neural Network Model

A prediction model based on a neural network was developed. The architecture of the neural network is as follows: 60 neurons in the input layer, where each set of 20 consecutive neurons corresponds to an amino acid representation after performing a

binarization process, a variable number of nodes in the hidden layer, and 9 neurons in the output layer, where each set of 3 consecutive neurons corresponds to the classes of input amino acids (see figure 9).

The neural network was developed to work with association rules with a maximum size of three amino acids for the evaluation pattern process (section D.1).

The verification phase of the proposed model is based on the following steps. i) Reception of a query sequence ii) Pattern Evaluation iii) Obtaining results by the neural network iv) Decision making using a voting system v) Secondary structure prediction, where steps from ii) to v) constitute an iterative loop over all the amino acids in the query sequence.

Step ii) returns a set of patterns corresponding to all the combinations, where the studied amino acid has an interaction given a size window. For these patterns, the neural network is used to get a prediction and a voting system is used to accumulate the individual contribution of each pattern. Finally, the contributions of each class are accumulated and the class with the highest contribution is reported.

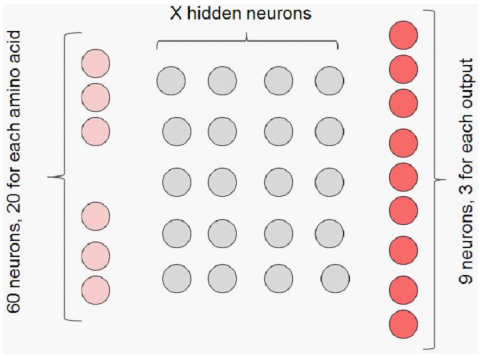

Fig. 9. Architecture of the developed neural network

IV. EXPERIMENTATION

The main goals of the experiments that were carried out were: *i)* To evaluate the association rules as a method for extraction of secondary structure information from protein sequence in order to build a sequence - structure layer; *ii)* To analyze and compare the results obtained using the CB513 data set and the SCOP-SFR data sets in a prediction experiment; *iii)* To clarify the limitations and advantages of using association rules in secondary structure prediction *iv)* To study the behavior of simple secondary structure predictors based on association rules.

A.1 Experimental framework

In order to accomplish such objectives, a set of experiments were carried out; eight of them are reported in table I. Id identifies a specific running experiment. The parameters of each experiment are as follows:

Id 1: A support of 0.2% and a confidence of 50% were used as the A priori algorithm parameters, 820 association rules extracted from CB513 were used, and the models were tested on the data set CB513.

Id 2: A support of 0.2% and a confidence of 50% were used as the A priori algorithm parameters, 820 association rules extracted from CB513 were used, and the models were tested on the data set SCOP-SFR.

Id 3: A support of 0.2% and a confidence of 50% were used as the A priori algorithm parameters, 424 association rules extracted from SCOP-SFR were used, and the models were tested on the data set CB513.

Id 4: A support of 0.2% and a confidence of 50% were used as the A priori algorithm parameters, 424 association rules extracted from SCOP-SFR were used, and the models were tested on the data set SCOP-SFR.

Id 5: A support of 0.2% and a confidence of 50% were used as the A priori algorithm parameters, 1244 association rules extracted from CB513 and SCOP-SFR were used, and the models were tested on the data set CB513.

Id 6: A support of 0.2% and a confidence of 50% were used as the A priori algorithm parameters, 1244 association rules extracted from CB513 and SCOP-SFR were used, and the models were tested on the data set SCOP-SFR.

Id 7: A support of 0.2% and a confidence of 25% were used as the A priori algorithm parameters, 3643 association rules extracted from CB513 and SCOP-SFR were used, and the models were tested on the data set CB513.

Id 8: A support of 0.2% and a confidence of 25% were used as the A priori algorithm parameters, 3643 association rules extracted from CB513 and SCOP-SFR were used, and the models were tested on the data set SCOP-SFR.

Additionally, A_0 represents the accuracy of the studied model given a confusion matrix and A1 represents the accuracy of the model, only taking into account classes α and C.

A.2 Experimental results

Analyzing the results of the experiments, it can be stated that the models have the characteristics of secondary structure prediction models belonging to the first and second generations. In this work, experimentally the results of Rost and Sander [29] were proved. Specifically, even though great efforts were made to increase the prediction accuracy of the association rule methods, they will only reach approximately 65% prediction accuracy. Methods studied in this paper have a prediction accuracy around 53%, which could be improved adding some characteristics used in first and second generation methods. The main problem of using association rules is the poor accuracy of β-sheet prediction, which is slightly better than random. Another problem was the short number of predicted helix and sheet elements represented by association rules; this is understood by the fact that a short number of association rules were generated with respect to the possible combinations in an alphabet of 20 amino acids.

TABLE I
EXPERIMENTAL RESULTS

ID	ASSOCIATION RULES (AR)	NEURAL NETWORKS	AR + NN	NN ERROR TRAINING
1	C β α C | 25182 1068 9743 β | 10082 3277 5700 α | 11178 1207 16682 A_0= 53.66% A_1= 64.34%	C β α C | 6 1824 34163 β | 0 3585 15474 α | 1 1561 27505 A_0= 36.66%	C β α C | 25182 745 10066 β | 10082 2416 6561 α | 11178 994 16895 A_0= 52.89% A_1= 64.67%	
2	C β α C | 41588 1872 17040 β | 17654 5693 11000 α | 20303 2224 32285 A_0= 53.16% A_1= 64.06%	C β α C | 14 3076 57410 β | 0 6482 27865 α | 0 2863 51949 A_0= 39%	C β α C | 41588 1259 17653 β | 17654 4374 12319 α | 20303 1822 32687 A_0= 52.55% A_1= 64.41%	
3	C β α C | 26253 806 8934 β | 11293 2495 5271 α | 12763 1047 15257 A_0= 52.31% A_1= 63.8%	C β α C | 0 857 35136 β | 0 2286 16773 α | 0 901 28166 A_0= 36.2%	C β α C | 26253 459 9281 β | 11293 1661 6105 α | 12763 674 15630 A_0= 51.76% A_1= 64.37%	
4	C β α C | 43667 1314 15519 β | 19653 4774 9920 α | 22639 1807 30366 A_0= 52.65% A_1= 64.2%	C β α C | 0 1479 59024 β | 0 4086 30261 α | 0 1590 53222 A_0= 38.29%	C β α C | 43667 763 16070 β | 19653 3079 11615 α | 22639 1156 31017 A_0= 51.96% A_1= 64.76%	
5	C β α C | 25183 1101 9709 β | 10082 3343 5634 α | 11178 1215 16674 A_0= 53.73% A_1= 64.33%	C β α C | 1 1921 34071 β | 0 3740 15319 α | 1 1651 27415 A_0= 37.03%	C β α C | 25183 766 10044 β | 10082 2490 6487 α | 11178 1047 16842 A_0= 52.91% A_1= 64.59%	
6	C β α C | 41589 1902 17009 β | 17654 5826 10867 α | 20303 2249 32260 A_0= 53.23% A_1= 64.04%	C β α C | 0 3244 57256 β | 0 6732 27615 α | 0 3015 51797 A_0= 39.10%	C β α C | 41589 1306 17605 β | 17654 4499 12194 α | 20303 1912 32597 A_0= 52.57% A_1= 64.33%	
7	C β α C | 5771 4632 25590 β | 881 5978 12200 α | 933 2682 25452 A_0= 44.22% A_1= 47.99%	C β α C | 5519 3418 27056 β | 840 4914 13305 α | 781 1983 26303 A_0= 43.67%	C β α C | 7407 3196 25390 β | 1175 4841 13043 α | 1177 1933 25957 A_0= 45.41% A_1= 51.28%	
8	C β α C | 9688 7513 43299 β | 1658 10220 22469 α | 1724 4942 48146 A_0= 45.47% A_1= 50.15%	C β α C | 9322 5545 45633 β | 1555 8462 24330 α | 1339 3558 49915 A_0= 45.23%	C β α C | 12423 5192 42885 β | 2223 8283 23841 α | 2138 3475 49199 A_0= 46.7% A_1= 53.43%	

The experiments carried out using the two data sets produced similar results (Id 1 to 6). It is important to mention that the SCOP-SFR data set was reported more recently than CB513.

Association rules are a good methodology to extract structure information from a protein data set to build an accuracy predictor, because from the experiments it can be stated that the association rules keep general information from a set of data representing the known fold space. In experiments 1 to 6, the accuracy is almost the same, even though different data sets were used in the training process. Moreover, comparing the results of experiments 5 and 6, with experiments 1 to 4, it is clear that the amount of association rules does not determine the accuracy.

Association rules based models could be sensitive to the number of redundant information, for example, in experiments 7 and 8, the neural network training error is higher than 0.5. Even if the application of filters to avoid redundant information is an easy process, the definitions of the parameters of good association rules are difficult to get.

The extraction of secondary structure information from protein sequence using association rules could be thought of as independent of the data set, if such data set represents a known fold space and it does not produce redundant association rules.

V. CONCLUSIONS

In this work, a data mining technique for association rule extraction was developed. The focus is on the use of association rules as a method for extraction of secondary structure information from protein sequence. Despite of the limitations of association rules as predictive methods, they are a significant source of information for extraction of secondary structure information from protein sequence in order to build a sequence — structure layer. This has been shown in different studies, where accuracy prediction methods have been developed based on frequent patterns as part of a sequence — structure layer. Additionally, it is important to mention that association rules give some insights about secondary structure prediction features to be used in learning algorithms.

The data mining methodology developed in this research is feasible and useful in the exploration of information from protein sequence.

The use of hash tables provides an excellent computational technique to model association rules, because the number of collisions is reduced to zero, it avoids the data redundancy and the insertion; in addition, erasing and search of association rules is performed efficiently.

The fixed size sliding window to study association rules is a limitation of these secondary structure prediction methods, but it highly decreases the computational resources to perform an A priori algorithm.

In this work, the problems with first and second generation methods are experimentally explored, clarifying the advantages and limitations of using association rules. As a conclusion, association rules are good to support secondary structure prediction methods, but they are limited predictors by themselves.

Future work will focus on the building of a four generation predictor with a sequence-structure layer based on association rules to experimentally evaluate the contribution of this approach to the accuracy of protein secondary structure prediction.

REFERENCES

[1] F Birzele. Data Mining for Protein Secondary Structure Prediction, Institut für Informatik XII, Jan 2005

[2] Chou P. Y. and Fasman G. D. (1974) Prediction of protein conformation, *Biochemistry*, 13, 211-245

[3] Granier J. Osguthorpe D. J. and Robson B. (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.*, 120, 97-120

[4] Rooman M. J. and Wodak S. J. (1991) Weak Correlation between Predictive Power of Individual Sequence Patterns and Overall Prediction Accuracy in Proteins. *Proteins*, 9, 69-78.

[5] Strelets V. B. (1995) New Machine Learning Technique for Analysis and Prediction of Sequence and Structure Features: Protein Secondary Structure Prediction.

[6] Frishman D. and Argos P. (1996) Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. *Protein Engineering*, 2, 133-142

[7] Zvelebil M. J., Barton G. J., Taylor W. R. and Sternberg M. J. (1987) Prediction of protein secondary structure and active sites using the alignment of homologous sequences, *J. Mol. Biol.*, 195, 957-961

[8] Rost B. and Sander C. (1993) Prediction of Protein Secondary Structure at Better than 70% Accuracy. *J. Mol. Biol.*, 232, 584-599

[9] Jones D. T. (1999) Protein Secondary Structure Prediction Based on Position specific Scoring Matrices. *J. Mol. Biol.*, 292, 195-202

[10] Meiler J and Baker D. (2003) Coupled prediction of protein secondary and tertiary structure, *Proc. Natl. Acad. Sci.*, 21, 12105-12110

[11] Rost B. and Eyrich V. (2001) EVA: large - scale anaylsis of secondary structure prediction, *Proteins*, 45, 192-199S.

[12] J. Martin, J-F Gibrat, F. Rodolphe, Choosing the Optimal Hidden Markov Model for Secondary-Structure Prediction, IEEE Intelligent Systems Vol 20 No 6, pp 19-25, 2005.

[13] Cuff, J. A. and Barton G. J. (1999) Evaluation and Improvement of Multiple Sequence Methods for Protein Secondary Structure Prediction. *Proteins*, 34, 508-519

[14] F. Birzele, and S. Kramer, A new representation for protein secondary structure prediction based on frequent patterns. Oxford Journals, Vol 22 No 21, Pp 2628-2634, 2006

[15] D. L. Nelson and M. Cox. Lehninger. Principles of Biochesmistry, Palgrave Macmillan, Chapter 3, 2004.

[16] A. Sali, E. Shakhnovich, M. Karplus, Kinetics of Protein Folding: A lattice model study of the requirements for folding to the native state, J. Mol. Biol., 235, 1614-1636, 1994.

[17] P-N Tan, M Steinbach, V Kumar, Introduction to Data Mining, Pearson Addison Wesley, Chapter 2, 2006.

[18] Rooman M. J. and Wodak S. J. (1991) Weak Correlation between Predictive Power of Individual Sequence Patterns and Overall

Prediction Accuracy in Proteins. *Proteins*, 9, 69-78

[19] Strelets V. B. (1995) New Machine Learning Technique for Analysis and Prediction of Sequence and Structure Features: Protein Secondary Structure Prediction.

[20] Agrawal R, Imielinski T, Swami AN. "Mining Association Rules between Sets of Items in Large Databases." SIGMOD. June 1993, 22(2):207-16

[21] Borgelt C. Efficient Implementations of Apriori and Eclat, 1st Workshop of Frequent Item Set Mining Implementations (FIMI 2003, Melbourne, FL, USA).

[22] Jones D: Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* 1999, 292(2):195-202.

[23] Przybylski D, Rost B: Alignments grow, secondary structure prediction improves. *Proteins* 2002, 46(2):197-205.

[24] Ward J, McGuffin L, Buxton B, Jones D: Secondary structure prediction with support vector machines. *Bioinformatics* 2003, 19(13):1650-5.

[25] Kim H, Park H: Protein secondary structure prediction based on an improved support vector machines approach. *Protein Eng* 2003, 16(8):553-560.

[26] Hu HJ, Pan Y, Harrison R, Tai PC: Improved protein secondary structure prediction using support vector machine with a new encoding scheme and an advanced tertiary classifier. *IEEE Trans Nanobioscience* 2004, 3(4):265-271.

[27] JMartin, J Gibrat and F Rodolphe, Analysis of an optimal hidden Markov model for secondary structure prediction, *BMC Structural Biology* 2006, 6:25.

[28] Koh I, Eyrich V, Marti-Renom M, Przybylski D, Madhusudhan M, Eswar N, Grana O, Pazos F, Valencia A, Sali A, Rost B: EVA: evaluation of protein structure prediction servers. *Nucleic Acids Res* 2003, 31(13):3311-5

David Becerra. Earned his bachelor's degree in computer science and engineering from the National University of Colombia at Bogota, and is currently a master student in computer science at the same university.

Giovanni Cantor. Earned his bachelor's degree in computer science and engineering from the National University of Colombia at Bogota, and is currently a master student in computer science at the same university.

Luis F Niño. B.Sc. Computer Systems Engineering and M.Sc. Mathematics at the National University of Colombia at Bogota. He earned a M.Sc. degree and a Ph.D degree in Mathematics with concentration in Computer Science from the University of Memphis, USA. He is currently an Associate Professor of the Systems and Computer Engineering Department at the National University of Colombia.

Jonatan Gomez. Computer Engineeer awarded by the Universidad Nacional de Colombia. Master of Sciences (Mathematics) awarded by the Universidad Nacional de Colombia and Master of Sciences (Mathematics) concentration in Computer Sciences awarded by the University of Memphis (USA). Ph. D. (Mathematics) concentration in Computer Science awarded by the University of Memphis (USA).

Leonardo Bobadilla. Earned his bachelor's degree in computer science and engineering and his master degree in statistics from the National University of Colombia at Bogota; currently he is a Phd student in computer science at the University of Illinois at Urbana-Champaign.

Universidad Nacional de Colombia Sede Medellín
Facultad de Minas

Escuela de Ingeniería de Sistemas

Pregrado

- Ingeniería de Sistemas e Informática.

Posgrado

- Doctorado en Ingeniería-Sistemas.
- Maestría en Ingeniería de Sistemas.
- Especialización en Sistemas con énfasis en:
  - Ingeniería de Software.
  - Investigación de Operaciones.
  - Inteligencia Artificial.
- Especialización en Mercados de Energía.

Áreas de Investigación

- Ingeniería de Software.
- Investigación de Operaciones.
- Inteligencia Artificial.

Escuela de Ingeniería de Sistemas
Dirección Postal:
Carrera 80 No. 65 - 223 Bloque M8A
Facultad de Minas. Medellín - Colombia
Tel: (574) 4255350 Fax: (574) 4255365
Email: esistema@unalmed.edu.co
http://pisis.unalmed.edu.co/