AUTOMATED ANNOTATION OF KEYWORDS FOR PROTEINS OF THE NEWCASTLE VIRUS DISEASE

by

ANTOINE RABABY

M.S., Computer Science, Lebanese American University, 2007

Thesis submitted in partial fulfillment of the requirements for the Degree of Master of Science in Computer Science

Division of Computer Science and Mathematics

LEBANESE AMERICAN UNIVERSITY

June 2007

LEBANESE AMERICAN UNIVERSITY

School of *Arts* and Sciences

Thesis Approval

I.D.#: 199435810

Student Name	**ANTOINE RABABY**
Thesis Title:	***AUTOMATED ANNOTATION OF KEYWORDS FOR PROTEINS OF THE NEWCASTLE VIRUS DISEASE***
Program:	**Computer Science**
Division /Dept:	**Computer Science and Mathematics**
School:	**School of Arts and Sciences, Byblos**
Approved by:	
Thesis Advisor:	***DANIELLE AZAR*** [redacted]
Member	***Haidar Harmanani*** [redacted]
Member	***CHADI NOUR*** [redacted]
Member	
Date:	JUNE 12, 2007

Plagiarism Policy Compliance Statement

I certify that I have read and understood LAU's Plagiarism Policy. I understand that failure to comply with this Policy can lead to academic and disciplinary actions against me.

This work is substantially my own, and to the extent that any part of this work is not my own I have indicated that by acknowledging its sources.

Name: Antoine Rababy

Signature: [redacted]

Date: 22/06/2007

To my parents

Acknowledgment

First of all, I would like to thank Dr. Danielle Azar who is the supervisor of my masters thesis at LAU. She did a very good job supervising me and she contributed to my thesis with many good suggestions. I also want to thank Dr. Haidar Harmanani, who is the chairperson of the department of computer science at the Lebanese American University and Dr. Chadi Nour who is an assistant professor of mathematics at LAU for accepting of being members in my thesis committee.

I would also like to thank all the people who supported me by answering my questions, among those are Mr. Gregoire Rossier and Ms Elisabeth Gasteiger at the Swiss Institute of Bioinformatics as well as Dr. Alexander Kanapin and Ms Daniela Wieser from The European Bioinformatics Institute (EBI).

I would like to express my sincere gratitude to the Lebanese American University whose financial support during my graduate studies made it all possible.

Finally, a very special thank you to my parents and especially to my mother who was the first person to encourage me to apply for graduate studies in computer science. I quote her: “Start your masters thesis in computer science now that you can, you will finish it without even knowing that you have started it. Nobody knows what the future holds”.

Abstract

The number of newly discovered proteins has increased drastically during the last two decades. Curators are no longer capable of manually annotating them. Therefore there is a great need to automate this process. Rule generation for protein annotation in databases such as Uniprot, Prosite, Interpro has been tackled by many scientists and researchers and has proven to be a reliable and successful method for correctly and accurately annotating proteins regarding certain fields (for example the keywords field). Our study of the organism "Newcastle Virus Disease" showed that data coming from Swiss-Prot was accurate (checked by human experts) while data coming from TrEMBL is not reliable and incomplete. We propose to automate the process of annotating proteins related to the Newcastle virus disease regarding their keywords field in both the Swiss-Prot and TrEMBL database. The rules generated have been applied to most of the proteins from SwissProt database and the results were promising. As a matter of fact 95% of the proteins were accurately annotated with the exact keyword(s). As for TrEMBL database our rules have annotated the proteins which were originally unannotated and improved or completed the annotation of proteins for which annotation was incomplete. These obtained results were again tested against the data in SwissProt and were found to be between 90% and 100% valid and correct.

Contents

List of Figures

List of Tables

Chapter1

Biological Background

1.1 Introduction

All living organisms are made up of units called cells. They start with at least one cell and the remaining cells are generated from previous ones.

It is widely believed that all cells are descendants from the first cell which existed 4.2 billion years ago [15]. The human body starts from one cell and contains at maturity approximately one hundred thousand billion cells.

There are two basic types of cells: Prokaryotic and eukaryotic. Prokaryotic cells are small, primitive cells without organelles (more about this in sec.1.2) examples are bacteria and algae, while Eukaryotic cells are larger, more advanced and contain organelles. Examples are humans, animals, fungi, etc .

1.2 Cell structure

All cells consist of three parts: the cell membrane, the nucleus and the cytoplasm. Cytoplasm is a special fluid containing special organelles, "which are discrete structures of a cell having specialized functions" [18]. Examples or organelles are: Mitochondrion which is responsible of energy production, ribosome which is responsible for translation of RNA into proteins, Golgi Apparatus which is responsible for sorting and modification of proteins, Lysosome which is responsible for breaking down large molecules and the Endoplasmic Reticulum which is responsible for modifying and folding new proteins [18].

The cell membrane separates the outside of the cell (the extracellular) from the inside of the cell (the intracellular) [15]. It is responsible for maintaining the integrity of the cell and controlling the passage of material into and out of the cell.

The nucleus is formed of a nuclear membrane around a fluid nucleoplasm. It is the control center of the cell. Chromatin inside the nucleus contains DeoxyriboNucleic Acid (DNA), the genetic material of the cell. The nucleolus is responsible for fabricating ribosomes and contains ribonucleic acid (RNA). The nucleus controls the structure and functioning of the cell (Fig.1 [15]).

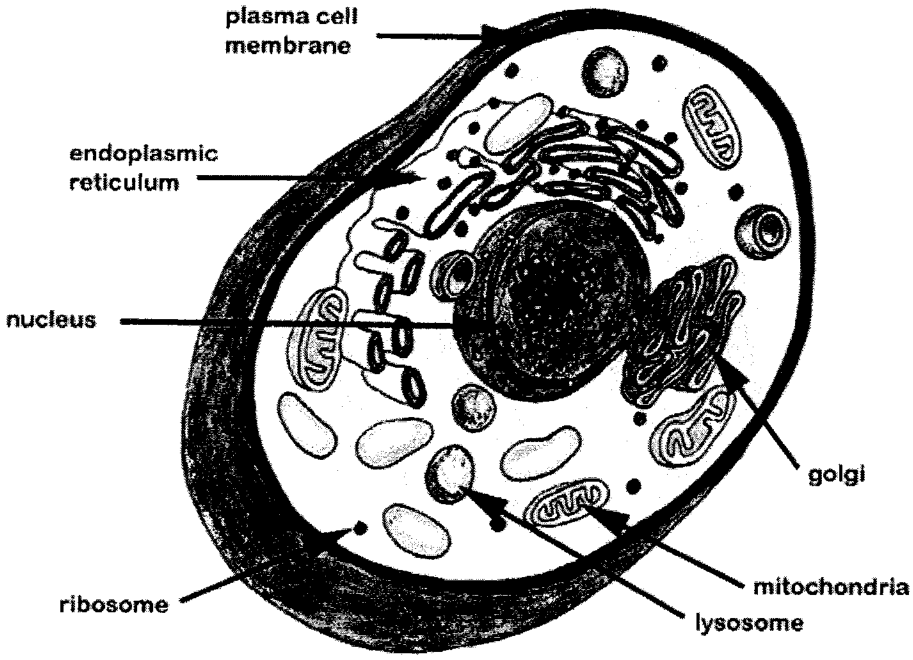

Figure 1: Cell Structure [15]

1.3 Cell function

As explained earlier, the nucleus of the cell contains the big double stranded helix molecule of heredity: the DNA. When a cell divides into two new cells, the DNA is duplicated via *Mitosis* where each new cell receives an exact copy of the DNA of the parent cell [15]. During cell division, Chromatin is organized into chromosomes. Each chromosome contains a DNA molecule composed of many genes, which are individual segments of DNA that contain the instructions necessary for proteins synthesis.

The number of chromosomes differs from one organism to another. Humans for example have 46 chromosomes organized in 23 pairs, whereas a frog has 26 chromosomes (13 pairs). Each member of the pair comes from one parent, so both members may be the same size or shape but definitely do not carry the exact same information. This is why a child looks pretty much like his father and his mother. The 23 pairs of human chromosomes are estimated to include about 10,000 genes and each gene codes for one protein. Cells transport food and oxygen and they are limited in size therefore when the capacity of food and oxygen to be transported by existing cells is exceeded, the division process takes place in order to compensate. Thus more cells are produced which induces more food and oxygen transportation [21].

As stated earlier, all the genetic information is transmitted from the parent to the child cell. The two nucleic acids responsible for accomplishing this task are DNA which stores genetic information and RNA which allows that information to be made use of in the cell. There are four nitrogen bases found in DNA: Adenine, Cytosine, Guanine and Thymine and four in RNA: Adenine, Cytosine, Guanine and Uracil [16].

Table1 shows an example of a DNA and RNA sequence.

A DNA sequence: A-C-T-G-G-A-C-A-T-G ……
An RNA sequence: A-C-U-G-G-A-C-A-U-G …….

Table 1: examples of DNA and RNA sequences.

Each human cell has 46 molecules of DNA and each of these molecules is made of 50 to 250 million bases housed in a chromosome.

The DNA inside a chromosome is formed of genes. A gene is any defined part of the DNA containing coded information of this DNA that allows the cell to produce new proteins. There are approximately between 50,000 and 100,000 genes inside each chromosome and each gene contains 20,000 to 250,000 chemical bases [13].

In a DNA sequence every three consecutive bases form a codon and every codon either codes for an amino acid or is a stopping codon. For example codon TGT codes for the amino acid Cysteine while codon TAA has a different biological meaning like "STOP Transcription HERE". For example the sequence CCCTGTGGAGCCACACCCTAG…. is coded into the following part of a protein: Proline-Cysteine-Glycine-Alanine-Threonine-Proline [17]. Table1 shows all the amino acids and their physicochemical properties. These properties are:

- Mass which is the atomic mass unit of each amino acid,

- pI which is the isoelectric point, it is the Ph (measure of the acidity or alkalinity of a solution) at which a molecule carries no net electric charge[2] [18].
- pK1, which is the dissociation constant that refers to the carboxyl (-COOH) group of the amino acid [25].
- pK2, which is the dissociation constant that refers to the amino (-NH3) group of the amino acid [25].

Proteins are made of amino acids and they are the building blocks of our body; they make new cells and destroy old ones, they break down food to release energy. The process of manufacturing proteins can be visualized in Figure 1 and can be summarized as follows:

1 REPLICATION: DNA is duplicated before a cell divides.

2 TRANSCRIPTION: when proteins are needed the corresponding genes are transcribed in RNA.

3 PROCESSING: RNA is first processed so that non-coding parts are removed.

4 TRANSPORTATION: RNA is transported out of the nucleus.

5 TRANSLATION: finally proteins are built based on the code in the RNA.

At the end of this laborious process, proteins are created.

[2] The molecule is neutral it carries no positive or negative charges.

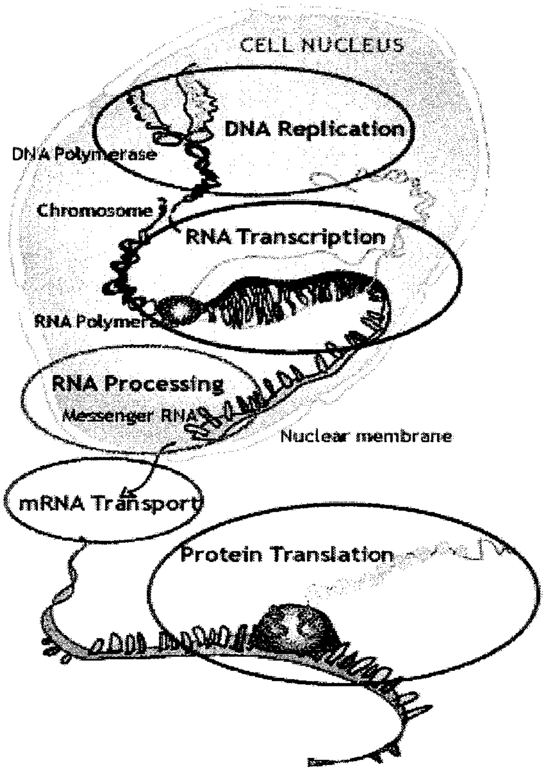

Figure 2: Proteins Fabrication Process [14].

Amino Acid	**Abbrev.**	**Mass**	**pI**	**pK_1 (α-COOH)**	**pK_2 (α-$^{+}NH_3$)**
Alanine	A	89.09404	6.01	2.35	9.87
Cysteine	C	121.15404	5.05	1.92	10.70
Aspartic acid	D	133.10384	2.85	1.99	9.90

Glutamic acid	E	147.13074	3.15	2.10	9.47
Phenylalanine	F	165.19184	5.49	2.20	9.31
Glycine	G	75.06714	6.06	2.35	9.78
Histidine	H	155.15634	7.60	1.80	9.33
Isoleucine	I	131.17464	6.05	2.32	9.76
Lysine	K	146.18934	9.60	2.16	9.06
Leucine	L	131.17464	6.01	2.33	9.74
Methionine	M	149.20784	5.74	2.13	9.28
Asparagine	N	132.11904	5.41	2.14	8.72
Proline	P	115.13194	6.30	1.95	10.64
Glutamine	Q	146.14594	5.65	2.17	9.13
Arginine	R	174.20274	10.76	1.82	8.99
Serine	S	105.09344	5.68	2.19	9.21
Threonine	T	119.12034	5.60	2.09	9.10
Selenocysteine	U	169.06			
Valine	V	117.14784	6.00	2.39	9.74

Tryptophan	W	204.22844	5.89	2.46	9.41
Tyrosine	Y	181.19124	5.64	2.20	9.21

Table 1: list of amino acids and their chemical properties

1.4 Proteins structure and functions

It is very important for scientists to be able to predict the protein's function from its sequence and especially from its 3D structure because proteins perform specific biochemical functions according to their amino acids sequence, which determines the unique 3D structure of each protein [23].

When a new protein sequence is discovered, the next thing to do is to compare it with already discovered sequences. Since genes are made of sequences of proteins scientists might find similarities between a newly discovered gene and one we know more about. Some of these similarities are listed below.

1. Genes may share high sequence similarity across their entire length.
2. Genes may show sequence similarity that is limited to a certain region. For example the protein encoded by the gene may share a well characterized DNA-binding domain with other proteins, while other parts are different.

3. Genes may share similar motifs, also known as common amino acid sequences whose folded structure is known. It is the case of Zinc fingers and Leucine zippers shown in Figure 3 below. Sequences residing between motifs can differ greatly from one protein to another, and the folded structure of these areas can be unknown, yet the known motif will fold into similar shapes [13].

All the above information helps to identify the new protein's physicochemical properties.

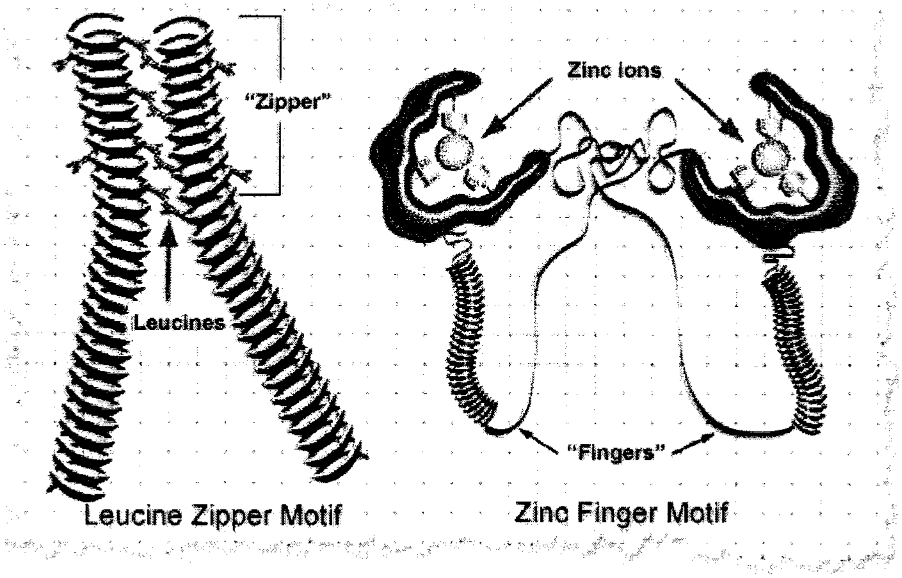

Figure 3: Zinc-Finger and Leucine zipper motifs

In case the new gene shares no similarities with any other known gene, it will be classified as "unique." Background information about a wide variety of known proteins helps understanding new ones.

When a protein is listed in the Swiss-Prot database, many relevant information such as its name, origin, amino acid sequence and physicochemical properties keywords are entered in the database if they are known and if not the field is left blank. Proteins are stored in the database according to many criteria such as: taxonomy, name, sequence similarity and keywords. The Keywords field in the database provides a way to connect a protein to the database by considering its biological, physical and chemical properties. We are concerned with the keywords field, because it is our main link between the biological information of a protein and its best fit place in the database. It stores this protein in its best fit place in the database. So If two or more proteins have a high matching percentage of keywords than this is a very good reason to state that these proteins share similar biological and physico-chemical properties and therefore they should belong to same category in the database.

Chapter 2

Related Work

The literature abounds with work that aims at annotating proteins automatically.

In [1] the authors generate rules to automatically annotate proteins in the SwissProt database regarding their keywords, the annotation is based on whether they belong to a certain InterPro family or not and to the taxonomy of the organism in which the protein was found. However, the technique used in this work results in rules that have a higher accuracy when used on small databases. Larger databases induced an increase in the number of rules, hence a higher number of conflicting rules and thus lower overall accuracy.

In [2] the authors have built a web application called: MineBlast,. This is a web service for literature search and presentation.

In [3] the authors present and describe WEKA a data mining tool. Rules can be generated using Id3 and J48 algorithms [20].

In [4] the authors have developed a system called Xanthippe, based on a simple exclusion mechanism and a decision tree approach using C4.5 algorithm (a descendant of Id3). It automatically generates annotation on proteins in Uniprot containing erroneous data. The system automatically flags and detects a large portion of this erroneous data, therefore increasing its accuracy.

In [5] the author presents a knowledge system for protein function annotation called: RuleMiner. The information is retrieved from Swissprot and protein family based sequence classification databases. Rules generated are based on sequence conservation, motifs and domains of proteins.

In [6] the authors introduce the Swiss-prot database which is an annotated protein sequence database created at the department of Medical biochemistry at the university of Geneva in Switzerland and assisted by the European Molecular Biology Laboratory (EMBL) since 1987. This database is divided in two parts: The core data and the Annotation. The authors also present the computer supplement for Swiss-Prot: TrEMBL which is used to accelerate the process of annotation. They explain in details its structure and its relation with other databases. Similarities to information (such as sequence similarities) in other databases such as PFAM, INTEPRO, PROSITE can be accessed from Swiss-prot.

In [7] the authors present the Universal Protein Resource (Uniprot), which is a centralized resource for protein sequences and functional information. It is created by uniting the Swissprot, TrEMBL and PIR (Protein Information Resource) [25]. The Uniprot knowledgebase is a comprehensive, fully classified, annotated protein sequence knowledgebase with many cross references. It consists of two major parts: TheUniprot/Swissprot which is a fully, manually annotated database. TrEMBL (translated EMBL), which is a very large protein database generated by computer translation of the genetic information from the EMBL[2] database." [18].

After the year 2004, a big effort was done in order to use automatic annotation with minimum human interaction.

[2] It is a molecular biology research institution supported by 19 European countries.

In [8] the authors describe and explain the importance of using the INTERPRO database in automatic protein annotation and genome analysis. In automatic protein annotation INTERPRO has provided accurate characterization of sequences which are candidates for functional annotation which "consists in attaching biological information to genomic elements." [18]. The rules based on InterPro characterization are stored in a RuleBase database. This process is applied on unannotated sequences which are then stored in the TrEMBL protein sequence database. INTERPRO is also used for comparative and statistical analysis on whole genomes. It has been extremely useful in proteome analysis therefore complementing the information in the CluSTr which is: "a database offering an automatic classification of Uniprot Knowledgebase proteins into groups of related proteins." [24].

In [9] the authors have developed the High-quality Automated and Manual Annotation (HAMAP) of microbial proteomes[3], which aims at integrating manual and automatic annotation methods to accelerate the curation process while preserving the quality of the database annotation. They apply automatic annotation only to entries belonging to manually defined orthologous[4] families and to ones with no identifiable similarities. They have built in a system that enforces errors checking and can even spot problematic cases. They have integrated there work in Swiss-prot and can be accessed at: http://www.expasy.org/sprot/hamap/

[3] It is a collection of proteins found in particular cell type.

[4] That have evolved from a common ancestor.

In [10] the authors have developed Ambiante para anotac ao Automatica e Comparac ao de Genomas (AC3 System). It is an environment for agent-based annotation and comparison of genomes. The authors have presented a case study where they show how they have used this system to annotate proteins related to the organism Mycoplasma pneumoniae. In this paper the authors have annotated proteins with some attributes and they relied on the keywords field in Swiss-Prot, since it gives several hints to experts about proteins functions and structure. The obtained results were satisfactory.

In [11] the authors present an approach in order to automatically annotate keywords using WEKA (a machine learning decision tree builder tool) for proteins related to mycoplasmataceae. This paper is based on article [1]. Since the rules generated in [1] and the obtained results were highly successful the authors have decided to conduct a similar approach limiting it to a specific organism instead of the whole database. The approach was applied to the mycoplasmataceae organism. They have achieved more accurate rules and results due to a decrease in number of contradictory and erroneous rules.
They have based their work on the work previously done by Kretschmann et Al [1]. They generate rules to automate proteins annotation related to Mycoplasmataceae in SwissProt using C4.5 algorithm using the keywords field. At first they have gathered information for proteins related to M. pneumoniae with 1539 instances, 714 Interpro entry numbers and 130 keywords. They have then decided to increase the size of their training set to include the whole Mycoplasmataceae family. This way, the number of proteins related to

the Mycoplasmataceae family decreased to 786 instances and the Interpro entry number increased to 807 due to the increase in the quantity of organisms, with the number of keywords remaining almost the same They collected for each protein all relevant keywords and InterPro entry numbers. They have created as many datasets as there are keywords with.

In some of the recent work listed above, authors have used InterPro database which was recently created (more details later) in order to automatically annotate proteins related to different organism regarding their keywords fields in the SwissProt database. Interpro is formed of the following databases: PRINTS, PROSITE, Pfam, ProDom SMART and TIGRFAMs [12]. Related sequences from each of these databases are unified into single InterPro entries. Each InterPro entry has a unique accession or entry number, some functional descriptions, literature references, and links to refer back to relevant database(s). For each InterPro entry is a list corresponding to all the matches from Swiss-Prot and TrEMBL databases. This database can be accessed easily from the Swiss-prot website.

In this work we have decided to use Interpro because it unites all other databases. We have also decided to apply Kretschmann et Al's and Bazzan et Al approach on a different organism.

Chapter 3

Problem statement

3.1 Problem statement

The number of newly discovered proteins has been increasing since the creation of the Uniprot database two decades ago. When a new protein is discovered, curators try to classify it by inscribing it in its correct place in the database. In the early years and until the creation of Trembl (the computer annotated supplement to Swissprot (the Swiss protein database)), specific information related to proteins were stored in fields such as entry id (primary key number of the protein), name and origin of the protein, keywords (words describing biological properties of the protein), relation to other database, and sequence (the chain of amino acids forming the protein). When a new protein is discovered curators try to annotate it by assigning correct information to its fields and fit the protein in the best place (in the SwissProt database) and relate it to other existing proteins in all other databases based on biological properties and sequence similarities as described in [10].

Curators perform annotation manually with some automated help by using TrEMBL. But since protein number is huge nowadays, and is constantly increasing, there is a great need for automation of protein annotation and classification .

The size of the database is the major point that makes this an interesting problem. As a matter of fact, the number of proteins already discovered and stored in the several databases mainly in Swiss-prot and PROSITE is 3,346,675 entries as of 4 May 1, 2007. So as new proteins are discovered curators will have to rely on the information already stored in the database in order to correctly identify this new protein and store it in its most appropriate place. This is needed for large projects such as the human genome project (HGP), which was completed by the U.S. department of energy and the national institute of health from 1990 until 2003.

In this thesis, we propose to apply the approach used by Bazzan et Al [12] on the Newcastle virus disease. This virus is poorly annotated in the Swiss-prot database. As a matter of fact, 30 % of 2,631 sequences of this virus have their keywords field annotated.

3.2 Data preparation

In our work, we collect all information related to the 2,631 proteins from the SwissProt data base. We then transform the data by creating a boolean attribute for each Interpro entry field. For each protein, we set the attribute[5] to True if the protein has the respective Interpro entry, False otherwise.

We also create a boolean class label that is set to True whenever the protein is annotated with a specific keyword, and to False otherwise. Table 3 shows an example of a data set with 2 sequences. The attributes are Intepro entries, the classification label indicates whether the protein should be annotated with keyword A.

[5] An attribute is a property of an instance that may be used to determine its classification.

	IPR000776	IPR005454	IPR010292	IPR000776	Annotate with keyword
P012765	True	False	True	True	Yes
Q023944	False	True	False	False	No

Table 3: data set where each row describes a protein, attributes are interpro entry numbers and the last column shows the classification label

We use the program WEKA (a comprehensive machine learning and data mining tool kit, that includes classification, regression, clustering and decision tree generation) [19] with the data described above and run C4.5 in order to build decision trees which build a causal relationship between the fact that an InterPro number exists in the respective protein field and whether the protein is annotated with a certain keyword. Next, we describe in detail the decision tree building mechanism to solve the classification problem.

3.3 The classification problem and decision tree building mechanism

A classification problem is "one of separating a large class of objects into smaller classes, and giving a criterion for determining whether a particular object is or is not in a particular class" [18]. Decision trees have been extensively used for such type of problems. Figure 4 shows an example of a decision tree. A node in a tree encodes an attribute (an Interpro entry number in our case) and arcs leaving a node are labeled with possible values of the attribute (true or false in our case). Leaf nodes are labeled with classifications (true or false in our case).

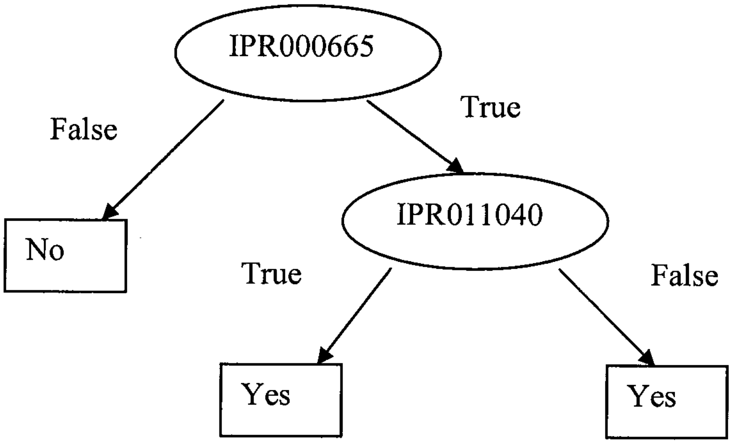

Figure 4: example of a decision tree.

ID3 is one algorithm that builds decision trees. It was developed by Quinlan in 1993[20]. C4.5 is a descendant of ID3 that we use in our work. The algorithm works as follows: Given a set of instances, *S*, and a set of attributes, *A*, ID3 calculates the information gain values for all attributes and selects the one with the highest value to be the root of the tree. It then creates a branch from this node for every value of this attribute. The new attributes list of each branch is then adjusted by subtracting from the initial list the parent of the node attribute and the above process is then repeated to child nodes until either all the attributes have been exhausted or all examples have been perfectly classified. For more information on how the algorithm computes the information gain for each attribute, the reader can refer to [20].

As already mentioned, the attributes that we chose to describe our data are: INTERPRO entry numbers or ID's. INTERPRO is a consortium of member databases such as PROSITE , Pfam, Prints, ProDom, SMART and TIGRFAMs. Proteins that are believed to share highly similar sequences are assigned to the same INTERPRO ID.

3.4 Our approach

A protein is identified by its (sequence signature) the sequence of amino acids that forms it. This protein is assigned useful biological key terms according to its amino acids sequence. This identifies the protein's physicochemical properties. This suggests a strong relationship between the protein signature and the corresponding keywords associated with it [24].

Since the INTERPRO database is based on relations among other databases which classify proteins based on their sequence signatures and block signatures (similar sequences of amino acids identified as blocks of amino acids), it is therefore very important and relevant to try to define rules that annotate a certain protein with a specific keyword based on their INTERPRO id.

We have chosen to work on "Newcastle Virus Disease" because our study showed that most of the keywords field of this organism are poorly annotated or not annotated at all. We have decided to collect data from the SwissProt/TrEMBL databases for all the proteins related to this organism.

We have formed 45 datasets one for each keyword (since we need to generate separate rules for each keyword). Attributes are the same for all datasets and for all keywords since we are working with a fixed organism (the Newcastle Virus Disease). Tables 3 and 4 show a part of the data set for keyword Zinc. A row in the tables describe a protein in terms of its Interpro entry numbers and its annotation with the keyword "Zinc".

	IPR007086	IPR004897	IPR000477	IPR034220	Annotate keyword Zinc
Protein1	True	True	True	False	Yes
Protein2	False	False	False	False	No
Protein3	False	False	False	False	No
Protein4	False	False	False	True	No

Table 4: Data where proteins are described in terms of their interpro entry numbers and their annotation with the keyword Zinc.

	IPR007086	IPR004897	IPR000477	IPR034220	Annotate keyword Zinc-finger
Protein1	True	false	True	False	Yes
Protein2	False	False	False	False	No
Protein3	False	False	False	False	Yes
Protein4	False	False	False	True	No

Table 5: Data where proteins are described in terms of their interpro entry numbers and their annotation with the keyword finger-Zinc.

In the next chapter, we present rules that we obtained with C4.5.

Chapter 4

Data, Results and Discussion

4.1 Experiment Setup

Our data set consists of 2,880 entries each of which describes a protein of the Newcastle Virus Disease[3].

We collected data using SRS - a tool used to search for an organism and all its related proteins in a certain database. We created a table where each row describes a protein in terms of its entry number, relevant keywords and Interpro entry (Table 6).

Of the 2880 proteins found in EBI, 2631 were retained and 249 eliminated because they were missing Interpro numbers and keywords.

Protein entry numbers	Relevant keywords	Interpro entry numbers
P12572	Viral Matrix Protein	IPR011234
P34567	Zinc	IPR034220
O45677	Hydrolase	IPR034220
Q32354	None	IPR000678 IPR001145

Table 6: proteins as they appear in the database.

[3] The data was retrieved from Uniprot and EBI.

We have generated three sets of rules. To build the first set, we created 40 files, one for each keyword. In each file the 2631 instances (proteins) are described by a set of 41 attributes. These attributes indicate relevant Interpro entry numbers. We have chosen randomly 50 instances out of the 2631 instances to be our testing data set so they were not used in the generation of the rules. Later on, the generated rules were tested on these 50 instances to see how well they perform. To build the second set of rules, we considered only proteins belonging to Swissprot database. We used them to generate rules that were later on tested on the proteins from TrEMBL. The third set of rules were generated using TrEMBL as training data and the rules were then tested on data collected from Swissprot. Results and accuracies are shown in the next chapter.

Similarly to [19] we ran C4.5 in three different ways: in normal mode, with 10-fold cross validation and with the pruning option.

4.2 Results and Comparison with Bazzan et al.

The rules constructed from the randomly created training data set had accuracies between 94% and 100%. Bazzan et al. [11] achieved an accuracy ranging between 85.37% and 100% when they generated rules from the training data set cross validation. We achieved an accuracy ranging between 88% and 100% on our testing sets.

In their initial study, Bazzan et al. have discovered that the keywords "Complete proteome" and "hypothetical protein", were of no interest in an automatic annotation. Therefore, they repeated the same experiments filtering out these keywords. On the other hand we have, instead, kept all keywords even those for which we got lower accuracies.

Bazzan et al. have predicted the correct annotation for keyword "cell division" with an accuracy of about 75%. We have predicted the correct annotation for keyword "Hemagglutinin" with an accuracy of about 90%.We have achieved in some cases higher accuracies because our list of Interpro entry numbers is much smaller than Bazzan et al's. Table 7 summarizes the results that we obtained and lists them alongside Bazzan et al.'s.

Keywords	**Bazzan et al's**		**Our results**	
	Training	**Testing**	**Training**	**Testing**
Acyltransferase	99.24%	--	--	--
Hemmaglutinin	--	--	98.15%	99.67%
Coiled-coil	99.49%	--	99.7294%	88%
Electron transport	100%	--	99.9612%	100%
Glycosidase	99.75%	--	99.9611%	200%
Hydrolase	94.44%	--	99.9611%	100%
Metal-binding	99.24%	--	98.456%	99.624%
Oxydoreductase	97.85%	--	99.9612%	100%
Signal	97.58%	--	99.7294%	88%
Transferase	91.35%	--	--	--
Envelope protein	--	--	98.141%	95.9184%
Transmembrane	97.20%	--	99.9612%	100%
Zinc	97.58%	--	98.456%	99.624%
Zinc-finger	99.62%	--	99.9611%	100%
Complete proteome	85.37%	--		
Isomerase Plasmid	--	--	99.9612%	100%
Copper			99.9612%	100%
Cell division	99.36%	--	--	--

Table 7 : Bazzan et al.'s versus ours on some keywords

Chapter 5

Decision Trees and Rules obtained with C4.5 on the Newcastle Virus Disease

5.1 Introduction

In this chapter we show some of the rules that we obtained along with their accuracy (Equation 1).

$$\text{Accuracy} = \frac{TP}{TP + FP} \qquad \text{(Equation 1)}$$

Where *TP* indicates the number of true positives (number of instances correctly classified to have a certain keyword) and *FP* indicates the number of false positives (number of instances incorrectly classified to have a certain keyword) [11]. Hence, accuracy is the percentage of correctly classified instances in the whole database.

5.2 Some rules for annotating proteins with Keywords

5.2.1 Rule1: annotating a protein with keyword Hemagglutinin

```
IPR000665 = True
| IPR011040 = True: yes
| IPR011040 = False: yes
IPR000665 = False: no
```

This rule indicates that a protein should be annotated with the keyword Hemagglutinin only if it contains IPR000665.

The same rule can be displayed in the form of a decision tree as follows: (figure 4).

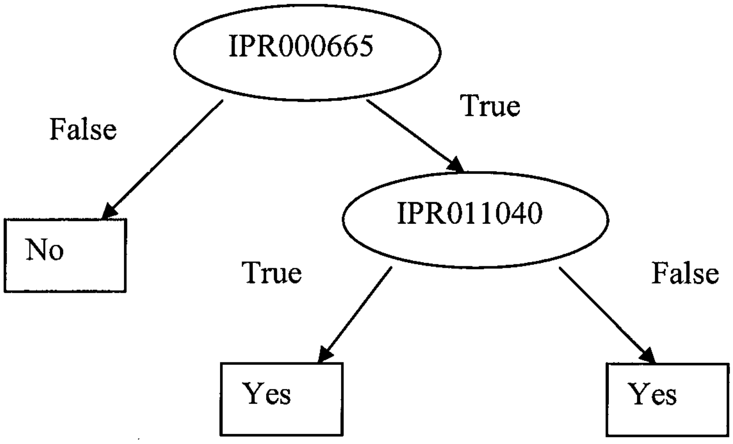

Figure 4: Decision tree for keyword Hemagglutinin

This rule has an accuracy of 98.150 % on the training set and 99.6721% on the testing set.

5.2.2 Rule2: annotating a protein with keyword Isomerase Plasmid

IPR000989 = True: yes
IPR000989 = False: no

The same rule can be displayed in the form of a decision tree as follows: (figure 5).

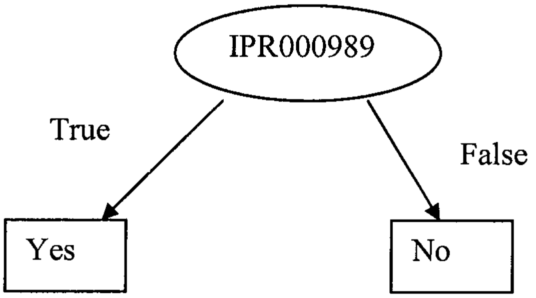

Figure 5: Decision tree for Keyword Isomerase Plasmid

This rule has an accuracy of 99.9612% on the training set and 100% on the testing set.

5.2.3 Rule3: annotating a protein with keywords: Coiled-Coil, Fusion Protein, Lipoprotein, Palmitate and Signal.

IPR000776 = True: yes
IPR000776 = False: no

The same rule can be displayed in the form of a decision tree as follows: (figure 6).

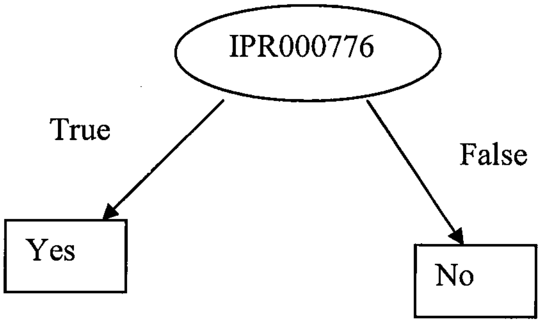

Figure 6: Decision tree for above keywords

This rule has an accuracy of 99.7294% on the training set and 88% on the testing set.

5.2.4 Rule4: annotating a protein with Keyword Glycoprotein and Signal-anchor.

```
IPR000776 = True: yes
IPR000776 = False
| IPR000665 = True: yes
| IPR000665 = False: no
```

The same rule can be displayed in the form of a decision tree as follows: (figure 7).

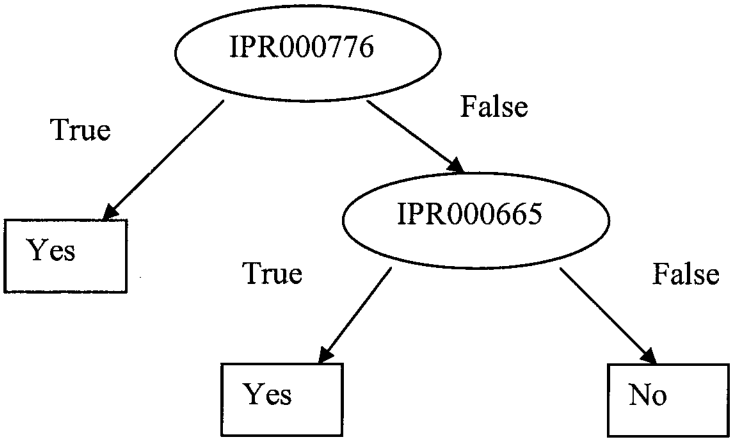

Figure 7: Decision tree for keywords Glycoprotein and Signal-Anchor

This rule has an accuracy of 98.791% on the training set and 96.9697% on the testing set.

5.2.5 Rule5: annotating a protein with keywords:ATP-binding, Methyltransferase, mRNA capping, mRNA processing, Multifunctional enzyme, Nucleotide-binding, Nucleotidyltransferase, S-adenosyl-L-methionineTransferase, RNA replication, RNA directed RNA polymerase.

IPR001016 = True: yes
IPR001016 = False: no

The same rule can be displayed in the form of a decision tree as follows: (figure 8).

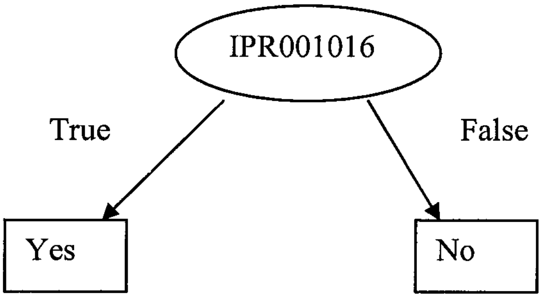

Figure 8: Decision tree for above keywords

This rule has an accuracy of 99.961% on the training set and 100% on the testing set.

5.2.6 Rule6: annotating a protein with Keywords Zinc and Metal-binding.

```
IPR007086 = True: yes
IPR007086 = False
| IPR004897 = True: yes
| IPR004897 = False
| | IPR000477 = True: yes
| | IPR000477 = False: no
```

The same rule can be displayed in the form of a decision tree as follows: (figure 9).

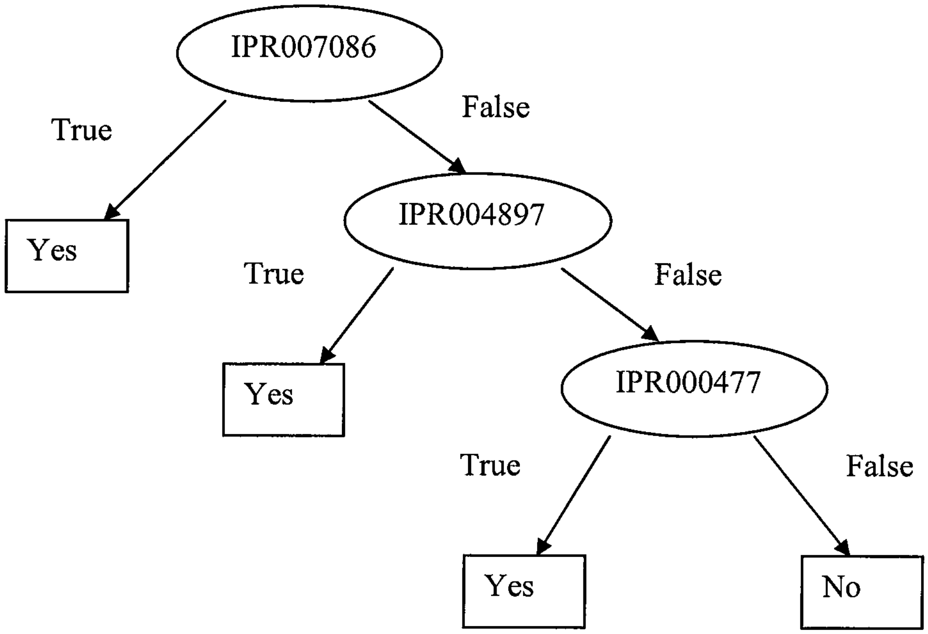

Figure 9: Decision tree for keywords Zinc and Metal-Binding

This rule has an accuracy of 98.456% on the training set and 99.6721% on the testing set.

5.2.7 Rule7: annotating a protein with Keyword Zinc-Finger.

IPR007086 = True: yes
IPR007086 = False
| IPR000477 = True: yes
| IPR000477 = False: no

The same rule can be displayed in the form of a decision tree as follows: (figure 10).

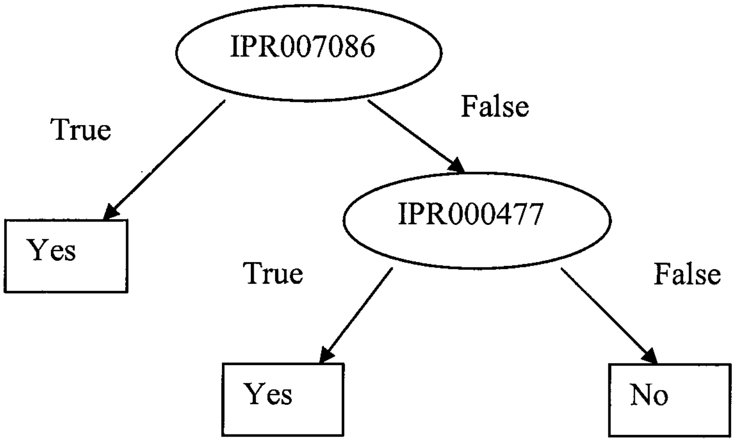

Figure 10:Decision tree for keyword Zinc-Finger

This rule has an accuracy of 99.9611% on the training set and 100% on the testing set.

5.2.8 Rule8: annotating a protein with keyword Interferon antiviral system evasion.

IPR004897 = True: yes
IPR004897 = False: no

The same rule can be displayed in the form of a decision tree as follows: (figure 11).

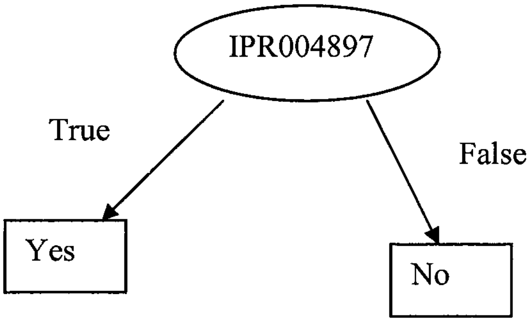

Figure 11: Decision tree for keyword Interferon AntiViral System Evasion

This rule has an accuracy of 99.1692% on the training set and 100% on the testing set.

5.2.9 Rule9: annotating a protein with Keywords: Copper, Electron transport, Heme, Inner membrane, Mitochondrion, Oxidoreductase, Respiratory chain, Transport and Transmembrane.

IPR000883 = True: yes
IPR000883 = False: no

The same rule can be displayed in the form of a decision tree as follows: (figure 12).

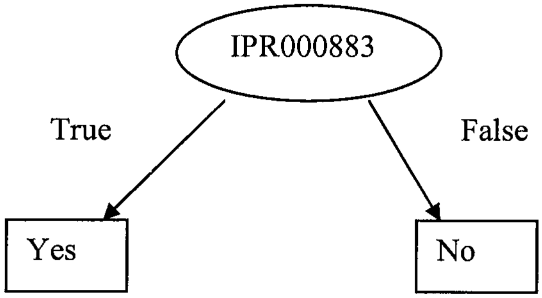

Figure 12: Decision tree for keywords mentioned above

This rule has an accuracy of 99.9612% on the training set and 100% on the testing set.

5.2.10 Rule10: annotating a protein with keywords: GlycosidaseHydrolase and Hydrolase.

IPR000665 = True: yes
IPR000665 = False: no

The same rule can be displayed in the form of a decision tree as follows: (figure 13).

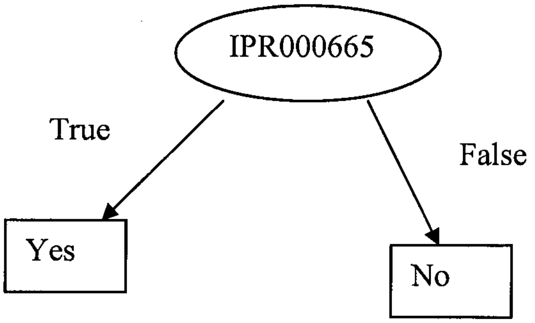

Figure 13:Decision tree for keywords Glycosidase Hydrolase and Hydrolase

This rule has an accuracy of 99.961% on the training set and 100% on the testing set.

5.2.11 Rule11: annotating a protein with keywords Alternative initiation and Phosphorylation.

IPR004897 = True: yes
IPR004897 = False: no

The same rule can be displayed in the form of a decision tree as follows: (figure 14).

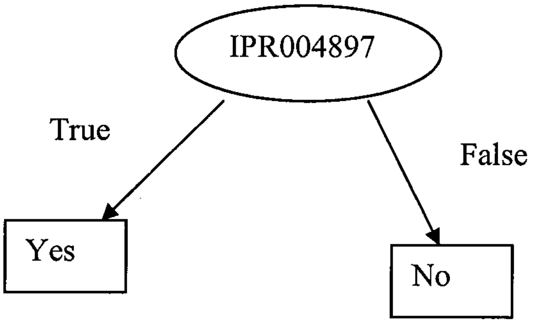

Figure 14:Decision tree for keywords Alternative Initiation and Phosphorylation

This rule has an accuracy of 99.9606% on the training set and 96.4192% on the testing set.

5.2.12 Rule12: annotating a protein with keyword Viral Matrix Protein.

IPR000982 = True: yes
IPR000982 = False: no

The same rule can be displayed in the form of a decision tree as follows: (figure 15).

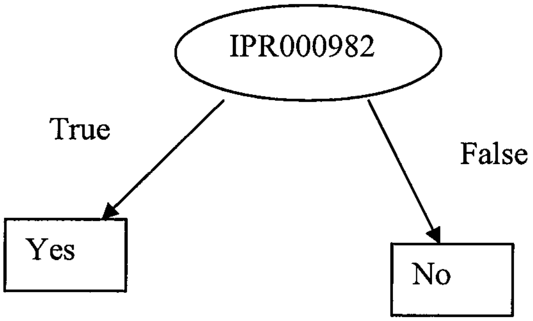

Figure 15:Decision tree for keyword Viral Matrix Protein

This rule has an accuracy of 98.4884% on the training set and 98% on the testing set.

5.2.13 Rule13: annotating a protein with keyword Envelope Protein

```
IPR000665 = True: yes
IPR000665 = False
| IPR000776 = True: yes
| IPR000776 = False
| | IPR000982 = True: yes
| | IPR000982 = False
| | | IPR005166 = True: yes
| | | IPR005166 = False: no
```

The same rule can be displayed in the form of a decision tree as follows: (Figure 16).

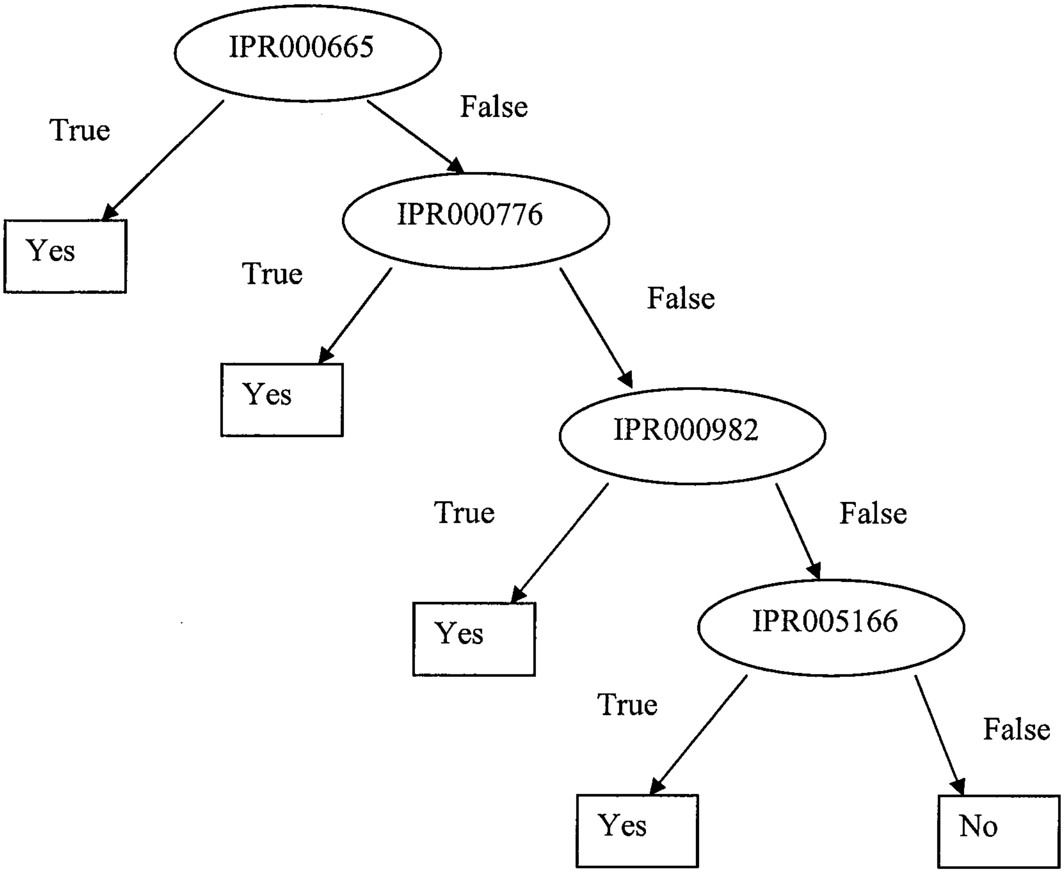

Figure 16: Decision tree for keyword Envelope Protein

This rule has an accuracy of 98.141% on the training set and 95.9184% on the testing set.

5.2.14 Rule14: annotating a protein with keyword Virion Protein

```
IPR000982 = True: yes
IPR000982 = False
| IPR002021 = True: yes
| IPR002021 = False
| | IPR000776 = True: yes
| | IPR000776 = False
| | | IPR001016 = True: yes
| | | IPR001016 = False
| | | | IPR000665 = True: yes
| | | | IPR000665 = False: no
```

The same rule can be displayed in the form of a decision tree as follows: (figure 17).

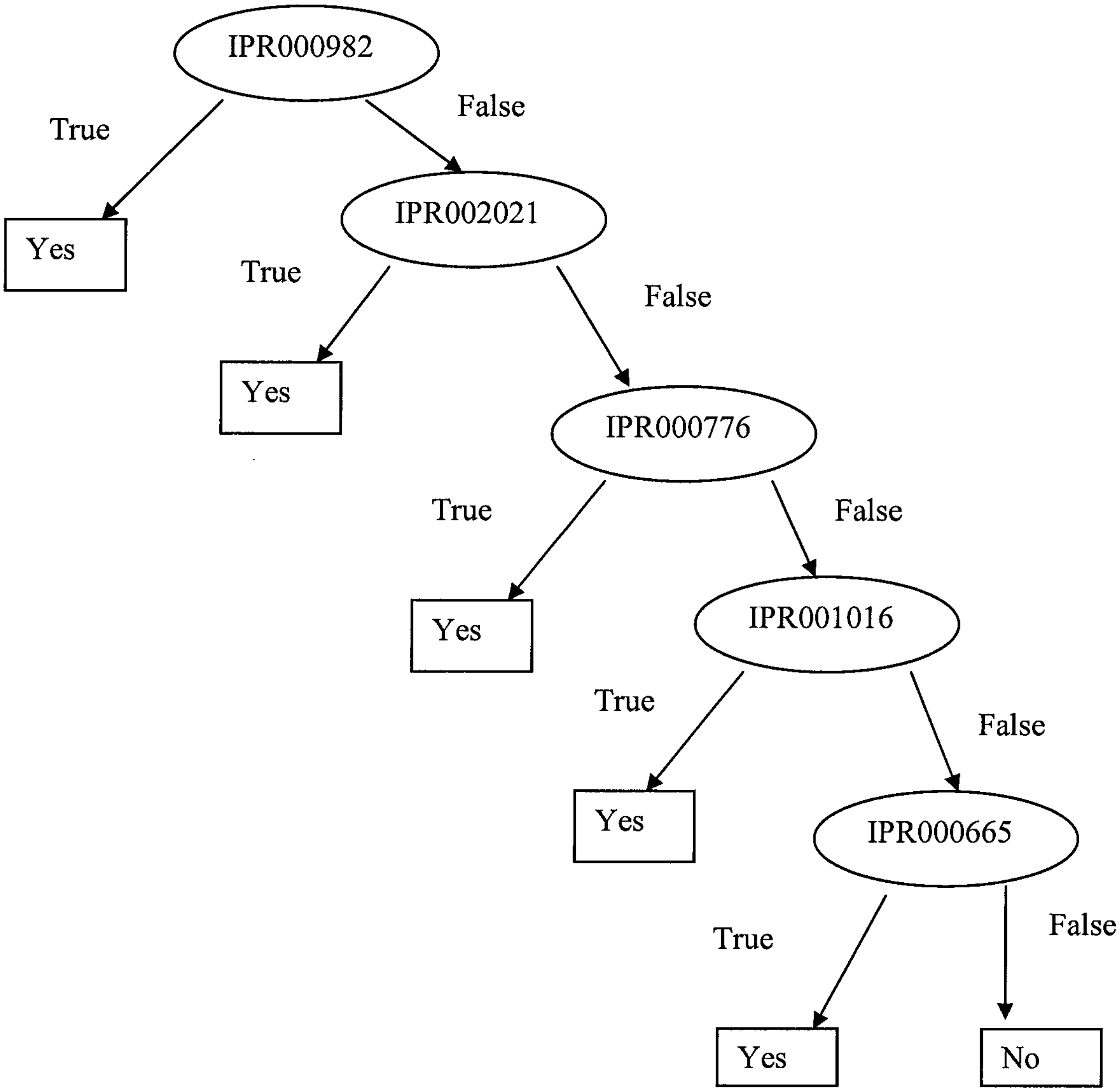

Figure 17:Decision tree for keyword Virion Protein

This rule has an accuracy of 98.8105% on the training set and 96.296% on the testing set.

5.2.15 Rule15: annotating a protein with keyword Viral Nucleo-Protein.

IPR002021 = True: yes
IPR002021 = False: no

The same rule can be displayed in the form of a decision tree as follows: (figure 18).

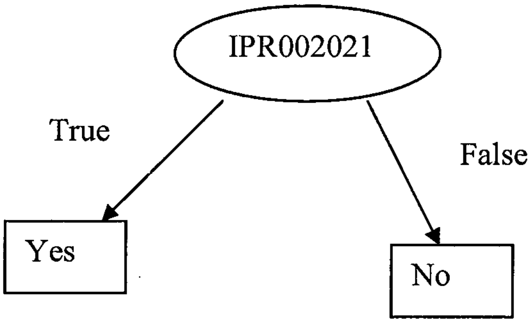

Figure 18:Decision tree for keyword Viral Nucleo Protein

This rule has an accuracy of 99.3769% on the training set and 100% on the testing set.

5.2.16 Rule16: annotating a protein with keyword RNA Directed DNA polymerase and Endonuclease.

IPR000477 = True: yes
IPR000477 = False: no

The same rule can be displayed in the form of a decision tree as follows: (figure 19).

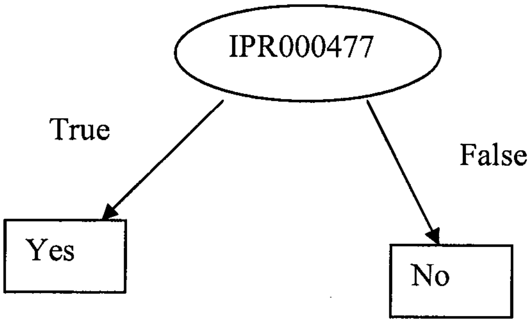

Figure 19: Decision tree for keyword RNA Directed DNA Polymerase and Endonuclease

This rule has an accuracy of 99.9612% on the training set and 100% on the testing set.

Chapter 6

Results validation and testing

In this chapter, we describe the annotation predicted by the rules for each protein and we compare it to the actual annotation found in Swissprot and/or TrEMBL. Table 8 summarizes the results.

	Proteins primary number acordin to Swissprot	Intrepro entry numbers	keywords annotations predicted by our program	keywords annotations as listed in SwissProt/TrEMBL
P1	P35740	IPR00665 and IPR011040	Glycosydase-Hydrolase, Hydrolase, Envelope protein, Hemagglutinin, Signal-anchor, virion protein, Transmembrane, membrane and Glycoprotein.	Hydrolase, Envelope protein, Hemagglutinin, Signal-anchor, virion protein, Transmembrane, membrane and Glycoprotein
P2	P12572	IPR000776	Cleavage on pair of basic residues, coiled coil, envelope protein, fusion protein, glycoprotein, lipoprotein, Membrane, Palmitate, Signal, Transmembrane, Virion protein	Cleavage on pair of basic residues, coiled coil, envelope protein, fusion protein, glycoprotein, lipoprotein, Membrane, Palmitate, Signal, Transmembrane, Virion protein

P3	P16073	IPR004897	Alternative Initiation, Phosphorylation, RNA Replication, Zinc, Metal-binding, Interferon anti viral system evasion and RNA editing.	Alternative Initiation, Phosphorylation, RNA Replication
P4	Q06428	IPR004897	Alternative Initiation, Phosphorylation, RNA Replication, Zinc, Metal-binding, Interferon anti viral system evasion and RNA editing.	Interferon anti viral system evasion, metal-binding, zinc and RNA editing
P5	O90339	IPR000776	Cleavage on pair of basic residues, coiled coil, envelope protein, fusion protein, glycoprotein, lipoprotein, Membrane, Palmitate, Signal, Transmembrane, Virion protein	None (protein belongs to TrEMBL).
P6	Q9YN79	IPR000665 and IPR01140	Glycosydase-Hydrolase, Hydrolase, Envelope protein, Hemagglutinin, Signal-	Envelope protein, Hemagglutinin,Transmembrane and hydrolase

			anchor, virion protein, Transmembrane, membrane and Glycoprotein.	
P7	P11206	IPR000982	Virion protein, envelope protein and viral matrix protein	Virion protein, envelope protein and viral matrix protein.
P8	P11205	IPR014023 and IPR001016	ATP-binding, Methyltransferase,mRNA capping,mRNA processing, Multifunctional enzyme,Nucleotide-binding,Nucleotidyltransferase,S- adenosyl-L-methionine,Transferase, virion protein, RNA directed RNA polymerase and RA replication.	ATP-binding, Methyltransferase,mRNA capping, mRNA processing, Multifunctional enzyme,Nucleotide-binding,Nucleotidyltransferase,S- adenosyl-L-methionine,Transferase, RNA replication, RNA directed RNA polymerase, transferase and virion protein

Table 8: Protein (P1-P8) description, predicted and actual annotations

We can conclude the following about the proteins P1-P8 listed in Table 9:

P1: Our program has annotated correctly all the 8 keywords already listed for this protein (P35740) and in addition it has annotated this protein with an additional keyword: Glycosydase-Hydrolase.

P2. Our program has correctly annotated all the keywords for this protein.

P3. Our program has correctly annotated all three existing keywords for this protein and in addition to 4 other new keywords.

P4. Our program has correctly annotated all four existing keywords for this protein and in addition to 3 other new keywords.

P 5. Our program has predicted the annotation of this protein with keywords listed above. Instead there are no listed keywords for this protein in SwissProt database.

P 6. Our program has correctly predicted the annotation of the 4 keywords that are already listed for this protein in SwissProt and in addition it has annotated this protein with four new keywords.

P 7. Our program has correctly annotated this protein with all three keywords listed in SwissProt database.

P 8. Our program has correctly annotated this protein with all the keywords that are already in the database.

The results show that for all the cases shown above, for proteins that belong to SwissProt database we have achieved for most keyword 95% correct annotation. For proteins belonging to TrEMBL we have achieved for most keywords 90% correct annotation, while the actual percentage of annotated proteins in TrEMBL is 20%. Our program completes the incomplete annotation and annotates all unannotated proteins in TrEMBL. For example, consider the case of protein P5 where the protein is a fusion protein (possibly a glycoprotein) with primary entry number O90339, data coming from TrEMBL states that this protein contains IPR000776 but the keywords field is empty. Our program has predicted the annotation of this protein with these keywords: Cleavage on pair of basic residues, coiled coil, envelope protein, fusion protein, glycoprotein, lipoprotein, Membrane, Palmitate, Signal, Transmembrane, Virion protein.

Below, we show the accuracy of the rules for some keywords.

Hemagglutinin:

1. When running C4.5 on Swissprot data set the Accuracy found is: 100%
2. When running C4.5 on the training set (Swissprot and TrEMBL) the Accuracy found is: 98.456%
3. When running C4.5 on the testing set (Swissprot and TrEMBL) the Accuracy found is: 99.6721%
4. When running C4.5 on TrEMBL data set the Accuracy found is: 97.286%

Rules were extracted from the Swissprot data set and tested on TrEMBL data set.

The accuracy was found to be 86.65%.

Rules were extracted from TrEMBL data set and tested on Swissprot data set.

The accuracy was found to be 100%.

Zinc:

1. When running C4.5 on Swissprot data set the Accuracy found is: 100%
2. When running C4.5 on the training set (Swissprot and TrEMBL) the Accuracy found is: 98.456%
3. When running C4.5 on the testing set (Swissprot and TrEMBL) the Accuracy found is: 99.6721%
4. When running C4.5 on TrEMBL data set the Accuracy found is: 100%

Rules were extracted from Swissprot data set and tested on TrEMBL data set.

We were unable to calculate the accuracy because all instances that contain IPR004897 do not contain the keyword Zinc.

Rules were extracted from TrEMBL data set and tested on Swissprot data set.

We were also unable to calculate the accuracy because all instances that contain IPR007086 and IPR000477 do not contain the keyword Zinc.

Copper, Electron transport, Heme, Inner membrane, Mitochondrion, Oxidoreductase, Respiratory chain, Transport and Transmembrane.

1. When running C4.5 on Swissprot data set no rules were generated no Accuracy was found.
2. When running C4.5 on the training set (Swissprot and TrEMBL) the Accuracy found is: 100%
3. When running C4.5 on the testing set (Swissprot and TrEMBL) the Accuracy found is: 100%
4. When running C4.5 on TrEMBL data set the Accuracy found is: 100%

No rules were extracted from Swissprot data set, therefore we were unable to calculate the accuracy.

Rules were extracted from TrEMBL data set and tested on Swissprot data set.

We were unable to calculate the accuracy because all the instances that contain IPR000883 do not contain the above keywords.

Lipoprotein, Coiled-coil, Fusion, Palmitate and Signal.

1. When running C4.5 on Swissprot data set the Accuracy found is: 75%
2. When running C4.5 on the training set (Swissprot and TrEMBL) the Accuracy found is: 99.713%
3. When running C4.5 on the testing set (Swissprot and TrEMBL) the Accuracy found is: 88%
4. When running C4.5 on TrEMBL data set the Accuracy found is: 100%

Rules were extracted from Swissprot data set and tested on TrEMBL data set.

The accuracy was found to be almost 0%. All proteins containing IPR000776 in TrEMBL were not annotated with above keywords.

No rules were extracted from TrEMBL data set, therefore we were unable to calculate the accuracy.

So to sum up,

- Our generated rules have accurately predicted the annotation of 95% of all the proteins that are listed in SwissProt.
- Our generated rules have accurately and correctly predicted the annotation of 85% of all the proteins that are listed in TrEMBL.
- Our generated rules have enhanced the data in TrEMBL by annotating most of the empty keywords fields with accurate and reliable keywords.

Chapter 7

Conclusion and Future Work

The goal of this thesis is to construct rules that allow automatic annotation of proteins with specific keywords. For this, we have retrieved Interpro entry numbers and keywords related to all the proteins from the SwissProt database for the Newcastle Virus Disease organism. We have used this data with C4.5 to generate rules for each keyword. We have tested these rules on unseen proteins. The rules that we generated have a training accuracy of at least 90% and a testing accuracy of 95%. In addition our work can be used to enhance the data in TrEMBL database by annotating unannotated keywords fields. Our technique was implemented only for proteins belonging to Newcastle Virus Disease. Therefore its accuracy was only tested on these specific proteins. It will be interesting to test it on proteins belonging to other organisms. We believe that it will perform well mainly due to the fact that the Newcastle Virus Disease was relatively poorly annotated.

REFERENCES

[1]. Kretschmann, E., Fleischmann, W., & Apweiler, R. (2001). Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT. *Bioinformatics, 17* (10), 920-926.

[2]. Dieterich, G., Kärst, U., Wehland, J., & Jänsch, L. (2005). MineBlast: A literature presentation service supporting protein annotation by data mining of BLAST results. *Bioinformatics, 21*(16), 3450-3451.

[3]. Frank, E., Hall, M., Trigg, L., Holmes, G., & Witten, Ian H. (2004). Data mining in bioinformatics using Weka. *Bioinformatics, 20* (15), 2479-2481.

[4]. Wieser, D., Kretschmann, E., & Apweiler, R. (2004). Filtering erroneous protein annotation. *Bioinformatics, 20* (Suppl. 1).

[5]. Yu, Gong-Xin. (2004). RuleMiner: A knowledge system for protein function annotations. *Bioinformatics Journal and Computational Biology*, *2*, 615-37.

[6] Bairoch, A., & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acid Research, 28*, (1).

[7]. Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., et al. (2005). The universal protein resource (UniProt). *Nucleic Acids Research, 33.*

[8]. Biswas, M., O'Rourke, J. F., Camon, E., Fraser, G., Kanapin, A., Karavidopoulou, Y., et al. (2002). Applications of InterPro in protein annotation and genome analysis. *Briefings in Bioinformatics, 3*(3), 285-295.

[9]. Gattiker, A., Michoud, K., Rivoire, C., Auchincloss, A.H., Coudert, E., Lima, T., et al. (2003). Automated annotation of microbial proteomes in SWISS-PROT. *Computational Biology and Chemistry,* 27 (1), 49-58.

[10]. Bazzan, A., Dos Santos, C. (2002). *Using the AC3 system for annotation of keyword: A case study.* Instituto de Informatica Universidade Federal de Rio Grande do Sul.

[11]. Bazzan, A., Engel, P.M., Schroeder, L.F., & Da Silva, S. C. (2002). Automated annotation of keywords for proteins related to mycoplasmataceae using machine learning techniques. *Bioinformatics, 18 (Suppl.* 2), S35-S43.

[12]. Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, et al. (2002). InterPro: An integrated documentation resource for protein families, domains and functional sites. *Briefings in Bioinformatics 3*(3), 225-235.

[13]. *The basics and beyond* (2007). Retrieved February 23, 2007, from University of UTAH, Genetic science Learning website: http://learn.genetics.utah.edu.

[14]. *Nobelprize.org*. Retrieved February 10, 2007, from: http://nobelprize.org.

[15]. Carpi, A (1999). *The Cell.* Retrieved February 10, 2007, from John Jay college of Criminal Justice website: http://web.jjay.cuny.edu/~acarpi/NSC/13-cells.htm.

[16]. Wyskiel, N. *Cell structure and DNA.* Retrieved February 23, 2007, from Yale-New Haven Teachers Institute website: http://www.yale.edu/ynhti/curriculum/units/1982/7/82.07.02.x.html

[17]. Freudenrich, C. *How DNA works*. Retrieved February 15, 2007 from: http://science.howstuffworks.com/dna5.htm.

[18]. *Wikipedia: The free encyclopedia.* Retrieved February 12, 2007 from: http://en.wikipedia.org/wiki/mainpage

[19]. *Data mining software in Java.* Retrieved February 24, 2007, from University of Waikato website: http://www.cs.waikato.ac.nz/ml/weka

[20]. *Tutorial (4): Id3, decision trees tutorial.* Retrieved February 15, 2007, from: http://decisiontrees.net/?q=node/27

[21]. Jambeck, P., & Gibas, C. (2001). Predicting protein structure and function from sequence. In *Developing bioinformatics computer skills* (Chap. 10). [n.p.]: [n.p.]. Retrieved February 24, 2007, from O'Reilly Safari Books Online: http://safari.oreilly.com/1565926641/bioskills-CHP-10

[22]. *PIR protein information resource.* Retrieved February 25, 2007, from Georgetown University website: http://pir.georgetown.edu/

[23]. Kretschmann, E. & Apweiler, R. *Clustr Documentation.* Retrieved February 26, 2007, from European Bioinformatics Institute website: http://www.ebi.ac.uk/clustr/documentation.html

[24]. *Dissociation constant.* Retrieved February 25, 2007, from: http://www.answers.com/topic/dissociation-constant

[25]. *InterPro tutorial.* Retrieved February 25, 2007, from European Bioinformatics Institute website: http://www.ebi.ac.uk/interpro/tutorial.html

Appendix A

Used Databases

A.1 Uniprot Knowledgebase

It is the main central database of protein sequences. It consists of two major parts: SwissProt and TrEMBL databases.

A.2 SwissProt

It can be considered as a reliable and consistent source of information regarding proteins since it is well annotated and checked for consistency and accuracy by professional curators.
It was created in 1986 and is mainly used by scientists and biologists from all over the world in order to retrieve important data regarding all existing proteins. It serves as well as a main source of information for other databases for searching for sequences similarities, protein families and possible functional groups.

A.3 TrEMBL

This database is the computer annotated part of the Uniprot. When generated rules are created after applying a certain automated annotation process to the SwissProt database, the outcome or results of this process is stored in the TrEMBL database. During the last decade too many studies were conducted on automatic annotation of SwissProt database and this led to heavy data which needed to be stored somewhere before it was checked (which was rarely done). So the TrEMBL database contains miscellaneous data that needs checking and therefore can not be used as a reliable source of information.
As stated on the UniProt official website: "It contains all the translations of EMBL nucleotide sequence entries not yet integrated in Swiss-Prot".

A.4 InterPro

This database is a combination of several other databases which are Prosite, ProDom, Pfam, Prints, SMART and TIGRFAMs and these databases contain proteins signatures information. InterPro focus on the strength of each of these databases in order to come up with a solid combined information. Proteins that are listed in SwissProt and TrEMBL can be assigned to one or more InterPro groups. The number of InterPro groups nowadays is approximately 11000.