ملخص

يلعب الموت المبرمج دوراً أساسياً في التوازن الطبيعي للخلايا. إن عملية الموت المبرمـج ‘‘ Apoptosis’’ هي عملية إنتحار ذاتي للخلية سواءً كانت الخلية مصابة أو معطلة. ويمثل حدوث أي خلل في هذه العملية خطر جسيم على صحة الجسم ككل، فنقصان في هذه العمليـة ينتهـي بالسـرطان، في حين أن زيادة الموت المبرمج قد ينتهي بـأمراض خطيـرة كضـمور الخلايـا العصـبية والزهيمـر. وتنفيذ هذا النوع من القتل الخلوي هو مسؤولية مجموعة من الإنزيمات تدعى كسبيززز.

توجد الكاسبيزات ‘‘Caspases’’ فى كل الكائنات الحيوانية. و قد تم التعـرف علـى 15 نـوع من الكاسبيزات لغاية اليوم، تم ربط اكثر مـن ثلثيهـا بعمليـة مـوت الخليـة المبرمـج. تلحـق بكـل نـوع مختلف من الكاسبيزات رقم للتمييـز (مثـل كاسـبيز-9 و كاسـبيز- 3). تحلـل هـذه الانزيمـات (تقطـع) سلاسل البروتين فى اماكن محددة. فإذا علمنا أن سلاسل البروتين تتكون من تتابعـات مـن الاحمـاض الامينية يبلغ عددها 20 حمض أميني، فإن هذه الانزيمات لا تحلل السلسـلة إلا بعـد الحمـض الامينـي الاسبارتك ‘‘Aspartic acid (D)’’ و هذه صفة مميزة لهذه النوعية من الانزيمات.

إن الزيادة في عدد البروتينات القابلة للقطع والجهد والمـال المبـذول لمعرفـة ذلـك دفـع للقيـام بالبحث عن وسيلة محوسبة ‘‘computerized’’ تقوم بسهولة بالتنبؤ بقابلية حـدوث عمليـة القطـع أو لا. وهذا البحث يقدم واحداً من أقوى الأدوات المتوفرة حالياً لمعرفة البروتينات القابلة للقطع مـن قبـل كاسبيز -3، ومعرفة مكان القطع ودرجته. وتعتبر طريقة عمـل CAT3 نمـوذج يمكـن تطـبيقه علـى باقي الإنزيمات في المستقبل.

Chapter 1: Introduction

1.1 **Apoptosis**

All cells in multicellular organisms such as insects, plants, and human undergo many physiological processes to keep the organism alive. Cell division and cell death are two major processes that should occur in harmony to keep the homeostasis “internal environmental balance” of the organism.

The ability of the cell to self-destruct when it faces harsh conditions or stress is important for the multicellular organism as the ability to produce new cells. Because of the process of cell death, the over all number of cells in any adult healthy organism remains constant. This critical process is called apoptosis or programmed cell death.

Apoptosis or programmed cell death is a genetically highly conserved cellular function. It was firstly identified in 1972 by Andrew Wyllie and his colleagues who published a paper that described the cell death process of apoptosis [Kerr et al. 1972] . The term apoptosis is of Greek origin, and is used to describe the falling of a petal from a flower or the loss of a leaf from a tree [Bleackley and Heibein 2001] .

The general importance of apoptosis is to get rid of unorganized, infected, old, and damaged cells. These types of cells are the typical targets of apoptosis. The death of these cells would prevent the organisms from suffering from many diseases and keep it alive.

There appears to be a variety of situations where apoptosis is fundamentally important for the development and survival of multicellular organisms. In normal physiological conditions, apoptosis plays an important role in the formation of new anatomical structures [Vartanyan et al. 2005]. For example, there is a web of cells between the fingers and toes of the human fetus, giving the hands and feet paddle-like features. As development progresses, the fingers and toes are sculpted from this paddle, with the cells between the new digits dying via apoptosis. Also apoptosis is responsible for the death of tail cells of the tadpole during its transition into a frog [Duke et al. 1996]. In addition, apoptosis is also important in the formation of functionally neural network and the development of the neural system [Mazarakis et al. 1997].

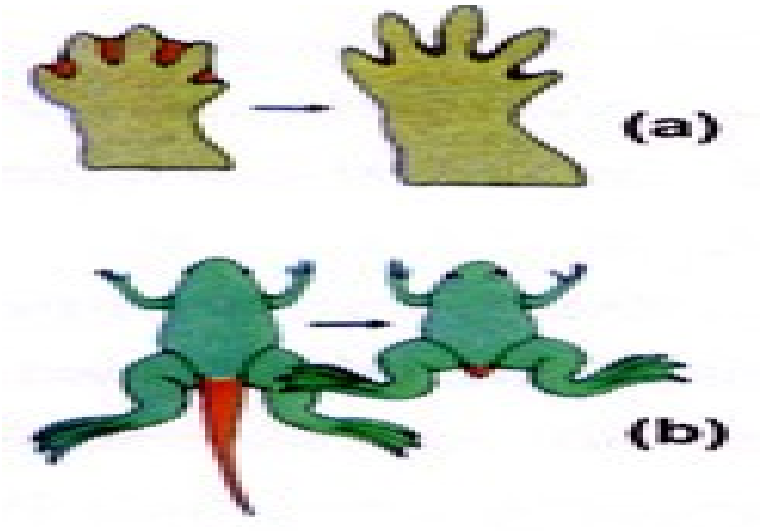

Figure 1.1 (a): Fetus hand development

Figure 1.1 (b): Tadpole tail destruction

On the other hand, uncontrolled and deregulated apoptosis will end with dangerous diseases such as cancer, AIDS, and Alzheimer's. Inhibition of the normal process of apoptosis can lead to the formation of tumors. In case of AIDS, the HIV virus infects a key cell in the immune system and destroys this cell by activating its apoptosis program. This in turn leads to the collapse of the whole immune system with well-known consequences. Apoptosis in the central nervous system is also thought to play a key role in loss of cells, which is a key feature of diseases such as Alzheimer's and other related diseases [Duke et al. 1996].

The main effectors of apoptosis are caspases. Their activation leads to different morphological and biochemical changes such as shrinkage, chromatin condensation, DNA fragmentation and plasma membrane blebbing [Chay et al. 2002]. These changes will make the cell that undergoes apoptosis to be a target for phagocytosis. Phagocytes; a group of white blood cells, recognize the external changes on the cell membrane, such as the externalization of phosphatidylserine and engulf the cell [Earnshaw et al. 1999].

1.2 **Caspases**

Caspases are a group of enzymes that belong to the family of cysteine proteases. The name "caspase" comes from: Cysteine-dependent ASPartyl-specific protease [Alnemri et al. 1996].

Up-to-date, 15 mammalian caspases have been described [Eckhart et al. 2005]. The name of caspases is referred to in the order of their publication. Therefore, the last caspase to be identified is caspase -15.

Caspases are synthesized as zymogens “inactive form”, which are activated by various triggers that cause the active caspase to express itself and carry out its specific function [Paszty et al. 2002]. Caspases activation usually occurs through proteolytic processing of the zymogen at conserved aspartic acid residues “Asp” or “D”. Two cleavages are required to convert the zymogen to an active caspase enzyme. The first cleavage separates the prodomain from the large subunit while the other separates the large and small subunits [Earnshaw et al. 1999] (Figure 1.2). The only mammalian enzymes that can activate caspases are the caspases themselves with the exception of Granzyme B. [Earnshaw et al. 1999]

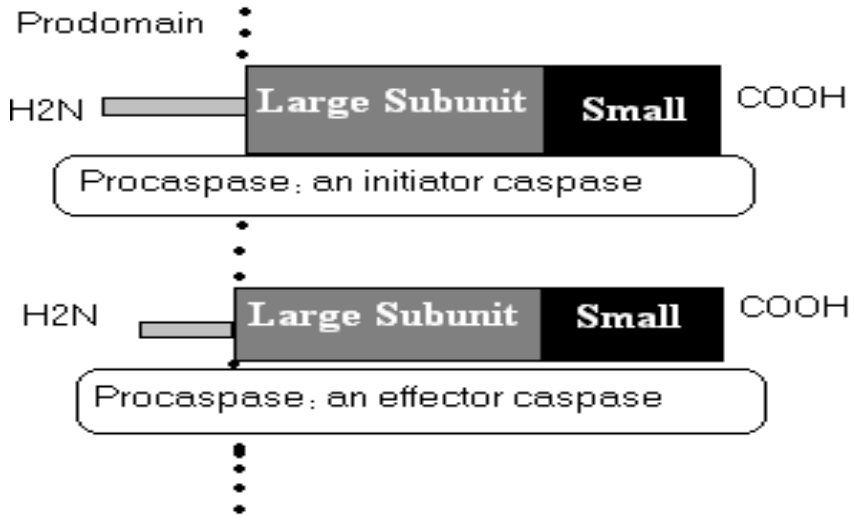

Figure 1.2: Zymogen structure

According to their prodomain, caspases are subdivided into two categories:

a) Initiator caspases: have long prodomain, examples are: caspase -8, -9 and -10.

b) Effector caspases: have short prodomain, examples are: caspase-3, -6 and -7.

While according to their function and structure caspases have two principal subfamilies:

a) Caspase -1 subfamily: includes caspases -1, -4, -5, and -13. Their main function is to control the inflammation.
b) Caspase-3 subfamily: includes caspases -3, -6,-7, -8, -9 and -10 as they are specialized in apoptosis.

Caspase -2 is structurally similar to the caspase -1 subfamily members but functionally it is involved in apoptosis. [Earnshaw et al. 1999]

Caspases play important roles in the initiation and execution of programmed cell death “apoptosis”. Both initiation and execution of apoptosis are carried out through cascade cleavage of substrates.

Although caspases major function is apoptosis, they are also involved in other important cellular process such as in inflammation, proliferation, cell cycle, and spermatogenesis [Los et al. 2001].

1.3 **Mechanism of action**

Caspases are activated from zymogens to active caspases by self cleavages. After activation caspases will cleave their wide range of substrates in the same manner of their self cleavage.

The cleavage of caspases is characterized by the presence of Aspartic acid 'D' residue in the P1 (Figure 1.3) position in the substrate. Although the cleavage of all substrates is executed after the amino acid 'D' at P1, the existence of 'D' alone does not make a peptide susceptible for cleavage by caspases.

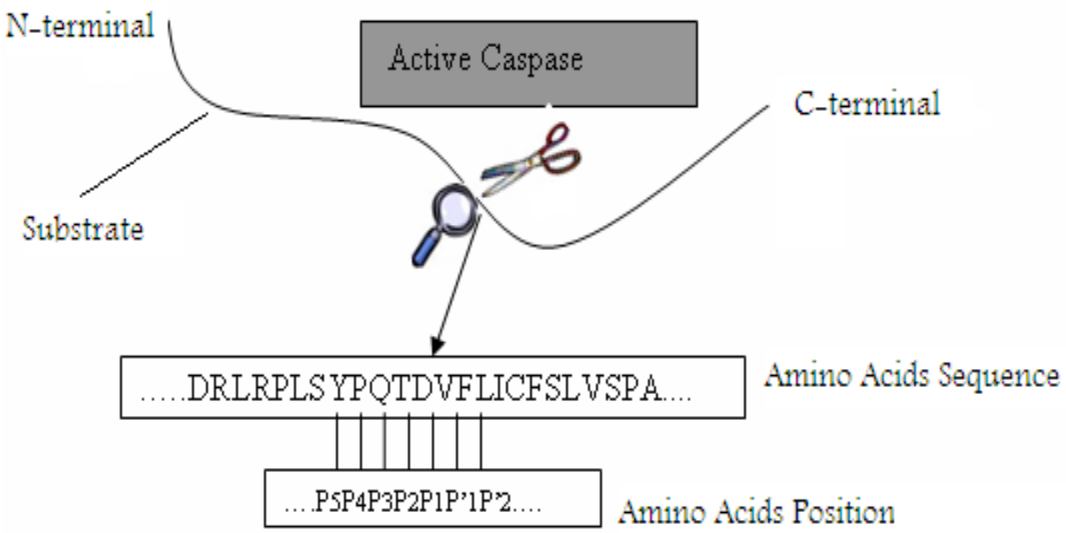

Figure 1.3: Caspase -3 cleavage process

However all caspases cleave their substrates after the same amino acid "D", the preferred three amino acids before the "D" at P1- at least- are varied between caspases [Backes et al. 2005]. Table 1.1 shows the classification of caspases according to their preferable cleavage site sequence "motif" [Garay-Malpartida et al. 2005; Backes et al. 2005]:

Table 1.1: Caspases preferable motif

	Caspases	Preferred motif/s
Group 1	1, 4, 5, 11, 12, 14	[W/L]EHD
Group 2	2, 3, 7	DExD
Group 3	6, 8, 9, 10	[L/V]E[T/H]D

Recently, it has been reported that the P1 position should be followed by a small amino acid such as Alanine 'A' or Glycine 'G' to be a good substrate of caspases [Stennicke et al. 2000]. Because of this, the recognition of at least five amino acids (P4-P3-P2-P1-P'1) (Figure 1.3) is necessary in the cleavage process carried out by all caspases.

The knowledge of the high affinity of caspases toward such motifs gives the researchers the ability to use such motifs as inhibitory peptides. For example, the motif DEVD found within poly (ADP-ribose) polymerase (PARP) is cleaved by caspase-3 [Lazebnik et al. 1994], and it has been used to create the tetrapeptide inhibitor Ac-DEVD-CHO that is a common inhibitor for caspase-3.

1.4 **Caspase -3**

Caspase -3 (interleukin-1beta converting enzyme/CED -3) is the main executer caspase that is responsible for the cleavage of many key proteins. Caspase -3 activation plays a central role in apoptosis. Activated Caspase -3 is responsible for the breakdown of several cellular components to DNA-repair and regulation [Earnshaw et al. 1999; Zhan et al. 2002].

Although the most preferable motif for caspase-3 is “DExD” [Thornberry et al. 1997], many substrates that are specific for caspase-3 found to have the unconventional motif “xxxD”. Yet, there is discussion about the significance of other conserved amino acids located outside the motif. The variability of amino acids in the cleavage site complicates the recognition and the prediction of these motifs.

Up-to-date, caspase -3 has more than 150 experimentally verified substrates. Some of these substrates do not have the conserved motif

"DExD" at their cleavage sites. On the other hand, some proteins that are not considered as caspases substrates may have "DExD" motifs in their primary protein sequence.

1.4.1 **Caspase -3 substrates**

Caspase -3 substrates are proteins with various functions and come from different protein families. Table 1.2 classifies many of caspase -3 substrates according to their function and location in the cell.

Table 1.2 Caspase -3 substrates

PROTEIN GROUPS	REFERENCE
Nuclear proteins	
Lamin A/C	[Rao et al. 1996]
Mdm2	[Pochampally et al. 1998]
U1 snRNP	[Casciola-Rosen et al. 1996]
SAF-A	[Kipp et al. 2000]
Protein kinases	
PKC ζ	[Smith et al. 2000]
PKC-Є	[Basu et al. 2002]
PKC-Θ	[Datta et al. 1997]
PKC-δ I	Persaud et al. 2005]
PKN	[Takahashi et al. 1998]
PKC-μ	[Haussermann et al. 1999]
PAK2	[Walter et al. 1998]
Mst1/2	[Graves et al. 1998]
HPK1	[Chen et al. 1999]
Apoptosis "direct"	
DFF /ICAD	[Inohara et al. 1998]
Bcl-2	[Bellows et al. 2000]
Caspase -9	[Zou et al. 2003]
Cytoplasmic proteins	
Beta-actin	[Song et al. 1997]
Gelsolin	[Kothakota et al. 1997]
Gas2	[Sgorbissa et al. 1999]
Beta-Catenin	[Steinhusen et al. 2000]
Keratin 18	[Caulin et al. 1997]
RAP1	[Cosulich et al. 1997]
Signal transduction pathways	
PP2A	[Santoro et al. 1998]
$cPLA_2$	[Luschen et al. 1998]
SREBP-1/2	[Wang et al. 1996]
RasGAP	[Yang and Widmann 2001]
IL-18	[Akita et al. 1997]
IL-16	[Zhang et al. 1998]
D4-GDI	[Na et al. 1996]
Regulation of cell cycle proliferation	
p21 waf1	[Gervais et al. 1998]
p27Kip1	[Eymin et al. 1999]
Nedd4	[Harvey et al. 1998]
DNA metabolism and repair	
PARP-1	[Lazebnik et al. 1994]
Rad51-A	[Flygare et al. 2000]
Topo- I	[Samejima et al. 1999]

Keeping in mind that some substrates are cleaved at more than one site. Although these substrates are cleaved by caspase -3, some of these

substrates also could be cleaved by other caspases mainly caspase -7 as it belongs to the same functional family “executors”.

The importance of caspase -3 substrates in many medical fields on one hand; and the difficulty to recognize these substrates on the other hand encourage the research of an algorithm that could predict the cleavage site of any given protein.

1.4.2 Available tools for prediction of Caspase -3 substrates cleavage site

The explosion of bioinformatics data and its availability on the web help many researchers develop bioinformatics tools that solve biological problems depending on scientific computing. In some fields a significant progress has been made, while in other fields work is still needed.

In general, tools that predict the cleavage site of many endopeptidase enzymes are few and have wide range of error. There are

only few available tools on the web that deal with the prediction of caspases substrates in general -as there is no specific tool only for caspase -3. Here we show the main two tools that are available:

1.4.2.1 GraBCas

This bioinformatics tool predicts the cleavage site for the caspases "1-9" and granzyme B substrates. GraBCas was developed according to experimentally substrates specificities determined by using positional scanning synthetic combinatorial libraries [Thornberry et al. 1997]. The amino acids in the positions P4, P3 and P2 were analyzed and a position specific scoring matrix (PSSM) was developed for each caspase and for granzyme B. Additional filter was used for caspase -3 and granzyme B. The filter depends on including additional sites (P6-P'2) in computing the score. Medium to large size amino acids (C, Q, I, M and V) at the position P'2 were excluded by a "low stringency" filter. While a "high stringency" filter selects hits with G at the same position [Backes et al. 2005]. GraBCas was written in Java and is available as an applet or as software at the following address:

http://wwwalt.med-rz.uniklinik-saarland.de/med_fak/humangenetik/software/index.html.

1.4.2.2 CaSPredictor

This tool is Visual Basic-programmed software that is not available on the web and can be obtained by direct contact with the author. CasPredictor developments depended on the analysis of natural caspases substrates. The software uses the Caspase Cleavage Site searcher algorithm "CCSearcher". The CCSearcher algorithm was developed based on three parameters. One of the parameters was the PEST index (I_{PEST}) which is computed by giving a value of 1 to the amino acids (S, T, P, E, D, N and Q) and 0 for other amino acids in the region from P19-P'16 [Garay-Malpartida et al. 2005].

$$I_{PEST} = \frac{P19+P18+P17+\ldots+P'16}{N}$$

N in general cases equal to 35.

Both tools: GraBCas and CaSPredictor; are general for all caspases. This generality increases the quantity of substrates to be predicted for all caspases; but decreases the specificity for any caspase

enzyme substrates as general perspectives are considered in calculating the final score. The focus on caspase -3 in one algorithm -as this thesis about – would decrease the error and be more specific in predicting any given protein.

1.5 Thesis review

The problem (Motivation): Caspase -3 is a highly selective enzyme that cleaves its peptide sequences (string) substrates only after aspartic acid "D" residue (character). Yet, this selectivity is rather not straightforward. Aspartic acid residue exists on average 6 times in a protein of 100 amino acids. However, only 2.5% of all aspartic acid residues can act as potential caspase -3 cleavage sites. In general, to know whether a given D residue is cleaved by caspase -3 or not, would costs about $10,000 with 6 months of lab works. Therefore, biologist are in a great need of a strong bioinformatics tool that can help them in predicting the right cleaved D using the protein sequence as input data instead of costly and time consuming lab work. The theoretical

prediction would minimize the number of D residues that should be experimentally verified, thus reducing significantly the time and cost of caspase -3 substrates discovery.

The goal (Objective): to build an algorithm that will predict if any given protein sequence (string) contains a caspase -3 cleavage site. The algorithm is specific for caspase -3. However, its principles could be easily applied to any other caspases substrates.

Thesis chapters:

- **Chapter 2 "Analysis and methods":** in this chapter, we analyzed caspases -3 substrates that are experimentally proven. The aim of analysis was to look for any common features among these substrates especially around their cleavage sites. The common features were used later to establish CAT3 scoring matrices.

- **Chapter 3 "Results":** In this chapter, we introduce the results of our analyses. We describe the important features that normally exist around D residue in order to be recognized as caspase -3 cleavage site. Based on these distinctive features, we constructed our scoring matrices. These matrices are the source from which each tested peptide sequence will retrieve its final score.
- **Chapter 4 "CAT3 algorithm":** this chapter explains the scoring system used in CAT3. An example is shown to explain exactly how CAT3 calculates a score for any given protein string.
- **Chapter 5 "Discussion and conclusion":** this chapter discusses our results. In addition, we discuss the results of comparing CAT3 accuracy with existing similar tools.
- **Chapter 6 "Appendices":** this chapter contains all the tables and codes we use in this thesis.

Chapter 2: Materials and Methods

2.1 Materials and software

The following softwares were used in data collection and analysis:

- Windows XP professional edition.
- Microsoft office 2003 professional edition.
- Internet Explorer.
- Perl: Active-state.
- Perl Builder 2.0.
- Perl2exe.
- EndNote X.

2.2 Data collection

PubMed literature database (www.pubmed.gov) was used to search for papers that describe human proteins, which act as caspase -3 natural substrates. Each paper was critically analyzed to specify the position of the cleavage site. The amino acid sequence of the proteins that were experimentally proven to include a specific caspase -3 cleavage site were obtained from Expasy homepage (www.expasy.org).

We could collect 144 experimentally proven caspase -3 substrates. The 144 obtained sequences were divided into two groups; 119 proteins, containing 136 cleavage sites, designated as matrix establishing group (MEG) and 25 proteins with 27 cleavage sites, which were designated the Test substrates group (TSG).

Data were organized in an excel sheet that contains the following fields:

- Name: the name of the substrate, or its Synonyms name “shortcut”.
- Accession: the reference code of the protein in the SWISS-PROT database
- Motif: the cleavage site {P4-P1} and its position in the protein sequence
- Pub-Med: a reference number of the article in the NCBI/Pub-Med database

The proteins primary sequences were saved for each substrate as a text file with its accession number as the name of the file.

2.2.1 Species database

Another database for some of the caspase -3 substrates were developed. While the first database contains only human substrates for caspase -3, this database contains information about caspase -3 substrates in different species. The species database compares the caspase -3 substrates that found in human with those found in other species. Mouse was the main species in the comparison, other species like rat, rabbit, pig, and bovine were also found in some substrates.

2.3 Analysis

The analysis of the caspase -3 substrates was taken over according to the following points:

1. Determination of the cleavage site(s) for each substrate
2. Determination of the region of interest outside the tetra-peptide cleavage site
3. Study the chemical properties of the selected regions
4. Study the distribution of each amino acid in the selected regions

All the analyses were mostly done using excel. Results of these analyses were demonstrated in charts and figures.

Analyses to revile essential amino acids or significant sections of caspase -3 substrates outside the tetra-peptide cleavage site were made by the following three approaches:

1. Comparing substrates among various species
2. Secondary structure analysis: the cleavage site was determined for each substrate. 50 amino acids to both the N-terminal and the C-terminal of the cleavage site were analyzed for their secondary structure using an online program GOR4
3. Chemical properties and amino acid analysis: the chemical properties were obtained by converting the amino acids to their corresponding chemical group according to (Figure 2.1).

Table 2.1: Amino acids chemical groups

Acidic		Basic			Polar					Nonpolar									
D	E	H	K	R	N	Q	S	T	Y	A	L	P	M	G	V	I	F	W	C

Different lengths have been taken to figure out the critical length of amino acids around the motif for the caspase-3 enzyme cleavage process. Figure 2 shows different lengths that were taken in the present analysis:

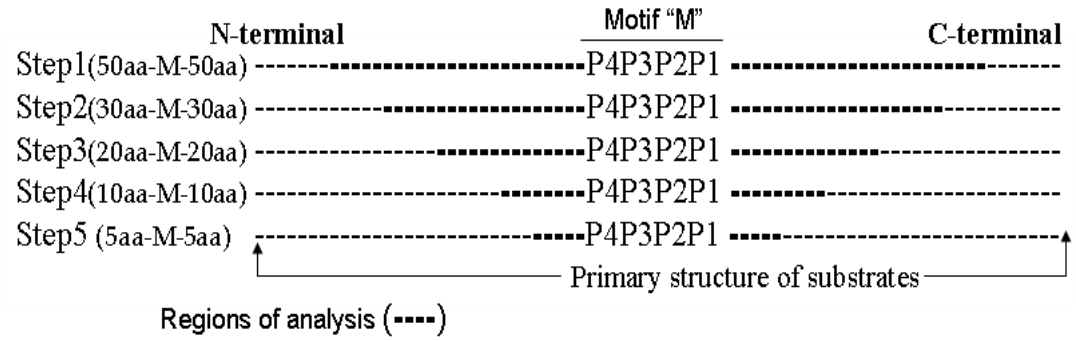

Figure 2.2: Steps in analyzing the regions before and after the motifs.

2.3.1 Selection of controls

Two types of controls were used with which our results can be compared to:

1. **Normal distribution of amino acids**

   There is a natural variation in amino acids frequency based on their codon usage. To compare the frequency of amino acids in the regions around the cleavage sites with normal amino acids frequencies the 144 substrates whole sequences were counted and the percentage for each amino acid was calculated. The final amino acids frequencies in the 144 caspase -3 substrates were very close to normal amino acids frequencies in human proteins.

2. **Uncleaved motifs in the same substrates**

   In order to test and verify the accuracy of our algorithm in predicting caspase -3 cleavage sites, MEG was analyzed for any Aspartic acid 'D' excluding the true cleavage sites from this analysis. These residues are considered to be useful negative controls since they were shown to be experimentally uncleaved.

Chapter 3: Results

The first two steps in analyses: species substrates and secondary structure analysis; of the regions surrounded the cleavage sites in MEG showed similarity between most of the substrates. Therefore, both

analyses steps could not be used to establish any significant marker to be used later in forming the score matrices.

Comparison between the regions surrounded the cleavage sites in human substrates, which we had in some of the substrates in MEG, with those in other species shows a high rate of similarity.

Analyzing some of the substrates in MEG using GOR4 shows a coiled structural feature in most of these substrates. GOR4 is an online[1] tool that predicts protein secondary structure. Figure 3 shows an example of GOR4 output for a random protein string.

[1] http://abs.cit.nih.gov/gor/

```
    150       160       170       180       190       200       210
      |         |         |         |         |         |         |
KPMSNFRFGENHAIMGVAFIWVMALACAAPPLVGWSRYIPQGMQCSCGALYFTLKPEINNESFVIYMFVV
ccccecccccchhhhhhhhhhhhhhhhcccccccccccccccccccccccccccccccccccccccccc
          ....
    220       230       240       250       260       270       280
      |         |         |         |         |         |         |
HFSIPLIVIFFCYGQLVFTVKEAAAQQQESATTQKAEKEVTRMVIIMVIAFLICWLPYAGVAFYIFTHQG
eeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeee

GOR4 :
   Alpha helix      (Hh) :    82 is  23.56%
   3₁₀ helix        (Gg) :     0 is   0.00%
   Pi helix         (Ii) :     0 is   0.00%
   Beta bridge      (Bb) :     0 is   0.00%
   Extended strand  (Ee) :   111 is  31.90%
   Beta turn        (Tt) :     0 is   0.00%
   Bend region      (Ss) :     0 is   0.00%
   Random coil      (Cc) :   155 is  44.54%
```

Figure 3 GOR4 output. GOR4 is a secondary structure prediction tool.

3.1 Chemical groups % of the regions before and after the motif

- Step1: (50-Motif-50) includes all the substrates that have at least 50 amino acids before and after the motif. Strings that have less than 50 amino acids were excluded in each graph.

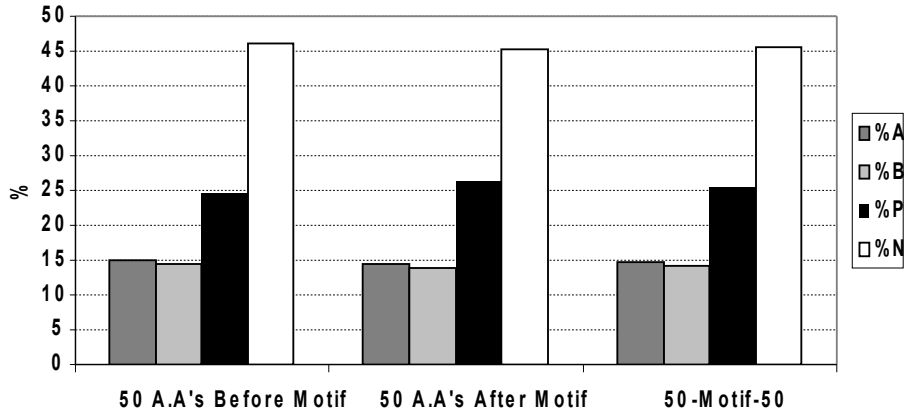

Figure 3.1 (a) Average % of 50 amino acids before and after the motif. A is for acidic, B for basic, P for polar, and N for nonpolar amino acids.

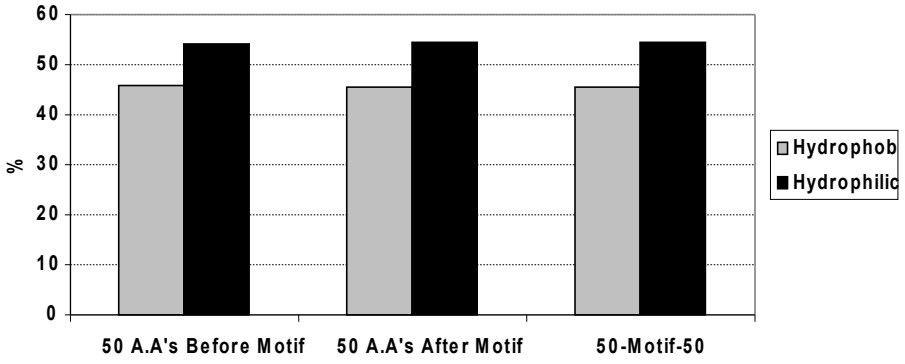

Figure 3.1 (b) Average % of Hydrophibicity of 50 amino acids before and after motif. Hydrophobic are nonpolar amino acids, while Hydrophilic are acidic, basic, and polar amino acids.

- Step 2: (30-Motif-30)

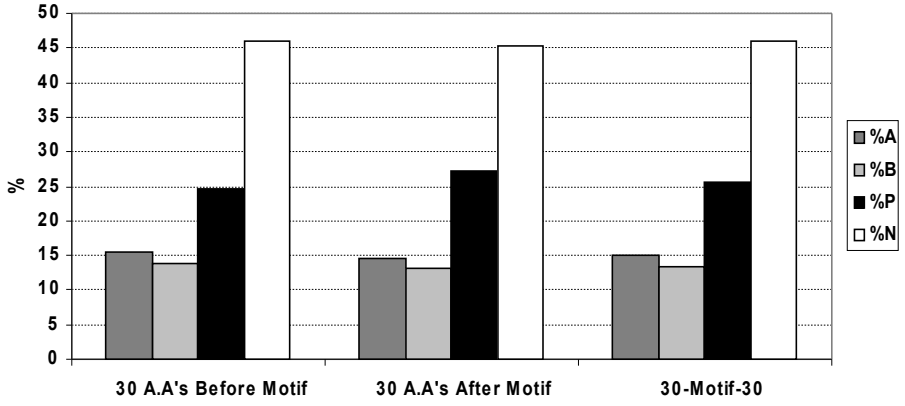

Figure 3.2 (a) Average % of 30 amino acids before and after motif. A is for acidic, B for basic, P for polar, and N for nonpolar amino acids.

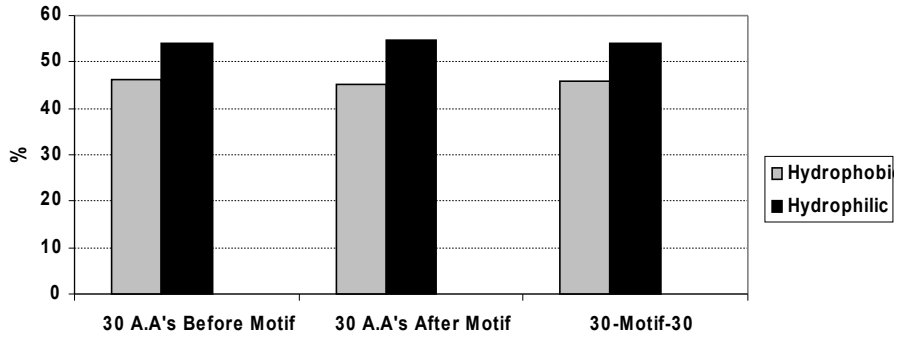

Figure 3.2 (b) Average % of Hydrophibicity of 30 amino acids before and after motif. Hydrophobic are nonpolar amino acids, while Hydrophilic are acidic, basic, and polar amino acids.

- Step 3: (20-Motif-20)

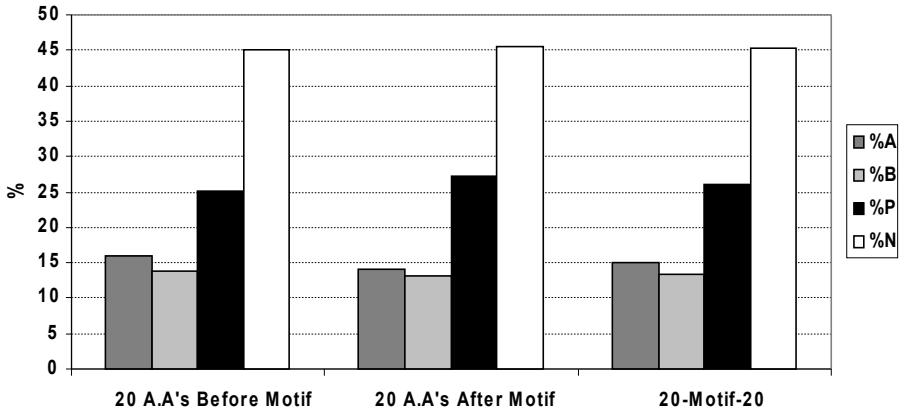

Figure 3.3 (a) Average % of 20 amino acids before and after motif. A is for acidic, B for basic, P for polar, and N for nonpolar amino acids.

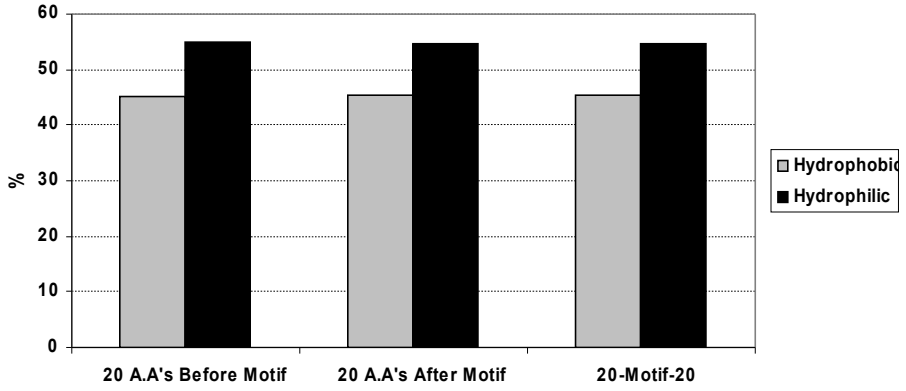

Figure 3.3 (b) Average % of Hydrophibicity of 20 amino acids before and after motif. Hydrophobic are nonpolar amino acids, while Hydrophilic are acidic, basic, and polar amino acids.

- Step 4: (10-Motif-10)

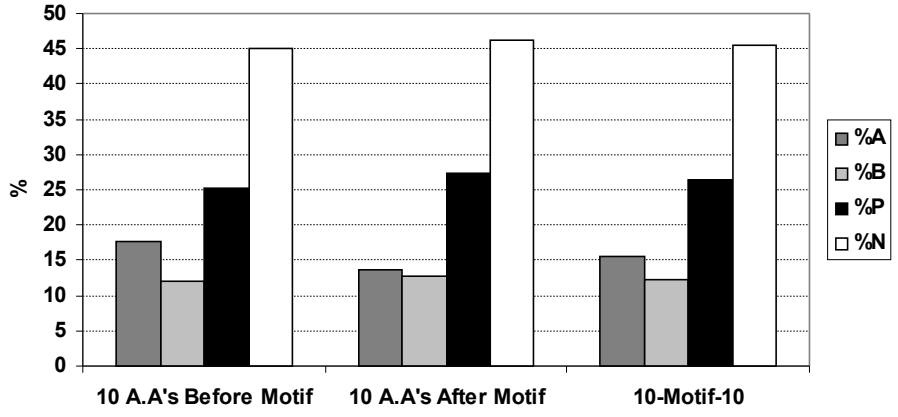

Figure 3.4 (a) Average % of 10 amino acids before and after motif. A is for acidic, B for basic, P for polar, and N for nonpolar amino acids.

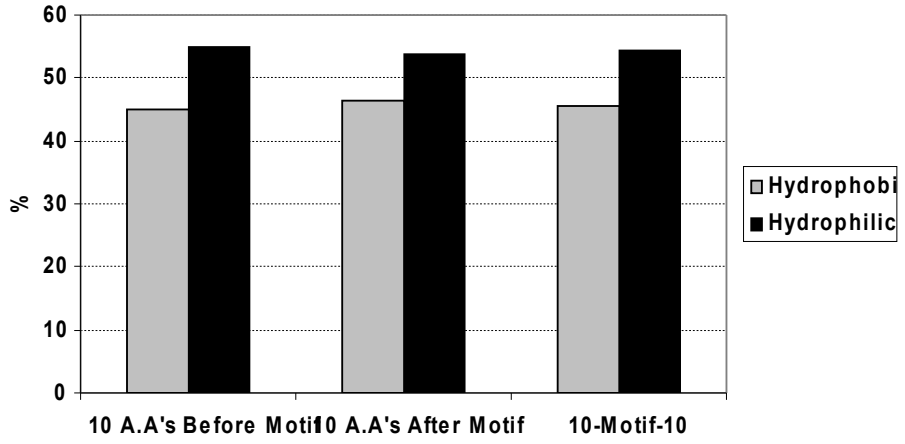

Figure 3.4 (b) Average % of Hydrophibicity of 10 amino acids before and after motif. Hydrophobic are nonpolar amino acids, while Hydrophilic are acidic, basic, and polar amino acids.

- Step 5: (5-Motif-5)

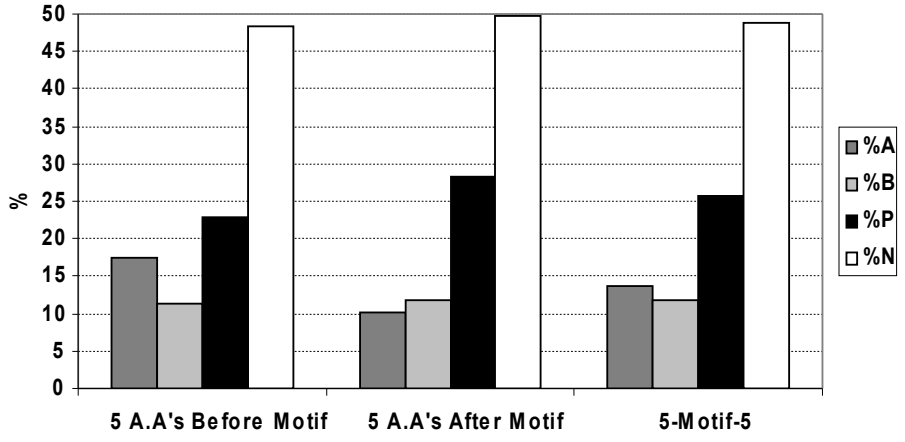

Figure 3.5 (a) Average % of 5 amino acids before and after motif. A is for acidic, B for basic, P for polar, and N for nonpolar amino acids.

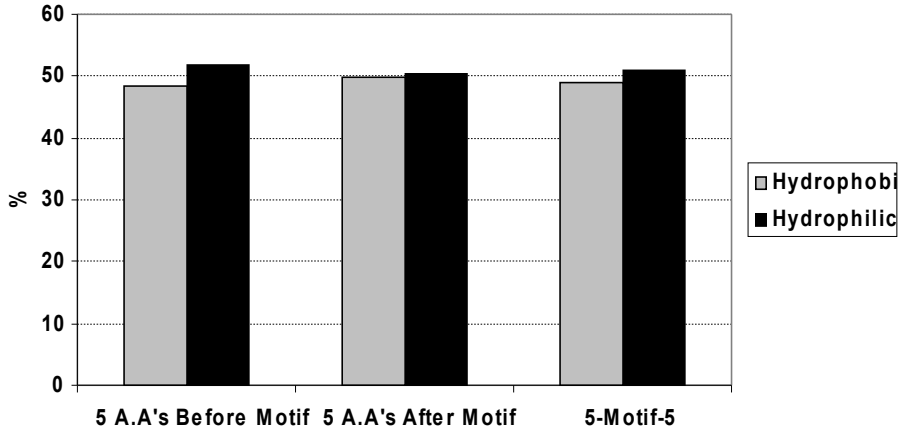

Figure 3.5 (b) Average % of Hydrophibicity of 5 amino acids before and after motif. Hydrophobic are nonpolar amino acids, while Hydrophilic are acidic, basic, and polar amino acids.

3.2 Amino acids content of the regions before and after the motif

This analysis was done only for the 4th and 5th steps (5, and 10 amino acids before and after the motif). The analysis for each amino acid percentage separately in the region surrounds the motif may give more insight knowledge about the amino acids requirements for caspase -3 recognition.

As a control for normal amino acids distribution in human proteins, we used the amino acids distribution in the entire sequence of the 119 proteins of caspase -3 substrates.

- **10-Motif-10**

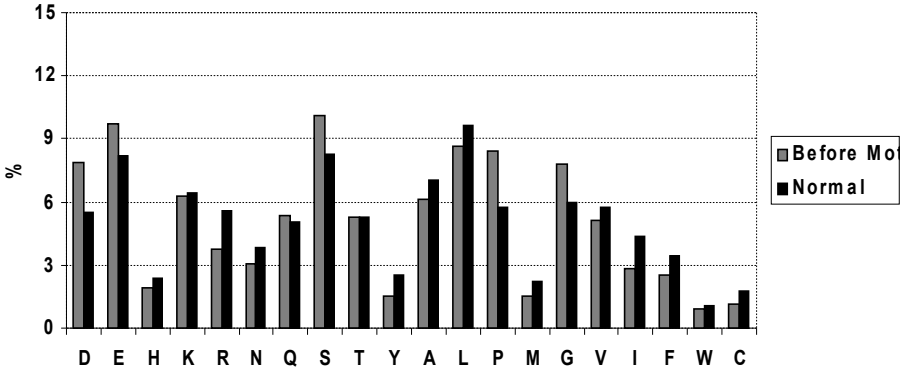

Figure 3.6 (a) Average % of each amino acids 10-before the motif vs. Normal amino acids%. The "Before motif" bars show the % of amino acids in the region surround the motif, while the "normal" bars show the % of amino acids distribution in all 119 substrates.

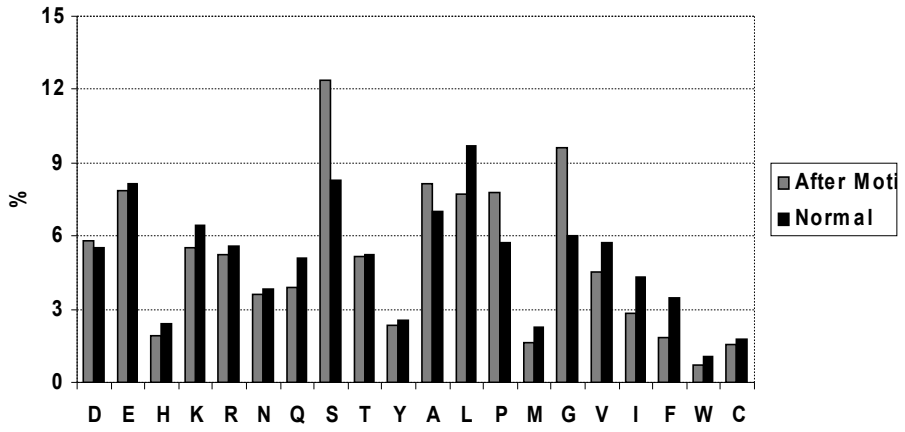

Figure 3.6 (b) Average % of each amino acids 10-after the motif vs. Normal Amino acids %. The "After motif" bars show the % of amino acids in the region surround the motif, while the "normal" bars show the % of amino acids distribution in all 119 substrates.

- 5-Motif-5

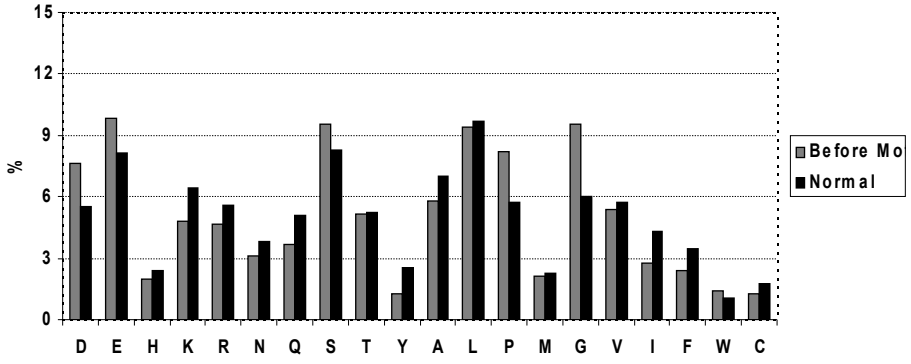

Figure 3.7 (a) Average % of each amino acid "5-before motif" vs. Normal%. The "Before motif" bars show the % of amino acids in the region surround the motif, while the "normal" bars show the % of amino acids distribution in all 119 substrates.

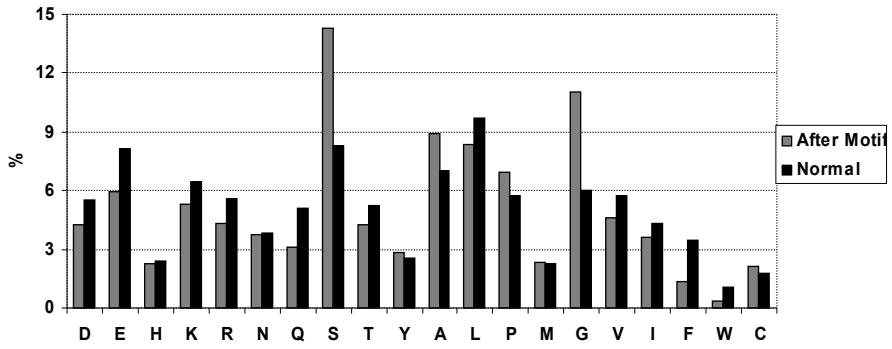

Figure 3.7 (b) Average % of each amino acid "5-after motif" vs. Normal %. The "After motif" bars show the % of amino acids in the region surround the motif, while the "normal" bars show the % of amino acids distribution in all 119 substrates.

3.3 Analysis of the amino acids content inside the motif

The 119 substrates of caspase-3 have 136 motifs as some proteins have more than one cleavage site. The cleavage site is composed of four amino acids P4, P3, P2, and P1. P1 is the amino acid where cleavage process takes place after (always Aspartic acid "D"). The following is an analysis for all 136-cleavage sites:

- Amino acids chemical groups

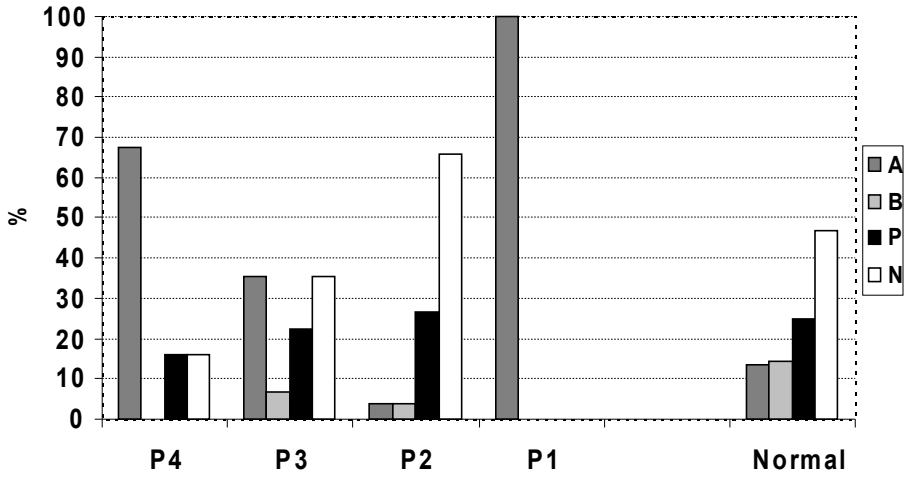

Figure 3.8 (a) % of amino acids chemical groups for the 136 motifs. A is for acidic, B for basic, P for polar, and N for nonpolar amino acids.

- Hydrophobic and Hydrophilic

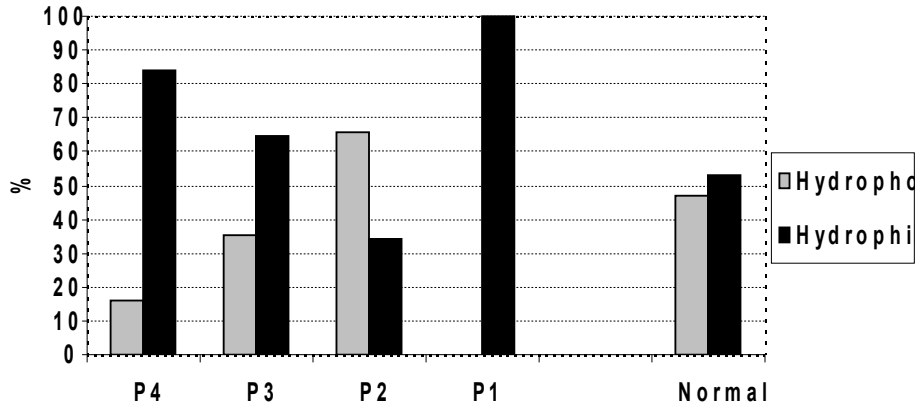

Figure 3.8 (b) Hydrophobic and Hydrophilic % of amino acids in Motifs. Hydrophobic are nonpolar amino acids, while Hydrophilic are acidic, basic, and polar amino acids.

3.4 Distribution matrix around the motif

Amino acids percentages in the positions (P14-P'10): Table 3.1 shows the % of each amino acid in each position in the region around the cleavage site. Fourteen amino acids were analyzed, four amino acid that form the motif "P4-P3-P2-P1" with 10 amino acids before the motif and 10 amino acids after the motif.

Table 3.1 Amino acids % in the positions (P14-P'10)

	Acidic		Basic			Polar					Nonpolar									
P(i)	D	E	H	K	R	N	Q	S	T	Y	A	L	P	M	G	V	I	F	W	C
P14	6.7	11.2	1.5	4.5	1.5	2.2	7.5	8.2	6.7	2.2	6.0	9.0	9.7	0.7	3.7	6.7	4.5	5.2	0.0	2.2
P13	8.9	6.7	1.5	8.1	3.7	3.7	8.9	9.6	8.1	3.0	7.4	6.7	7.4	2.2	5.2	3.0	3.7	0.0	0.7	1.5
P12	11.1	8.1	2.2	10.4	3.0	1.5	7.4	11.9	3.0	1.5	5.2	8.9	7.4	0.7	5.2	5.2	2.2	3.7	1.5	0.0
P11	5.1	11.0	0.7	7.4	2.2	1.5	3.7	15.4	5.9	0.0	5.9	7.4	11.0	1.5	9.6	5.1	4.4	2.2	0.0	0.0
P10	10.3	8.8	2.2	11.0	5.9	3.7	10.3	8.1	3.7	2.2	6.6	5.9	8.8	0.0	5.1	2.9	1.5	0.7	0.0	2.2
P9	7.4	11.0	4.4	4.4	5.1	2.9	2.9	8.1	4.4	0.7	8.1	10.3	5.1	2.2	11.0	6.6	0.7	1.5	1.5	1.5
P8	4.4	12.5	2.2	5.1	3.7	4.4	0.7	8.1	6.6	1.5	5.9	8.8	10.3	2.2	10.3	3.7	4.4	2.2	1.5	1.5
P7	6.6	8.8	2.9	6.6	8.8	1.5	0.7	9.6	5.9	1.5	5.1	12.5	5.9	2.2	9.6	5.1	0.7	3.7	0.7	1.5
P6	6.6	5.9	0.0	2.9	3.7	2.2	6.6	13.2	5.9	1.5	8.1	10.3	9.6	0.0	13.2	2.9	1.5	2.2	1.5	2.2
P5	14.0	12.5	0.0	3.7	2.9	5.1	5.1	8.8	5.1	1.5	3.7	7.4	5.9	3.7	5.1	5.9	5.1	2.2	1.5	0.7
P4	**66.9**	5.1	0.0	0.0	0.0	0.7	0.7	9.6	1.5	1.5	3.7	2.2	1.5	0.0	0.7	4.4	0.0	0.7	0.0	0.7
P3	3.7	**32.4**	3.7	1.5	2.9	2.2	2.9	9.6	3.7	1.5	5.9	8.8	0.7	2.9	2.9	8.1	2.9	2.9	0.0	0.7
P2	2.2	2.2	1.5	0.0	2.2	2.2	2.2	5.1	15.4	2.2	2.9	11.0	13.2	3.7	4.4	**22.1**	5.1	1.5	0.0	0.7
P1	**100.0**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
P'1	2.2	0.7	2.9	1.5	3.7	6.6	0.0	**23.5**	2.2	3.7	6.6	5.1	1.5	0.0	**27.2**	1.5	3.7	2.9	0.0	4.4
P'2	4.4	4.4	2.9	9.6	2.2	1.5	3.7	10.3	2.2	2.2	10.3	9.6	10.3	1.5	11.0	8.8	3.7	0.0	0.7	0.7
P'3	4.4	6.6	1.5	5.1	7.4	3.7	1.5	9.6	7.4	1.5	9.6	8.8	7.4	2.2	7.4	5.9	4.4	1.5	0.0	4.4
P'4	3.7	8.8	2.2	5.9	4.4	3.7	2.9	13.2	7.4	3.7	8.1	9.6	5.1	5.1	7.4	2.2	3.7	1.5	0.0	1.5
P'5	8.1	8.8	1.5	6.6	4.4	3.7	5.1	11.8	3.7	4.4	8.1	6.6	11.0	2.2	4.4	4.4	4.4	0.0	0.0	0.7
P'6	5.9	12.5	1.5	5.1	3.7	2.2	3.7	9.6	3.7	1.5	11.8	7.4	8.8	1.5	10.3	5.9	2.9	1.5	0.0	0.7
P'7	5.1	6.6	0.7	8.8	8.8	3.7	4.4	9.6	12.5	1.5	6.6	6.6	5.1	0.0	11.0	0.7	1.5	3.7	2.2	0.7
P'8	6.6	11.0	2.9	8.8	5.1	4.4	1.5	9.6	5.1	1.5	5.9	7.4	11.0	2.9	4.4	2.9	4.4	2.9	0.0	1.5
P'9	5.9	9.6	2.2	3.7	5.9	4.4	6.6	14.7	4.4	3.7	6.6	5.9	9.6	0.0	8.8	5.1	0.7	0.7	1.5	0.0
P'10	13.2	9.6	2.2	5.1	5.9	3.7	2.9	9.6	5.9	0.0	5.9	7.4	6.6	0.7	7.4	5.1	2.2	3.7	2.2	0.7

3.5 **Amino acids content at each position in the region (P9-P'5)**

Analysis of the percentage of each amino acid in the region (P14-P'10) was reduced to become only in the region (P9-P'5). The reason of this reduction is to focus more on the region close to P1. We call each string a 'P14' peptide as it is composed from 14 amino acids.

The 136 'P14' peptides that founded in the 119 proteins were separated in two groups, each group have its own score matrix. The two groups are:

- *DxxD* group: this group contains all the cleaved strings that their amino acid in P4 is "D" as also in P1. From 136 cleavage sites 91 'P14' peptides had *DxxD* motif forming about 67% from all cleaved strings.
- *xxxD* group: this group contains 45 'P14' peptides which is about 33% from all cleaved strings.

Both groups will give two different percentage tables that will be used in the final algorithm.

The following are the percentage tables for both groups:

Table 3.2 Amino acids % of *DxxD* substrates in the positions (P9-P'5)

	Acidic		Basic			Polar					Nonpolar									
DXXD Cleaved%	D	E	H	K	R	N	Q	S	T	Y	A	L	P	M	G	V	I	F	W	C
P9	8.8	12.1	3.3	4.4	5.5	2.2	1.1	8.8	3.3	1.1	6.6	9.9	5.5	1.1	12.1	8.8	0.0	2.2	2.2	1.1
P8	3.3	12.1	1.1	5.5	3.3	6.6	1.1	7.7	4.4	1.1	6.6	5.5	11.0	2.2	14.3	2.2	5.5	3.3	1.1	2.2
P7	5.5	11.0	3.3	4.4	8.8	1.1	1.1	12.1	6.6	1.1	4.4	14.3	5.5	2.2	7.7	5.5	0.0	2.2	1.1	2.2
P6	8.8	6.6	0.0	4.4	3.3	2.2	6.6	14.3	7.7	2.2	5.5	8.8	11.0	0.0	8.8	2.2	1.1	3.3	2.2	1.1
P5	12.1	15.4	0.0	4.4	3.3	3.3	6.6	7.7	2.2	1.1	4.4	7.7	6.6	4.4	6.6	4.4	3.3	3.3	2.2	1.1
P4	**100.0**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
P3	5.5	**28.6**	2.2	2.2	3.3	2.2	3.3	9.9	5.5	2.2	5.5	7.7	0.0	3.3	4.4	7.7	3.3	3.3	0.0	0.0
P2	3.3	1.1	2.2	0.0	2.2	2.2	2.2	5.5	16.5	3.3	3.3	12.1	11.0	5.5	3.3	**19.8**	4.4	2.2	0.0	0.0
P1	**100.0**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
P'1	3.3	1.1	2.2	1.1	2.2	7.7	0.0	**24.2**	3.3	4.4	4.4	7.7	0.0	0.0	**22.0**	2.2	5.5	3.3	0.0	5.5
P'2	4.4	4.4	3.3	11.0	2.2	1.1	2.2	11.0	3.3	0.0	9.9	11.0	7.7	2.2	12.1	9.9	3.3	0.0	1.1	0.0
P'3	4.4	8.8	2.2	2.2	7.7	4.4	0.0	12.1	8.8	2.2	7.7	8.8	8.8	2.2	6.6	3.3	3.3	1.1	0.0	5.5
P'4	2.2	13.2	3.3	4.4	3.3	4.4	1.1	11.0	8.8	3.3	6.6	11.0	7.7	4.4	6.6	2.2	3.3	1.1	0.0	2.2
P'5	9.9	11.0	1.1	7.7	5.5	4.4	2.2	11.0	4.4	5.5	5.5	7.7	7.7	2.2	4.4	5.5	3.3	0.0	0.0	1.1

Table 3.3 Amino acids % of *xxxD* substrates in the positions (P9-P'5)

	Acidic		Basic			Polar					Nonpolar									
XXXD Cleaved%	D	E	H	K	R	N	Q	S	T	Y	A	L	P	M	G	V	I	F	W	C
P9	4.4	8.9	6.7	4.4	4.4	4.4	6.7	6.7	6.7	0.0	11.1	11.1	4.4	4.4	8.9	2.2	2.2	0.0	0.0	2.2
P8	6.7	13.3	4.4	4.4	4.4	0.0	0.0	8.9	11.1	2.2	4.4	15.6	8.9	2.2	2.2	6.7	2.2	0.0	2.2	0.0
P7	8.9	4.4	2.2	11.1	8.9	2.2	0.0	4.4	4.4	2.2	6.7	8.9	6.7	2.2	13.3	4.4	2.2	6.7	0.0	0.0
P6	2.2	4.4	0.0	0.0	4.4	2.2	6.7	11.1	2.2	0.0	13.3	13.3	6.7	0.0	22.2	4.4	2.2	0.0	0.0	4.4
P5	17.8	6.7	0.0	2.2	2.2	8.9	2.2	11.1	11.1	2.2	2.2	6.7	4.4	2.2	2.2	8.9	8.9	0.0	0.0	0.0
P4	0.0	15.6	0.0	0.0	0.0	2.2	2.2	28.9	4.4	4.4	11.1	6.7	4.4	0.0	2.2	13.3	0.0	2.2	0.0	2.2
P3	0.0	**40.0**	6.7	0.0	2.2	2.2	2.2	8.9	0.0	0.0	6.7	11.1	2.2	2.2	0.0	8.9	2.2	2.2	0.0	2.2
P2	0.0	4.4	0.0	0.0	2.2	2.2	2.2	4.4	13.3	0.0	2.2	8.9	17.8	0.0	6.7	**26.7**	6.7	0.0	0.0	2.2
P1	**100.0**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
P'1	0.0	0.0	4.4	2.2	6.7	4.4	0.0	**22.2**	0.0	2.2	11.1	0.0	4.4	0.0	**37.8**	0.0	0.0	2.2	0.0	2.2
P'2	4.4	4.4	2.2	6.7	2.2	2.2	6.7	8.9	0.0	6.7	11.1	6.7	15.6	0.0	8.9	6.7	4.4	0.0	0.0	2.2
P'3	4.4	2.2	0.0	11.1	6.7	2.2	4.4	4.4	4.4	0.0	13.3	8.9	4.4	2.2	8.9	11.1	6.7	2.2	0.0	2.2
P'4	6.7	0.0	0.0	8.9	6.7	2.2	6.7	17.8	4.4	4.4	11.1	6.7	0.0	6.7	8.9	2.2	4.4	2.2	0.0	0.0
P'5	4.4	4.4	2.2	4.4	2.2	2.2	11.1	13.3	2.2	2.2	13.3	4.4	17.8	2.2	4.4	2.2	6.7	0.0	0.0	0.0

The general percentages of each amino acid in the 136 'P14' peptides, before separation, are showed in the next table:

Table 3.4 Amino acids % of all substrates in the positions (P9-P'5)

	Acidic		Basic			Polar					Nonpolar									
P(i)	D	E	H	K	R	N	Q	S	T	Y	A	L	P	M	G	V	I	F	W	C
P9	7.4	11.0	4.4	4.4	5.1	2.9	2.9	8.1	4.4	0.7	8.1	10.3	5.1	2.2	11.0	6.6	0.7	1.5	1.5	1.5
P8	4.4	12.5	2.2	5.1	3.7	4.4	0.7	8.1	6.6	1.5	5.9	8.8	10.3	2.2	10.3	3.7	4.4	2.2	1.5	1.5
P7	6.6	8.8	2.9	6.6	8.8	1.5	0.7	9.6	5.9	1.5	5.1	12.5	5.9	2.2	9.6	5.1	0.7	3.7	0.7	1.5
P6	6.6	5.9	0.0	2.9	3.7	2.2	6.6	13.2	5.9	1.5	8.1	10.3	9.6	0.0	13.2	2.9	1.5	2.2	1.5	2.2
P5	14.0	12.5	0.0	3.7	2.9	5.1	5.1	8.8	5.1	1.5	3.7	7.4	5.9	3.7	5.1	5.9	5.1	2.2	1.5	0.7
P4	**66.9**	5.1	0.0	0.0	0.0	0.7	0.7	9.6	1.5	1.5	3.7	2.2	1.5	0.0	0.7	4.4	0.0	0.7	0.0	0.7
P3	3.7	**32.4**	3.7	1.5	2.9	2.2	2.9	9.6	3.7	1.5	5.9	8.8	0.7	2.9	2.9	8.1	2.9	2.9	0.0	0.7
P2	2.2	2.2	1.5	0.0	2.2	2.2	2.2	5.1	15.4	2.2	2.9	11.0	13.2	3.7	4.4	**22.1**	5.1	1.5	0.0	0.7
P1	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
P'1	2.2	0.7	2.9	1.5	3.7	6.6	0.0	23.5	2.2	3.7	6.6	5.1	1.5	0.0	**27.2**	1.5	3.7	2.9	0.0	4.4
P'2	4.4	4.4	2.9	9.6	2.2	1.5	3.7	10.3	2.2	2.2	10.3	9.6	10.3	1.5	11.0	8.8	3.7	0.0	0.7	0.7
P'3	4.4	6.6	1.5	5.1	7.4	3.7	1.5	9.6	7.4	1.5	9.6	8.8	7.4	2.2	7.4	5.9	4.4	1.5	0.0	4.4
P'4	3.7	8.8	2.2	5.9	4.4	3.7	2.9	13.2	7.4	3.7	8.1	9.6	5.1	5.1	7.4	2.2	3.7	1.5	0.0	1.5
P'5	8.1	8.8	1.5	6.6	4.4	3.7	5.1	11.8	3.7	4.4	8.1	6.6	11.0	2.2	4.4	4.4	4.4	0.0	0.0	0.7

3.6 **Uncleaved strings**

Although the analysis of amino acids distribution adjacent to the cleavage sites showed importance of certain amino acids, analysis of the 'P14' peptides containing cleavage sites would not enable us to deduce the measurable weight for a given amino acid at a certain residue. Therefore, these results should be compared with a set of 'P14' peptide sequences that contain 'D' and they were experimentally shown to be uncleaved.

The cleaved proteins could serve as a very good control of uncleaved sites. Excluding all the natural cleavage sites "136" from these proteins gave us a rich database of uncleaved sites.

The sequences of these proteins have been searched for any aspartic acid 'D' and a same string length of amino acids. All the 136 cleavage sites were excluded from this database. The result was all uncleaved sites in a database of 5538 strings.

From those 5538 uncleaved strings, only strings that fulfill the requirements; strings of 14 amino acids (5-M[1]-5), where selected for analysis. Strings with length shorter than 14 amino acids were excluded. Out of 5538, we exclude 144 strings having 5394 'P14' peptides to analyze.

These 5394 'P14' peptides were divided into two groups: *DxxD* and *xxxD* groups. The *DxxD* group of uncleaved contains 333 'P14' peptides, while *xxxD* group contains 5061 'P14' peptides with a percentage of 6% and 94% for both groups respectively.

3.6.1 **Uncleaved data results**

[1] M: (motif)

The following tables show the percentages of each amino acid at each position in the 'P14' peptides.

- General percentages of uncleaved:" in 5349 strings"

Table 3.5: Amino acids % of all uncleaved strings in the positions (P9-P'5)

Uncleaved	Acidic		Basic			Polar					Nonpolar									
P(i)	D	E	H	K	R	N	Q	S	T	Y	A	L	P	M	G	V	I	F	W	C
P9	6.0	9.0	2.8	7.0	5.8	4.1	4.6	8.0	4.5	2.8	6.1	9.9	5.2	2.2	5.5	6.3	4.1	3.4	0.8	1.8
P8	6.2	8.8	2.1	7.1	5.9	4.1	5.4	7.9	4.7	2.6	6.4	9.6	5.2	2.3	5.7	5.6	4.4	3.2	0.9	1.7
P7	6.0	8.1	2.2	6.9	5.5	4.0	4.4	8.1	5.2	3.2	6.1	10.5	4.8	2.3	5.7	5.6	5.1	3.3	1.1	1.9
P6	6.4	7.8	2.8	6.1	5.6	3.4	4.4	8.5	5.3	3.2	6.6	10.2	4.8	2.3	5.6	5.7	4.5	3.9	1.1	1.9
P5	6.3	7.7	2.3	6.7	5.9	4.1	5.9	8.1	5.1	2.3	7.0	9.2	5.3	1.9	5.8	6.0	4.6	3.4	1.0	1.4
P4	6.2	8.0	1.9	7.8	5.4	4.2	4.3	8.6	5.2	2.6	6.3	10.0	5.3	2.3	5.7	5.5	4.3	3.9	1.1	1.5
P3	7.2	7.4	2.7	5.5	4.5	4.2	4.2	8.3	4.6	2.7	6.7	**11.2**	5.2	2.2	6.2	6.1	4.8	3.6	1.1	1.6
P2	6.6	**10.2**	1.8	7.1	6.7	3.3	4.9	7.2	4.7	2.6	5.9	9.9	4.6	2.2	6.3	5.5	4.5	3.5	1.2	1.4
P1	**100.0**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
P'1	6.6	8.7	1.9	5.4	5.4	3.9	3.9	8.0	4.9	3.2	6.0	**10.5**	4.7	1.9	5.4	6.2	5.5	4.7	1.5	1.7
P'2	7.2	9.1	1.9	6.1	4.6	3.8	3.7	7.4	4.4	3.2	6.3	9.9	4.3	2.5	5.5	7.0	5.4	4.0	1.7	1.8
P'3	7.6	8.3	2.2	7.3	6.8	4.4	4.3	7.8	4.7	2.6	6.5	8.8	4.9	2.5	5.5	5.5	4.3	3.5	1.0	1.5
P'4	6.7	8.8	2.5	7.0	5.7	3.9	4.8	7.6	5.0	2.6	7.2	9.7	4.1	2.2	5.2	6.0	4.3	3.8	1.4	1.6
P'5	6.5	8.1	2.2	7.2	6.0	3.6	4.5	7.2	4.6	2.5	6.0	10.6	5.0	2.3	5.7	6.2	5.2	3.5	1.3	1.8

- *DxxD* group of uncleaved: "333 strings"

Table 3.6: Amino acids % of uncleaved strings of *DxxD* type in the positions (P9-P'5)

Uncleaved	Acidic		Basic			Polar					Nonpolar									
DxxD	D	E	H	K	R	N	Q	S	T	Y	A	L	P	M	G	V	I	F	W	C
P9	7.5	9.3	2.7	6.6	6.6	3.0	3.9	10.5	5.1	3.9	5.7	8.4	3.9	1.5	3.9	5.4	4.8	5.4	0.6	1.2
P8	8.4	7.8	0.9	6.6	7.2	3.3	7.8	5.1	3.9	2.7	7.5	12.6	4.2	3.0	5.4	3.6	4.8	2.7	0.9	1.5
P7	9.9	7.2	2.1	9.0	5.7	4.8	5.1	8.7	5.4	2.4	6.3	8.4	3.9	1.5	4.8	2.1	5.7	5.1	0.6	1.2
P6	10.5	11.4	2.1	3.6	3.0	4.2	3.6	7.5	3.6	3.6	5.7	14.7	3.0	1.5	3.0	6.3	6.9	2.7	1.2	1.8
P5	7.8	8.4	2.1	6.6	8.1	2.1	3.9	8.4	4.8	1.5	5.4	9.9	3.3	3.3	7.8	6.0	3.6	3.9	1.8	1.2
P4	**100.0**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
P3	**12.3**	6.6	3.0	6.0	5.4	4.5	3.0	6.3	3.3	2.7	4.8	12.0	5.1	1.2	3.0	6.6	5.7	5.1	2.1	1.2
P2	**12.6**	**12.0**	1.5	6.6	5.1	4.2	1.5	5.1	3.9	5.1	2.4	9.3	2.4	2.7	4.5	7.2	5.4	5.1	2.4	0.9
P1	**100.0**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
P'1	**10.5**	8.1	2.4	4.8	4.8	5.7	4.2	7.8	4.5	2.7	5.7	6.3	2.1	3.0	6.6	6.3	7.2	4.5	1.8	0.9
P'2	**11.1**	9.9	0.6	6.6	3.3	3.9	4.5	6.3	3.6	3.6	3.6	**11.7**	5.7	2.7	3.0	8.7	4.5	4.2	1.2	1.2
P'3	9.0	7.8	1.8	7.2	7.8	4.5	5.1	6.6	5.1	2.7	4.2	6.9	7.8	3.3	4.8	5.1	2.7	4.5	1.8	1.2
P'4	9.3	8.4	2.4	6.9	6.0	3.0	4.5	6.3	6.6	3.0	4.2	8.1	4.2	2.7	5.1	6.3	3.6	4.2	3.3	1.8
P'5	6.6	11.1	1.8	6.9	5.4	3.9	3.6	9.0	6.0	2.4	5.4	10.8	4.5	2.1	4.8	5.1	6.0	3.6	0.6	0.3

- *xxxD* group of uncleaved: "5061 strings"

Table 3.7 Amino acids % of uncleaved strings of xxxD type in the positions (P9-P'5)

Uncleaved	Acidic		Basic			Polar					Nonpolar									
xxxD	D	E	H	K	R	N	Q	S	T	Y	A	L	P	M	G	V	I	F	W	C
P9	5.8	9.0	2.8	7.0	5.7	4.1	4.7	7.9	4.5	2.8	6.1	10.0	5.3	2.3	5.6	6.3	4.1	3.3	0.8	1.8
P8	6.1	8.8	2.2	7.2	5.8	4.1	5.3	8.1	4.8	2.6	6.3	9.4	5.2	2.3	5.7	5.7	4.4	3.2	0.9	1.7
P7	5.8	8.2	2.3	6.8	5.5	3.9	4.4	8.0	5.2	3.2	6.1	10.6	4.8	2.3	5.7	5.8	5.1	3.2	1.1	1.9
P6	6.2	7.5	2.8	6.3	5.7	3.3	4.5	8.5	5.5	3.2	6.6	9.9	4.9	2.4	5.8	5.7	4.3	3.9	1.1	1.9
P5	6.2	7.6	2.3	6.7	5.7	4.3	6.0	8.1	5.2	2.3	7.1	9.1	5.5	1.8	5.7	6.0	4.7	3.4	0.9	1.5
P4	**0.0**	8.5	2.1	8.3	5.7	4.5	4.6	9.1	5.6	2.7	6.7	10.6	5.6	2.5	6.0	5.9	4.5	4.2	1.2	1.6
P3	6.9	7.4	2.7	5.5	4.5	4.1	4.3	8.4	4.7	2.6	6.9	**11.1**	5.2	2.3	6.4	6.1	4.8	3.5	1.0	1.7
P2	6.2	**10.1**	1.9	7.1	6.8	3.2	5.1	7.4	4.7	2.5	6.2	**9.9**	4.7	2.1	6.4	5.3	4.5	3.4	1.2	1.4
P1	**100.0**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
P'1	6.3	8.8	1.9	5.5	5.4	3.8	3.9	8.0	5.0	3.2	6.0	**10.8**	4.9	1.9	5.3	6.1	5.4	4.7	1.4	1.7
P'2	7.0	9.0	2.0	6.1	4.7	3.8	3.7	7.5	4.4	3.2	6.4	9.8	4.2	2.5	5.7	6.9	5.5	4.0	1.8	1.8
P'3	7.5	8.3	2.2	7.3	6.7	4.4	4.3	7.9	4.7	2.6	6.6	8.9	4.7	2.5	5.5	5.6	4.4	3.4	0.9	1.5
P'4	6.5	8.8	2.5	7.0	5.6	4.0	4.8	7.7	4.9	2.6	7.4	9.8	4.1	2.2	5.2	6.0	4.3	3.7	1.2	1.6
P'5	6.4	7.9	2.2	7.2	6.0	3.6	4.5	7.1	4.5	2.5	6.0	10.6	5.1	2.4	5.7	6.3	5.1	3.5	1.3	1.9

3.7 **Score matrices**

Score matrices were generated from the subtraction of uncleaved strings matrix values from those of cleaved at the positions (P9-P'5). The result was three score matrices:

- General score matrix:

Table 3.8 General Score Matrix

General SM	D	E	H	K	R	N	Q	S	T	Y	A	L	P	M	G	V	I	F	W	C
P9	1.4	2.0	1.6	-2.6	-0.6	-1.1	-1.7	0.1	-0.1	-2.1	2.0	0.4	0.0	0.0	5.5	0.4	-3.4	-2.0	0.6	-0.3
P8	-1.8	3.7	0.1	-2.0	-2.3	0.3	-4.7	0.2	1.9	-1.2	-0.5	-0.8	5.1	-0.1	4.6	-1.9	0.0	-1.0	0.5	-0.3
P7	0.6	0.7	0.7	-0.3	3.3	-2.5	-3.7	1.5	0.7	-1.7	-1.0	2.0	1.1	-0.1	3.9	-0.5	-4.4	0.3	-0.3	-0.4
P6	0.2	-1.9	-2.8	-3.2	-1.9	-1.1	2.2	4.8	0.5	-1.7	1.5	0.1	4.8	-2.3	7.6	-2.8	-3.0	-1.7	0.4	0.3
P5	7.6	4.8	-2.3	-3.0	-3.0	1.0	-0.7	0.7	0.0	-0.8	-3.3	-1.8	0.5	1.8	-0.6	-0.1	0.5	-1.2	0.5	-0.7
P4	60.7	-2.9	-1.9	-7.8	-5.4	-3.5	-3.6	1.0	-3.8	-1.1	-2.6	-7.7	-3.8	-2.3	-4.9	-1.1	-4.3	-3.2	-1.1	-0.7
P3	-3.5	25.0	0.9	-4.0	-1.6	-1.9	-1.3	1.3	-0.9	-1.2	-0.8	-2.4	-4.4	0.7	-3.3	2.0	-1.9	-0.6	-1.1	-0.9
P2	-4.4	-8.0	-0.4	-7.1	-4.4	-1.1	-2.7	-2.1	10.8	-0.4	-3.0	1.1	8.7	1.5	-1.9	16.6	0.6	-2.0	-1.2	-0.7
P1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
P'1	-4.4	-8.0	1.1	-3.9	-1.7	2.7	-3.9	15.5	-2.7	0.5	0.6	-5.4	-3.3	-1.9	21.8	-4.7	-1.8	-1.8	-1.5	2.7
P'2	-2.8	-4.7	1.1	3.4	-2.4	-2.4	0.0	2.9	-2.2	-1.0	4.0	-0.3	6.0	-1.1	5.5	1.8	-1.7	-4.0	-1.0	-1.0
P'3	-3.2	-1.7	-0.7	-2.1	0.6	-0.7	-2.9	1.8	2.6	-1.2	3.1	0.1	2.5	-0.3	1.9	0.3	0.1	-2.0	-1.0	2.9
P'4	-3.0	0.1	-0.3	-1.1	-1.2	-0.2	-1.9	5.7	2.3	1.1	0.9	-0.1	1.0	3.0	2.1	-3.8	-0.6	-2.3	-1.4	-0.2
P'5	1.6	0.8	-0.7	-0.6	-1.6	0.0	0.7	4.5	-0.9	1.9	2.1	-3.9	6.0	-0.1	-1.3	-1.8	-0.8	-3.5	-1.3	-1.0

General P(i) diff.= general P(i) cleaved- general P(i) uncleaved

- *DxxD* score matrix:

Table 3.9 *DxxD* Score Matrix

DxxD SM	D	E	H	K	R	N	Q	S	T	Y	A	L	P	M	G	V	I	F	W	C
P9	1.3	2.8	0.6	-2.2	-1.1	-0.8	-2.8	-1.7	-1.8	-2.8	0.9	1.5	1.6	-0.4	8.2	3.4	-4.8	-3.2	1.6	-0.1
P8	-5.1	4.3	0.2	-1.1	-3.9	3.3	-6.7	2.6	0.5	-1.6	-0.9	-7.1	6.8	-0.8	8.9	-1.4	0.7	0.6	0.2	0.7
P7	-4.4	3.8	1.2	-4.6	3.1	-3.7	-4.0	3.4	1.2	-1.3	-1.9	5.9	1.6	0.7	2.9	3.4	-5.7	-2.9	0.5	1.0
P6	-1.7	-4.8	-2.1	0.8	0.3	-2.0	3.0	6.8	4.1	-1.4	-0.2	-5.9	8.0	-1.5	5.8	-4.1	-5.8	0.6	1.0	-0.7
P5	4.3	7.0	-2.1	-2.2	-4.8	1.2	2.7	-0.7	-2.6	-0.4	-1.0	-2.2	3.3	1.1	-1.2	-1.6	-0.3	-0.6	0.4	-0.1
P4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
P3	-6.8	22.0	-0.8	-3.8	-2.1	-2.3	0.3	3.6	2.2	-0.5	0.7	-4.3	-5.1	2.1	1.4	1.1	-2.4	-1.8	-2.1	-1.2
P2	-9.3	-10.9	0.7	-6.6	-2.9	-2.0	0.7	0.4	12.6	-1.8	0.9	2.8	8.6	2.8	-1.2	12.6	-1.0	-2.9	-2.4	-0.9
P1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
P'1	-7.2	-7.0	-0.2	-3.7	-2.6	2.0	-4.2	16.4	-1.2	1.7	-1.3	1.4	-2.1	-3.0	15.4	-4.1	-1.7	-1.2	-1.8	4.6
P'2	-6.7	-5.5	2.7	4.4	-1.1	-2.8	-2.3	4.7	-0.3	-3.6	6.3	-0.7	2.0	-0.5	9.1	1.2	-1.2	-4.2	-0.1	-1.2
P'3	-4.6	1.0	0.4	-5.0	-0.1	-0.1	-5.1	5.5	3.7	-0.5	3.5	1.9	1.0	-1.1	1.8	-1.8	0.6	-3.4	-1.8	4.3
P'4	-7.1	4.8	0.9	-2.5	-2.7	1.4	-3.4	4.7	2.2	0.3	2.4	2.9	3.5	1.7	1.5	-4.1	-0.3	-3.1	-3.3	0.4
P'5	3.3	-0.1	-0.7	0.8	0.1	0.5	-1.4	2.0	-1.6	3.1	0.1	-3.1	3.2	0.1	-0.4	0.4	-2.7	-3.6	-0.6	0.8

DxxD P(i) diff.= *DxxD* P(i) cleaved- *DxxD* P(i) uncleaved

- *xxxD* score matrix:

Table 3.10 *xxxD* Score Matrix

xxxD SM	D	E	H	K	R	N	Q	S	T	Y	A	L	P	M	G	V	I	F	W	C
P9	-1.4	-0.1	3.9	-2.6	-1.3	0.3	2.0	-1.2	2.2	-2.8	5.0	1.1	-0.8	2.2	3.3	-4.1	-1.9	-3.3	-0.8	0.4
P8	0.6	4.5	2.2	-2.7	-1.4	-4.1	-5.3	0.8	6.3	-0.4	-1.9	6.2	3.7	-0.1	-3.5	1.0	-2.1	-3.2	1.3	-1.7
P7	3.1	-3.7	0.0	4.3	3.4	-1.7	-4.4	-3.6	-0.7	-1.0	0.6	-1.7	1.8	-0.1	7.6	-1.4	-2.8	3.4	-1.1	-1.9
P6	-3.9	-3.1	-2.8	-6.3	-1.3	-1.1	2.2	2.6	-3.2	-3.2	6.7	3.4	1.7	-2.4	16.4	-1.2	-2.1	-3.9	-1.1	2.6
P5	11.6	-1.0	-2.3	-4.5	-3.5	4.6	-3.8	3.0	6.0	-0.1	-4.9	-2.5	-1.0	0.4	-3.4	2.9	4.2	-3.4	-0.9	-1.5
P4	0.0	7.0	-2.1	-8.3	-5.7	-2.2	-2.4	19.7	-1.1	1.7	4.4	-3.9	-1.2	-2.5	-3.8	7.5	-4.5	-2.0	-1.2	0.7
P3	-6.9	32.6	3.9	-5.5	-2.2	-1.9	-2.1	0.5	-4.7	-2.6	-0.2	0.0	-2.9	-0.1	-6.4	2.8	-2.5	-1.2	-1.0	0.6
P2	-6.2	-5.7	-1.9	-7.1	-4.5	-1.0	-2.9	-2.9	8.6	-2.5	-3.9	-1.0	13.1	-2.1	0.2	21.3	2.2	-3.4	-1.2	0.8
P1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
P'1	-6.3	-8.8	2.6	-3.2	1.2	0.7	-3.9	14.2	-5.0	-1.0	5.1	-10.8	-0.5	-1.9	32.4	-6.1	-5.4	-2.5	-1.4	0.5
P'2	-2.6	-4.6	0.3	0.6	-2.4	-1.6	3.0	1.4	-4.4	3.4	4.7	-3.1	11.3	-2.5	3.2	-0.3	-1.0	-4.0	-1.8	0.4
P'3	-3.1	-6.1	-2.2	3.8	-0.1	-2.2	0.2	-3.4	-0.3	-2.6	6.7	0.0	-0.2	-0.2	3.4	5.5	2.2	-1.2	-0.9	0.7
P'4	0.1	-8.8	-2.5	1.9	1.0	-1.7	1.8	10.1	-0.5	1.9	3.7	-3.1	-4.1	4.5	3.7	-3.8	0.1	-1.5	-1.2	-1.6
P'5	-2.0	-3.4	0.0	-2.8	-3.8	-1.4	6.6	6.2	-2.3	-0.3	7.3	-6.1	12.7	-0.1	-1.3	-4.1	1.5	-3.5	-1.3	-1.9

xxxD P(i) diff.= *xxxD* P(i) cleaved- *xxxD* P(i) uncleaved

Chapter 4: CAT3 Algorithm

In this chapter, we will describe the following:

- Programming CAT3 algorithm
- Description of the code and how it works
- Processing an example
- Efficiency of the CAT3:
  - Comparison between CAT3 and other cleavage tools
  - Sensitivity and Specificity of CAT3

4.1 Programming CAT3 algorithm

CAT3 was programmed using Perl language. A Perl editor software "called Perl Builder v.2" was helpful in writing the code. The editing software makes it easy to run and debug the algorithm. To simplify the use of CAT3, software "called Perl2exe" was used to make it as executable program that could be run under windows just by clicking its icon. The algorithm uses two sub-algorithms; in programming languages, they are called sub-modules.

The program for CAT3 starts with a user-friendly screen. The user can enter any protein sequence directly in a text field or as a file by browsing the targeted source. The input -protein sequence – should be in fasta format either it is a text input or a file.

The final output of CAT3 is a text file that its name is the name of the protein. The output text file will contains the following:

1. PROTEIN INFORMATION: all the information about the protein, which includes:
    - Protein Name: the SWISS-PROT id of the protein
    - Protein Length: the number of amino acids in the protein
    - Number of Aspartic acids in the protein: number of 'D's
    - Positions of the Aspartic acids: positions of the 'D's
2. CLEAVAGE PREDICTION ANALYSIS: all the Aspartic acids with their surrounding amino acids, 'P14' peptides, are shown in this part in an organized way. Each 'P14' peptide has its own score. The 'P14' peptide with the highest score is the probable cleavage site for the entered protein if its score is (> 30).

3. SIGNIFICANT CLEAVAGE SITE: this part shows the whole protein sequence with the mark '><' next to the 'D' where cleavage process mostly will take place.

COMMENTS: this part shows the criteria we use in deciding if the 'P14' peptide is probable to be cleaved or not according to its score. Contact information is also found in this part.

Despite the next section of this chapter explains the criteria that the algorithm considers in calculating the final score for any motif, a step-by-step explanation is founded in the code lines. The explanation lines were marked with'#' in the beginning of each line.

The complete code of CAT3 and its sub-codes are in Appendices chapter.

4.2 Description of the algorithm

The algorithm of CAT3 depends on three values or scores that their average summation will give the final score. The decision wither a protein is a caspase-3 substrate or not is taken according to the final score of its 'P14' peptides. The highest the final score is the higher the possibility of cleavage of that 'P14' peptide would be.

The three scores are:

1. **Score A:** this score is a specific score as it depends on the *DxxD* and *xxxD* score matrices but not on the general matrix of probability differences. Score A is the average of two scores: score1, and score2. Both scores retrieve their values from the corresponding matrix, *DxxD* or *xxxD*, depending on the 'P4' amino acid. Score1 and score2 are the average summation of the amino acids values in the positions (P5-P'2) and (P9-P'5) respectively. Because of its significant effect on cleavage process, the region (P5-P'2) has a double influence on score-A.

$$\textbf{Score A} = \frac{\textbf{Score1+Score2}}{\textbf{2}}$$

There are two cases of this score depending on the amino acid 'P4'.

Case 1: if 'P4' is not 'D' but any other amino acid:

$$\textbf{Score1 } \textit{xxxD} = \left(\frac{\textbf{P5+P4+P3+P2+P'1+P'2}}{\textbf{6}} \right. \left. \textbf{/ 21.5} \right) * \textbf{100\%}$$

$$\textbf{Score2 } \textit{xxxD} = \frac{(P9+P8+P7+P6+P5+P4+\ldots\ldots+P'5 \,/\, 14.9) * 100\%}{13}$$

Were 21.5 and 14.9 are the highest average scores that could be obtained from the *xxxD* matrix according to their amino acids positions.

Case 2: if the 'P4' is 'D': the values are calculated from the *DxxD* matrix. Here the P4 score does not have any value, as it is always 'D'; however, a value of 5 was added to the average score instead.

$$\textbf{Score1 } \textit{DxxD} = \frac{(\{P5+P3+P2+P'1+P'2\} + 5 \,/\, 13.4) * 100\%}{5}$$

$$\textbf{Score2 } \textit{DxxD} = \frac{(\{P9+P8+P7+P6+P5+P3+\ldots+P'5\} + 5 \,/ 9.3) * 100\%}{12}$$

The previous values: (13.4) and (9.3) are the highest average scores that could be obtained from the *DxxD* matrix.

2. **Score B:** this is the general score of the string (P9-P'5). The score is calculated from the general matrix of probability differences.

Score B= ({P9+P8+P7+P6+P5+P3+…+P'5} /13.4)* 100%

13

Were (13.4) is the highest average score that could be obtained from the highest amino acid value that found in the general matrix of probability difference.

3. **Score C** "Markov Score": not like the other scores, this score depends on the multiplication, not addition, of the probability of each amino acid in the positions (P6-P'1).

Score C= Ps/Pa

Ps "Probability of the string"= P(6)*P(5)*……*P('1)

The probability of each amino acid in the region (P6-P'1) was taken from its percentage value divided over 100. The percentage values of amino acids either calculated from *DxxD* or *xxxD* are in the tables (3.2) and (3.3) in chapter Results.

Pa "Probability of amino acid" = P(6)*P(5)*……*P('1)

The probability of each amino acid in any position is calculated from the probability of each amino acid in normal analysis.

D	E	H	K	R	N	Q	S	T	Y	A	L	P	M	G	V	I	F	W	C
0.06	0.08	0.02	0.06	0.06	0.04	0.05	0.08	0.05	0.03	0.07	0.10	0.06	0.02	0.06	0.06	0.04	0.03	0.01	0.02

Final score: the average of the three scores (A, B, C).

$$\textbf{Final Score} = \left(\frac{\textbf{Score A + Score B + Score C}}{\textbf{3}} \right. \textbf{/ 91) * 100\%}$$

Where 91 is the highest score obtained after running the program on MEG.

The final score could be negative or positive but in the final calculation the string of the protein is considered to be cleaved by caspase -3 if it has a final score >30. The higher the final score of a string means a higher probability for the cleavage to be happen. CAT3 shows separately in its output the highest score for the motif in the protein.

4.3 CAT3: processing an example

- Input (fasta format)

```
>sp|O02718|BCL2_BOVIN Apoptosis regulator Bcl-2 - Bos taurus (Bovine).
MAHAGGTGYDNREIVMKYIHYKLSQRGYEWDAGDAGAAPPGAAPAPGILSSQPGRTPAPS
RTSPPPPPAAAAGPAPSPVPPVVHLTLRQAGDDFSRRYRRDFAEMSSQLHLTPFTARERF
ATVVEELFRDGVNWGRIVAFFEFGGVMCVESVNREMSPLVDSIALWMTEYLNRHLHTWIQ
DNGGWDAFVELYGPSMRPLFDFSWLSLKALLSLALVGACITLGAYLGHK
```

- Scanning the primary sequence for “P14” peptides

Strings	Motif	Position
AHAGGTGYDNREIV	TGYD	10
LSQRGYEWDAGDAG	YEWD	31
RGYEWDAGDAGAAP	DAGD	34
HLTLRQAGDDFSRR	QAGD	92
LTLRQAGDDFSRRY	AGDD	93
DFSRRYRRDFAEMS	YRRD	101
TVVEELFRDGVNWG	LFRD	130
NREMSPLVDSIALW	PLVD	161
RHLHTWIQDNGGWD	WIQD	181
WIQDNGGWDAFVEL	GGWD	186
GPSMRPLFDFSWLS	PLFD	201

- Scoring each “P14” peptide

Score A	Score B	Score C
Specific score	General Score	Markov score

Table 4.1 Scoring table for a “P14” peptide

Pos.	a.a	Specific score	General score	Ps
P9	A	5	2.0	0.081
P8	H	2.2	0.1	0.022
P7	A	0.6	-1.0	0.051
P6	G	16.4	7.6	0.132
P5	G	-3.4	-0.6	0.051
P4	T	-1.1	-3.8	0.015
P3	G	-6.4	-3.3	0.029
P2	Y	-2.5	-0.4	0.022
P1	D	0	0.0	1
P'1	N	0.7	2.7	0.066
P'2	R	-2.4	-2.4	0.22
P'3	E	-6.1	-1.7	0.066
P'4	I	0.1	-0.6	0.037
P'5	V	-4.1	-1.8	0.044

The shaded cells in the table are the values each score (A, B, and C) is calculated from. The equation for each score is described before in this chapter. The "Pa" value for score "C" is the normal amino acids distribution in MEG; the following chart shows the Pa values:

D	E	H	K	R	N	Q	S	T	Y	A	L	P	M	G	V	I	F	W	C
0.06	0.08	0.02	0.06	0.06	0.04	0.05	0.08	0.05	0.03	0.07	0.10	0.06	0.02	0.06	0.06	0.04	0.03	0.01	0.02

Applying the values in the table in the corresponding equations we have:

Score A= -6.1, Score B= -1.8, Score C= 0.32

Final score= (A+B+C) / 91 * 100% =~ -2 ➔ no cleavage < 30.

3

4.4 Sensitivity and Specificity of CAT3

The accuracy of a classification algorithm can be defined by two criteria: sensitivity and specificity.

$$\text{Sensitivity} = \frac{TP}{TP + FP} \quad \text{while} \quad \text{Specificity} = \frac{TN}{TN + FN},$$

where

TP: is the number of true positives.

FP: is the number of false positives.

TN: is the number of true negatives.

FN: is the number of false positives [Liping Wei 1998].

For our CAT3 tool, sensitivity measures the ability of CAT3 to detect real caspase -3 cleavage sites, while specificity measures the ability of CAT3 to reject regions that are not caspase -3 cleavage sites.

In analyzing the test group –the first 22 substrates were analyzed here as the last 3 substrates were added latterly- we end with 886 'D' of each one resemble a different motif or probable cleavage site. According to CAT3 results and considering positive cleavage sites are motifs with a final score >30, the 886 motifs are classified as follows:

- Motifs with a final score >30 (44): 21 true cleavage sites (TP), and 23 false cleavage sites (FP).
- Motifs with a final score ≤ 30 (842): 839 are non cleavage sites (TN); and true cleavage sites are only 3 (FN).

Sensitivity= 21/ (21+26) * 100 = 48%.

Specificity= 839/ (839+3) *100= 100%.

The result of both sensitivity and specificity are directly affected by our criteria in considering positive cleavage sites with a final score >30. However, CAT3 algorithm was based one minimizing the number of results of probable cleavage sites. In general, the first cleavage site with a final score >30 is considered the main cleavage site of the protein. Therefore, the final sensitivity and specificity are calculated upon a new analysis having two conditions in considering a positive cleavage site:

1. The final score of the motif >30, and
2. The motif with the highest score is the only cleavage site in general, as some substrates have more than one cleavage site.

The second condition will consider all the motifs with a final score below the highest final score as non-cleavage sites. Analysis of the 886 motifs from the test group with the new condition classified them into:

- Motifs with a final score >30 and have the highest final score (24): 20 true cleavage sites (TP), and 4 false cleavage sites (FP).
- Others (862): 858 are non cleavage sites (TN); and 4 true cleavage sites (FN).

Sensitivity= 20/ (20+4) * 100 = 83%.

Specificity= 858/ (858+3) *100= 100%.

4.5 CAT3 predicts unknown cleavage sites

Referring to literature, many substrates are already known to be cleaved by caspases in general but their cleavage site is still unknown

[Fischer et al. 2003]. A group of 44 substrates with their cleavage sites is still unknown were analyzed using CAT3.

Although some literatures did not specify which caspase was responsible for the cleavage, we consider all substrates were cleaved by caspases -3. CAT3 recognized (35/44) substrates with their final score >30 which is about 80%. The highest final score was 82 for the substrate Desmocollin-3 at the motif (DEND↓238), while the lowest final score was for the substrate α-Tubulin at the motif (IQPD↓33). Table 6.5 (in chapter Appendices) shows all the 44 substrates with their final scores.

Chapter 5: Discussion and Conclusions

5.1 Discussion

Analyses of the regions that surrounded the cleavage sites according to their secondary structure and chemical properties showed unstructured and non-specific physiochemical properties in these regions. This may help in explaining the process of caspase -3 in binding and cleaving its target proteins.

We concluded that the region around the cleavage site must be more or less a loop structure that form as a joint between the two cleaved units. This coiled polypeptide region in most of the cleaved sites, and its peripheral location, as we assume, would assist caspase -3 cleavage processes [Garay-Malpartida et al. 2005].

The search of the region of interest looking for common features patterns -by analyzing these regions into their physiochemical properties and amino acids frequencies- shows that caspase -3 cleavage process depends mainly on short strings before and after the cleavage sites that do not exceeds more than 10 amino acids in length. Compared to normal, both regions 10 and 5 amino acids before and after the motif show significant amino acids differences. Therefore, we consider that this

region (up to 10 amino acids from the motif) plays an important role in caspase -3 recognizing the cleavage site.

Analysis of amino acids distribution in the regions 10 and 5 amino acids before and after the cleavage sites shows a high percentage of serine (S). This could be a kind of control of cleavage as phosphorylated serine residues affect proteolysis [Tozser et al. 2003].

To increase the efficiency and the specificity to the final algorithm we divided the 136 cleavage sites in MEG in two groups according to P4. All the 14 amino acids peptides that their P4 is 'D' (91/136) formed the DxxD group, while the rest (45/136) formed the xxxD group. This separation between these two groups strengthen the algorithm and decreases the error in prediction as some amino acids percentages were coupled with one group rather in the other.

To compare the efficiency of CAT3 with other cleavage tools that are mentioned in chapter 2, we run the three tools on the same group of

verified caspase -3 substrates “TEG” the contains 27 cleavage sites. Figure 4 shows the result of this comparison.

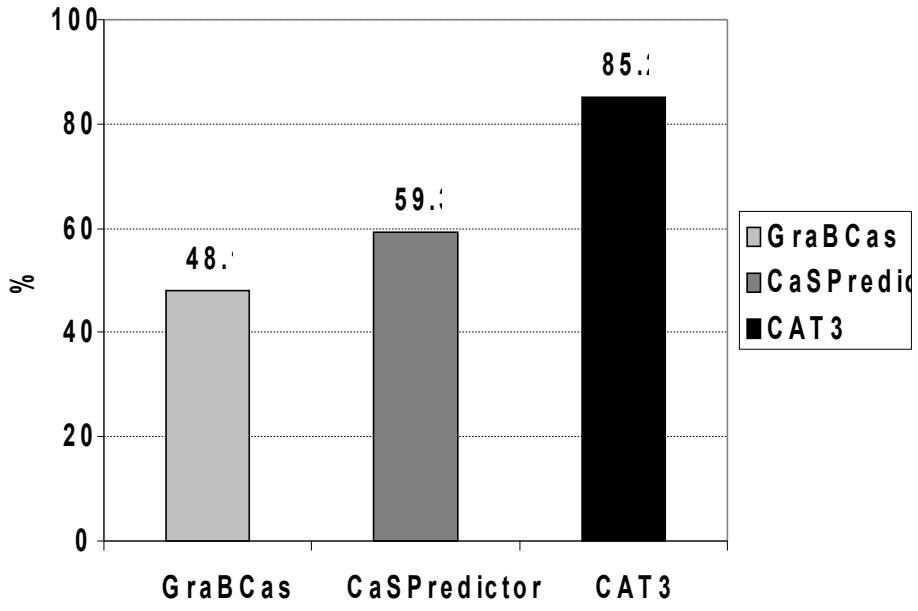

Figure 4 Comparison between CAT3 and other cleavage tools. “TEG” tested results from the three tools: GraBCas, CaSPredictor, and CAT3. GraBCas predicts (13/27), CaSPredictor predicts (16/27), and CAT3 predicts (23/27).

The undetected 15% (4/27) by CAT3 forms a false negative cleavage sites. These motifs together with new experimentally proven substrates

may form an advanced study for us to improve our algorithm code and its efficiency.

5.2 Conclusions

CAT3 is a powerful bioinformatics tool that predicts the cleavage site of the enzyme caspase-3 substrates. The significant efficiency of the code makes CAT3 one of the best tools in this field.

The usages of CAT3:

1. Prediction of cleavage sites: although there are some substrates that are experimentally proven to be cleaved by caspase -3, their exact cleavage site is unknown. These substrates are considered as excellent inputs for CAT3.
2. Prediction of new substrates: CAT3 could be used to scan all human proteins. Proteins with high scores "most of it" will form a group of new caspase -3 substrates "hypothetically".
3. Prediction of other caspases substrates: CAT3 could be used as a model for other caspases and proteases. Modifying the score matrices to the corresponding caspase by repeating the same analyses we did on caspase -3 substrates may help in predicting

other caspases substrates and their cleavage sites. We consider this work as a future work for this thesis.

References

Akita, K., T. Ohtsuki, Y. Nukada, T. Tanimoto, M. Namba, T. Okura, R. Takakura-Yamamoto, K. Torigoe, Y. Gu, M. S. Su, M. Fujii, M. Satoh-Itoh, K.

Yamamoto, K. Kohno, M. Ikeda, and M. Kurimoto. "Involvement of Caspase-1 and Caspase-3 in the Production and Processing of Mature Human Interleukin 18 in Monocytic Thp.1 Cells." *J Biol Chem* 272, no. 42 (1997): 26595-603.

Alnemri, E. S., D. J. Livingston, D. W. Nicholson, G. Salvesen, N. A. Thornberry, W. W. Wong, and J. Yuan. "Human Ice/Ced-3 Protease Nomenclature." *Cell* 87, no. 2 (1996): 171.

Atsumi, G., M. Murakami, K. Kojima, A. Hadano, M. Tajima, and I. Kudo. "Distinct Roles of Two Intracellular Phospholipase A2s in Fatty Acid Release in the Cell Death Pathway. Proteolytic Fragment of Type Iva Cytosolic Phospholipase A2alpha Inhibits Stimulus-Induced Arachidonate Release, Whereas That of Type Vi Ca2+-Independent Phospholipase A2 Augments Spontaneous Fatty Acid Release." *J Biol Chem* 275, no. 24 (2000): 18248-58.

Backes, C., J. Kuentzer, H. P. Lenhof, N. Comtesse, and E. Meese. "Grabcas: A Bioinformatics Tool for Score-Based Prediction of Caspase- and Granzyme B-Cleavage Sites in Protein Sequences." *Nucleic Acids Res* 33, no. Web Server issue (2005): W208-13.

Bae, S. S., D. K. Perry, Y. S. Oh, J. H. Choi, S. H. Galadari, T. Ghayur, S. H. Ryu, Y. A. Hannun, and P. G. Suh. "Proteolytic Cleavage of Phospholipase C-Gamma1 During Apoptosis in Molt-4 Cells." *Faseb J* 14, no. 9 (2000): 1083-92.

Basu, A., D. Lu, B. Sun, A. N. Moor, G. R. Akkaraju, and J. Huang. "Proteolytic Activation of Protein Kinase C-Epsilon by Caspase-Mediated Processing and Transduction of Antiapoptotic Signals." *J Biol Chem* 277, no. 44 (2002): 41850-6.

Bellows, D. S., B. N. Chau, P. Lee, Y. Lazebnik, W. H. Burns, and J. M. Hardwick. "Antiapoptotic Herpesvirus Bcl-2 Homologs Escape Caspase-Mediated Conversion to Proapoptotic Proteins." *J Virol* 74, no. 11 (2000): 5024-31.

Beyaert, R., V. J. Kidd, S. Cornelis, M. Van de Craen, G. Denecker, J. M. Lahti, R. Gururajan, P. Vandenabeele, and W. Fiers. "Cleavage of Pitslre Kinases by Ice/Casp-1 and Cpp32/Casp-3 During Apoptosis Induced by Tumor Necrosis Factor." *J Biol Chem* 272, no. 18 (1997): 11694-7.

Bhanumathy, C. D., S. K. Nakao, and S. K. Joseph. "Mechanism of Proteasomal Degradation of Inositol Trisphosphate Receptors in Cho-K1 Cells." *J Biol Chem* 281, no. 6 (2006): 3722-30.

Bischof, O., S. Galande, F. Farzaneh, T. Kohwi-Shigematsu, and J. Campisi. "Selective Cleavage of Blm, the Bloom Syndrome Protein, During Apoptotic Cell Death." *J Biol Chem* 276, no. 15 (2001): 12068-75.

Bleackley, R. C., and J. A. Heibein. "Enzymatic Control of Apoptosis." *Nat Prod Rep* 18, no. 4 (2001): 431-40.

Bratton, S. B., G. Walker, D. L. Roberts, K. Cain, and G. M. Cohen. "Caspase-3 Cleaves Apaf-1 into an Approximately 30 Kda Fragment That Associates with

an Inappropriately Oligomerized and Biologically Inactive Approximately 1.4 Mda Apoptosome Complex." *Cell Death Differ* 8, no. 4 (2001): 425-33.

Buendia, B., A. Santa-Maria, and J. C. Courvalin. "Caspase-Dependent Proteolysis of Integral and Peripheral Proteins of Nuclear Membranes and Nuclear Pore Complex Proteins During Apoptosis." *J Cell Sci* 112 (Pt 11) (1999): 1743-53.

Bushell, M., D. Poncet, W. E. Marissen, H. Flotow, R. E. Lloyd, M. J. Clemens, and S. J. Morley. "Cleavage of Polypeptide Chain Initiation Factor Eif4gi During Apoptosis in Lymphoma Cells: Characterisation of an Internal Fragment Generated by Caspase-3-Mediated Cleavage." *Cell Death Differ* 7, no. 7 (2000): 628-36.

Bushell, M., W. Wood, M. J. Clemens, and S. J. Morley. "Changes in Integrity and Association of Eukaryotic Protein Synthesis Initiation Factors During Apoptosis." *Eur J Biochem* 267, no. 4 (2000): 1083-91.

Byun, Y., F. Chen, R. Chang, M. Trivedi, K. J. Green, and V. L. Cryns. "Caspase Cleavage of Vimentin Disrupts Intermediate Filaments and Promotes Apoptosis." *Cell Death Differ* 8, no. 5 (2001): 443-50.

Casciola-Rosen, L., D. W. Nicholson, T. Chong, K. R. Rowan, N. A. Thornberry, D. K. Miller, and A. Rosen. "Apopain/Cpp32 Cleaves Proteins That Are Essential for Cellular Repair: A Fundamental Principle of Apoptotic Death." *J Exp Med* 183, no. 5 (1996): 1957-64.

Caulin, C., G. S. Salvesen, and R. G. Oshima. "Caspase Cleavage of Keratin 18 and Reorganization of Intermediate Filaments During Epithelial Cell Apoptosis." *J Cell Biol* 138, no. 6 (1997): 1379-94.

Chay, K. O., S. S. Park, and J. F. Mushinski. "Linkage of Caspase-Mediated Degradation of Paxillin to Apoptosis in Ba/F3 Murine Pro-B Lymphocytes." *J Biol Chem* 277, no. 17 (2002): 14521-9.

Chen, D., and Q. Zhou. "Caspase Cleavage of Bimel Triggers a Positive Feedback Amplification of Apoptotic Signaling." *Proc Natl Acad Sci U S A* 101, no. 5 (2004): 1235-40.

Chen, F., O. K. Arseven, and V. L. Cryns. "Proteolysis of the Mismatch Repair Protein Mlh1 by Caspase-3 Promotes DNA Damage-Induced Apoptosis." *J Biol Chem* 279, no. 26 (2004): 27542-8.

Chen, F., M. Kamradt, M. Mulcahy, Y. Byun, H. Xu, M. J. McKay, and V. L. Cryns. "Caspase Proteolysis of the Cohesin Component Rad21 Promotes Apoptosis." *J Biol Chem* 277, no. 19 (2002): 16775-81.

Chen, Y. R., R. Kori, B. John, and T. H. Tan. "Caspase-Mediated Cleavage of Actin-Binding and Sh3-Domain-Containing Proteins Cortactin, Hs1, and Hip-55 During Apoptosis." *Biochem Biophys Res Commun* 288, no. 4 (2001): 981-9.

Choi, E. K., N. F. Zaidi, J. S. Miller, A. C. Crowley, D. E. Merriam, C. Lilliehook, J. D. Buxbaum, and W. Wasco. "Calsenilin Is a Substrate for Caspase-3 That Preferentially Interacts with the Familial Alzheimer's Disease-Associated C-Terminal Fragment of Presenilin 2." *J Biol Chem* 276, no. 22 (2001): 19197-204.

Clem, R. J., T. T. Sheu, B. W. Richter, W. W. He, N. A. Thornberry, C. S. Duckett, and J. M. Hardwick. "C-Iap1 Is Cleaved by Caspases to Produce a Proapoptotic C-Terminal Fragment." *J Biol Chem* 276, no. 10 (2001): 7602-8.

Communal, C., M. Sumandea, P. de Tombe, J. Narula, R. J. Solaro, and R. J. Hajjar. "Functional Consequences of Caspase Activation in Cardiac Myocytes." *Proc Natl Acad Sci U S A* 99, no. 9 (2002): 6252-6.

Condorelli, F., P. Salomoni, S. Cotteret, V. Cesi, S. M. Srinivasula, E. S. Alnemri, and B. Calabretta. "Caspase Cleavage Enhances the Apoptosis-Inducing Effects of Bad." *Mol Cell Biol* 21, no. 9 (2001): 3025-36.

Cryns, V. L., Y. Byun, A. Rana, H. Mellor, K. D. Lustig, L. Ghanem, P. J. Parker, M. W. Kirschner, and J. Yuan. "Specific Proteolysis of the Kinase Protein Kinase C-Related Kinase 2 by Caspase-3 During Apoptosis. Identification by a Novel, Small Pool Expression Cloning Strategy." *J Biol Chem* 272, no. 47 (1997): 29449-53.

Datta, R., H. Kojima, K. Yoshida, and D. Kufe. "Caspase-3-Mediated Cleavage of Protein Kinase C Theta in Induction of Apoptosis." *J Biol Chem* 272, no. 33 (1997): 20317-20.

Duke, R. C., D. M. Ojcius, and J. D. Young. "Cell Suicide in Health and Disease." *Sci Am* 275, no. 6 (1996): 80-7.

Earnshaw, W. C., L. M. Martins, and S. H. Kaufmann. "Mammalian Caspases: Structure, Activation, Substrates, and Functions During Apoptosis." *Annu Rev Biochem* 68 (1999): 383-424.

Eckhart, L., C. Ballaun, A. Uthman, C. Kittel, M. Stichenwirth, M. Buchberger, H. Fischer, W. Sipos, and E. Tschachler. "Identification and Characterization of a Novel Mammalian Caspase with Proapoptotic Activity." *J Biol Chem* 280, no. 42 (2005): 35077-80.

Ellerby, L. M., R. L. Andrusiak, C. L. Wellington, A. S. Hackam, S. S. Propp, J. D. Wood, A. H. Sharp, R. L. Margolis, C. A. Ross, G. S. Salvesen, M. R. Hayden, and D. E. Bredesen. "Cleavage of Atrophin-1 at Caspase Site Aspartic Acid 109 Modulates Cytotoxicity." *J Biol Chem* 274, no. 13 (1999): 8730-6.

Eymin, B., O. Sordet, N. Droin, B. Munsch, M. Haugg, M. Van de Craen, P. Vandenabeele, and E. Solary. "Caspase-Induced Proteolysis of the Cyclin-Dependent Kinase Inhibitor P27kip1 Mediates Its Anti-Apoptotic Activity." *Oncogene* 18, no. 34 (1999): 4839-47.

Fischer, U., R. U. Janicke, and K. Schulze-Osthoff. "Many Cuts to Ruin: A Comprehensive Update of Caspase Substrates." *Cell Death Differ* 10, no. 1 (2003): 76-100.

Flygare, J., D. Hellgren, and A. Wennborg. "Caspase-3 Mediated Cleavage of Hsrad51 at an Unconventional Site." *Eur J Biochem* 267, no. 19 (2000): 5977-82.

Frame, M., K. F. Wan, R. Tate, P. Vandenabeele, and N. J. Pyne. "The Gamma Subunit of the Rod Photoreceptor Cgmp Phosphodiesterase Can Modulate the

Proteolysis of Two Cgmp Binding Cgmp-Specific Phosphodiesterases (Pde6 and Pde5) by Caspase-3." *Cell Signal* 13, no. 10 (2001): 735-41.

Franklin, C. C., C. M. Krejsa, R. H. Pierce, C. C. White, N. Fausto, and T. J. Kavanagh. "Caspase-3-Dependent Cleavage of the Glutamate-L-Cysteine Ligase Catalytic Subunit During Apoptotic Cell Death." *Am J Pathol* 160, no. 5 (2002): 1887-94.

Galvan, V., S. Chen, D. Lu, A. Logvinova, P. Goldsmith, E. H. Koo, and D. E. Bredesen. "Caspase Cleavage of Members of the Amyloid Precursor Family of Proteins." *J Neurochem* 82, no. 2 (2002): 283-94.

Garay-Malpartida, H. M., J. M. Occhiucci, J. Alves, and J. E. Belizario. "Caspredictor: A New Computer-Based Tool for Caspase Substrate Prediction." *Bioinformatics* 21 Suppl 1 (2005): i169-76.

Gastman, B. R., D. E. Johnson, T. L. Whiteside, and H. Rabinowich. "Caspase-Mediated Degradation of T-Cell Receptor Zeta-Chain." *Cancer Res* 59, no. 7 (1999): 1422-7.

Gentiletti, F., F. Mancini, M. D'Angelo, A. Sacchi, A. Pontecorvi, A. G. Jochemsen, and F. Moretti. "Mdmx Stability Is Regulated by P53-Induced Caspase Cleavage in Nih3t3 Mouse Fibroblasts." *Oncogene* 21, no. 6 (2002): 867-77.

Gervais, F. G., N. A. Thornberry, S. C. Ruffolo, D. W. Nicholson, and S. Roy. "Caspases Cleave Focal Adhesion Kinase During Apoptosis to Generate a Frnk-Like Polypeptide." *J Biol Chem* 273, no. 27 (1998): 17102-8.

Gervais, F. G., D. Xu, G. S. Robertson, J. P. Vaillancourt, Y. Zhu, J. Huang, A. LeBlanc, D. Smith, M. Rigby, M. S. Shearman, E. E. Clarke, H. Zheng, L. H. Van Der Ploeg, S. C. Ruffolo, N. A. Thornberry, S. Xanthoudakis, R. J. Zamboni, S. Roy, and D. W. Nicholson. "Involvement of Caspases in Proteolytic Cleavage of Alzheimer's Amyloid-Beta Precursor Protein and Amyloidogenic a Beta Peptide Formation." *Cell* 97, no. 3 (1999): 395-406.

Gervais, J. L., P. Seth, and H. Zhang. "Cleavage of Cdk Inhibitor P21(Cip1/Waf1) by Caspases Is an Early Event During DNA Damage-Induced Apoptosis." *J Biol Chem* 273, no. 30 (1998): 19207-12.

Gregorc, U., S. Ivanova, M. Thomas, V. Turk, L. Banks, and B. Turk. "Hdlg/Sap97, a Member of the Maguk Protein Family, Is a Novel Caspase Target During Cell-Cell Detachment in Apoptosis." *Biol Chem* 386, no. 7 (2005): 705-10.

Harvey, K. F., N. L. Harvey, J. M. Michael, G. Parasivam, N. Waterhouse, E. S. Alnemri, D. Watters, and S. Kumar. "Caspase-Mediated Cleavage of the Ubiquitin-Protein Ligase Nedd4 During Apoptosis." *J Biol Chem* 273, no. 22 (1998): 13524-30.

Haussermann, S., W. Kittstein, G. Rincke, F. J. Johannes, F. Marks, and M. Gschwendt. "Proteolytic Cleavage of Protein Kinase Cmu Upon Induction of Apoptosis in U937 Cells." *FEBS Lett* 462, no. 3 (1999): 442-6.

Henis-Korenblit, S., N. L. Strumpf, D. Goldstaub, and A. Kimchi. "A Novel Form of Dap5 Protein Accumulates in Apoptotic Cells as a Result of Caspase

Cleavage and Internal Ribosome Entry Site-Mediated Translation." *Mol Cell Biol* 20, no. 2 (2000): 496-506.

Hofmann, T. G., S. P. Hehner, W. Droge, and M. L. Schmitz. "Caspase-Dependent Cleavage and Inactivation of the Vav1 Proto-Oncogene Product During Apoptosis Prevents Il-2 Transcription." *Oncogene* 19, no. 9 (2000): 1153-63.

Houde, C., S. Roy, N. Leung, D. W. Nicholson, and N. Beauchemin. "The Cell Adhesion Molecule Ceacam1-L Is a Substrate of Caspase-3-Mediated Cleavage in Apoptotic Mouse Intestinal Cells." *J Biol Chem* 278, no. 19 (2003): 16929-35.

Itoh, M., H. Chiba, T. Noutomi, E. Takada, and J. Mizuguchi. "Cleavage of Bax-Alpha and Bcl-X(L) During Carboplatin-Mediated Apoptosis in Squamous Cell Carcinoma Cell Line." *Oral Oncol* 36, no. 3 (2000): 277-85.

Kahns, S., S. Lykkebo, L. D. Jakobsen, M. S. Nielsen, and P. H. Jensen. "Caspase-Mediated Parkin Cleavage in Apoptotic Cell Death." *J Biol Chem* 277, no. 18 (2002): 15303-8.

Katsuda, K., M. Kataoka, F. Uno, T. Murakami, T. Kondo, J. A. Roth, N. Tanaka, and T. Fujiwara. "Activation of Caspase-3 and Cleavage of Rb Are Associated with P16-Mediated Apoptosis in Human Non-Small Cell Lung Cancer Cells." *Oncogene* 21, no. 13 (2002): 2108-13.

Keller, S. H., and S. K. Nigam. "Biochemical Processing of E-Cadherin under Cellular Stress." *Biochem Biophys Res Commun* 307, no. 2 (2003): 215-23.

Kerr, J. F., A. H. Wyllie, and A. R. Currie. "Apoptosis: A Basic Biological Phenomenon with Wide-Ranging Implications in Tissue Kinetics." *Br J Cancer* 26, no. 4 (1972): 239-57.

Kim, K. W., H. H. Chung, C. W. Chung, I. K. Kim, M. Miura, S. Wang, H. Zhu, K. D. Moon, G. B. Rha, J. H. Park, D. G. Jo, H. N. Woo, Y. H. Song, B. J. Kim, J. Yuan, and Y. K. Jung. "Inactivation of Farnesyltransferase and Geranylgeranyltransferase I by Caspase-3: Cleavage of the Common Alpha Subunit During Apoptosis." *Oncogene* 20, no. 3 (2001): 358-66.

Kim, M., K. Murphy, F. Liu, S. E. Parker, M. L. Dowling, W. Baff, and G. D. Kao. "Caspase-Mediated Specific Cleavage of Bubr1 Is a Determinant of Mitotic Progression." *Mol Cell Biol* 25, no. 21 (2005): 9232-48.

Kipp, M., B. L. Schwab, M. Przybylski, P. Nicotera, and F. O. Fackelmayer. "Apoptotic Cleavage of Scaffold Attachment Factor a (Saf-a) by Caspase-3 Occurs at a Noncanonical Cleavage Site." *J Biol Chem* 275, no. 7 (2000): 5031-6.

Kook, S., S. R. Shim, S. J. Choi, J. Ahnn, J. I. Kim, S. H. Eom, Y. K. Jung, S. G. Paik, and W. K. Song. "Caspase-Mediated Cleavage of P130cas in Etoposide-Induced Apoptotic Rat-1 Cells." *Mol Biol Cell* 11, no. 3 (2000): 929-39.

Kothakota, S., T. Azuma, C. Reinhard, A. Klippel, J. Tang, K. Chu, T. J. McGarry, M. W. Kirschner, K. Koths, D. J. Kwiatkowski, and L. T. Williams. "Caspase-3-Generated Fragment of Gelsolin: Effector of Morphological Change in Apoptosis." *Science* 278, no. 5336 (1997): 294-8.

Kovacsovics, M., F. Martinon, O. Micheau, J. L. Bodmer, K. Hofmann, and J. Tschopp. "Overexpression of Helicard, a Card-Containing Helicase Cleaved During Apoptosis, Accelerates DNA Degradation." *Curr Biol* 12, no. 10 (2002): 838-43.

Lamkanfi, M., N. Festjens, W. Declercq, T. Vanden Berghe, and P. Vandenabeele. "Caspases in Cell Survival, Proliferation and Differentiation." *Cell Death Differ* 14, no. 1 (2007): 44-55.

Landais, I., H. Lee, and H. Lu. "Coupling Caspase Cleavage and Ubiquitin-Proteasome-Dependent Degradation of Ssrp1 During Apoptosis." *Cell Death Differ* 13, no. 11 (2006): 1866-78.

Lane, J. D., J. Lucocq, J. Pryde, F. A. Barr, P. G. Woodman, V. J. Allan, and M. Lowe. "Caspase-Mediated Cleavage of the Stacking Protein Grasp65 Is Required for Golgi Fragmentation During Apoptosis." *J Cell Biol* 156, no. 3 (2002): 495-509.

Lane, J. D., M. A. Vergnolle, P. G. Woodman, and V. J. Allan. "Apoptotic Cleavage of Cytoplasmic Dynein Intermediate Chain and P150(Glued) Stops Dynein-Dependent Membrane Motility." *J Cell Biol* 153, no. 7 (2001): 1415-26.

Law, S. F., G. M. O'Neill, S. J. Fashena, M. B. Einarson, and E. A. Golemis. "The Docking Protein Hef1 Is an Apoptotic Mediator at Focal Adhesion Sites." *Mol Cell Biol* 20, no. 14 (2000): 5184-95.

Lazebnik, Y. A., S. H. Kaufmann, S. Desnoyers, G. G. Poirier, and W. C. Earnshaw. "Cleavage of Poly(Adp-Ribose) Polymerase by a Proteinase with Properties Like Ice." *Nature* 371, no. 6495 (1994): 346-7.

Lee, S. B., D. Rodriguez, J. R. Rodriguez, and M. Esteban. "The Apoptosis Pathway Triggered by the Interferon-Induced Protein Kinase Pkr Requires the Third Basic Domain, Initiates Upstream of Bcl-2, and Involves Ice-Like Proteases." *Virology* 231, no. 1 (1997): 81-8.

Leo, E., Q. L. Deveraux, C. Buchholtz, K. Welsh, S. Matsuzawa, H. R. Stennicke, G. S. Salvesen, and J. C. Reed. "Traf1 Is a Substrate of Caspases Activated During Tumor Necrosis Factor Receptor-Alpha-Induced Apoptosis." *J Biol Chem* 276, no. 11 (2001): 8087-93.

Liping Wei, Jeffrey T. Chang and Russ B. Altman. "Statistical Analysis of Protein Structures Using Environmental Features for Multiple Purposes
" In *Computational Methods in Molecular Biology*, edited by Bernardi G., 205-25. Paris: ELSEVIER, 1998.

Liu, F., M. Dowling, X. J. Yang, and G. D. Kao. "Caspase-Mediated Specific Cleavage of Human Histone Deacetylase 4." *J Biol Chem* 279, no. 33 (2004): 34537-46.

Liu, W., and S. Linn. "Proteolysis of the Human DNA Polymerase Epsilon Catalytic Subunit by Caspase-3 and Calpain Specifically During Apoptosis." *Nucleic Acids Res* 28, no. 21 (2000): 4180-8.

Lo, S. S., S. H. Lo, and S. H. Lo. "Cleavage of Cten by Caspase-3 During Apoptosis." *Oncogene* 24, no. 26 (2005): 4311-4.

Los, M., C. Stroh, R. U. Janicke, I. H. Engels, and K. Schulze-Osthoff. "Caspases: More Than Just Killers?" *Trends Immunol* 22, no. 1 (2001): 31-4.

Lowe, M., J. D. Lane, P. G. Woodman, and V. J. Allan. "Caspase-Mediated Cleavage of Syntaxin 5 and Giantin Accompanies Inhibition of Secretory Traffic During Apoptosis." *J Cell Sci* 117, no. Pt 7 (2004): 1139-50.

Luciano, F., J. E. Ricci, and P. Auberger. "Cleavage of Fyn and Lyn in Their N-Terminal Unique Regions During Induction of Apoptosis: A New Mechanism for Src Kinase Regulation." *Oncogene* 20, no. 36 (2001): 4935-41.

Luschen, S., S. Ussat, M. Kronke, and S. Adam-Klages. "Cleavage of Human Cytosolic Phospholipase A2 by Caspase-1 (Ice) and Caspase-8 (Flice)." *Biochem Biophys Res Commun* 253, no. 1 (1998): 92-8.

Mazarakis, N. D., A. D. Edwards, and H. Mehmet. "Apoptosis in Neural Development and Disease." *Arch Dis Child Fetal Neonatal Ed* 77, no. 3 (1997): F165-70.

McGinnis, K. M., M. M. Whitton, M. E. Gnegy, and K. K. Wang. "Calcium/Calmodulin-Dependent Protein Kinase Iv Is Cleaved by Caspase-3 and Calpain in Sh-Sy5y Human Neuroblastoma Cells Undergoing Apoptosis." *J Biol Chem* 273, no. 32 (1998): 19993-20000.

Mehlen, P., S. Rabizadeh, S. J. Snipas, N. Assa-Munt, G. S. Salvesen, and D. E. Bredesen. "The Dcc Gene Product Induces Apoptosis by a Mechanism Requiring Receptor Proteolysis." *Nature* 395, no. 6704 (1998): 801-4.

Mejillano, M., M. Yamamoto, A. L. Rozelle, H. Q. Sun, X. Wang, and H. L. Yin. "Regulation of Apoptosis by Phosphatidylinositol 4,5-Bisphosphate Inhibition of Caspases, and Caspase Inactivation of Phosphatidylinositol Phosphate 5-Kinases." *J Biol Chem* 276, no. 3 (2001): 1865-72.

Morley, S. J., M. J. Coldwell, and M. J. Clemens. "Initiation Factor Modifications in the Preapoptotic Phase." *Cell Death Differ* 12, no. 6 (2005): 571-84.

Mukerjee, N., K. M. McGinnis, Y. H. Park, M. E. Gnegy, and K. K. Wang. "Caspase-Mediated Proteolytic Activation of Calcineurin in Thapsigargin-Mediated Apoptosis in Sh-Sy5y Neuroblastoma Cells." *Arch Biochem Biophys* 379, no. 2 (2000): 337-43.

Nachmias, B., Y. Ashhab, V. Bucholtz, O. Drize, L. Kadouri, M. Lotem, T. Peretz, O. Mandelboim, and D. Ben-Yehuda. "Caspase-Mediated Cleavage Converts Livin from an Antiapoptotic to a Proapoptotic Factor: Implications for Drug-Resistant Melanoma." *Cancer Res* 63, no. 19 (2003): 6340-9.

Nyormoi, O., Z. Wang, D. Doan, M. Ruiz, D. McConkey, and M. Bar-Eli. "Transcription Factor Ap-2alpha Is Preferentially Cleaved by Caspase 6 and Degraded by Proteasome During Tumor Necrosis Factor Alpha-Induced Apoptosis in Breast Cancer Cells." *Mol Cell Biol* 21, no. 15 (2001): 4856-67.

Paszty, K., A. K. Verma, R. Padanyi, A. G. Filoteo, J. T. Penniston, and A. Enyedi. "Plasma Membrane Ca2+Atpase Isoform 4b Is Cleaved and Activated by Caspase-3 During the Early Phase of Apoptosis." *J Biol Chem* 277, no. 9 (2002): 6822-9.

Persaud, S. D., V. Hoang, J. Huang, and A. Basu. "Involvement of Proteolytic Activation of Pkcdelta in Cisplatin-Induced Apoptosis in Human Small Cell Lung Cancer H69 Cells." *Int J Oncol* 27, no. 1 (2005): 149-54.

Pochampally, R., B. Fodera, L. Chen, W. Shao, E. A. Levine, and J. Chen. "A 60 Kd Mdm2 Isoform Is Produced by Caspase Cleavage in Non-Apoptotic Tumor Cells." *Oncogene* 17, no. 20 (1998): 2629-36.

Qi, H., P. Juo, J. Masuda-Robens, M. J. Caloca, H. Zhou, N. Stone, M. G. Kazanietz, and M. M. Chou. "Caspase-Mediated Cleavage of the Tiam1 Guanine Nucleotide Exchange Factor During Apoptosis." *Cell Growth Differ* 12, no. 12 (2001): 603-11.

Rao, L., D. Perez, and E. White. "Lamin Proteolysis Facilitates Nuclear Events During Apoptosis." *J Cell Biol* 135, no. 6 Pt 1 (1996): 1441-55.

Rheaume, E., L. Y. Cohen, F. Uhlmann, C. Lazure, A. Alam, J. Hurwitz, R. P. Sekaly, and F. Denis. "The Large Subunit of Replication Factor C Is a Substrate for Caspase-3 in Vitro and Is Cleaved by a Caspase-3-Like Protease During Fas-Mediated Apoptosis." *Embo J* 16, no. 21 (1997): 6346-54.

Ricci, J. E., L. Maulon, F. Luciano, S. Guerin, A. Livolsi, B. Mari, J. P. Breittmayer, J. F. Peyron, and P. Auberger. "Cleavage and Relocation of the Tyrosine Kinase P59fyn During Fas-Mediated Apoptosis in T Lymphocytes." *Oncogene* 18, no. 27 (1999): 3963-9.

Rissman, R. A., W. W. Poon, M. Blurton-Jones, S. Oddo, R. Torp, M. P. Vitek, F. M. LaFerla, T. T. Rohn, and C. W. Cotman. "Caspase-Cleavage of Tau Is an Early Event in Alzheimer Disease Tangle Pathology." *J Clin Invest* 114, no. 1 (2004): 121-30.

Samejima, K., P. A. Svingen, G. S. Basi, T. Kottke, P. W. Mesner, Jr., L. Stewart, F. Durrieu, G. G. Poirier, E. S. Alnemri, J. J. Champoux, S. H. Kaufmann, and W. C. Earnshaw. "Caspase-Mediated Cleavage of DNA Topoisomerase I at Unconventional Sites During Apoptosis." *J Biol Chem* 274, no. 7 (1999): 4335-40.

Satoh, S., M. Hijikata, H. Handa, and K. Shimotohno. "Caspase-Mediated Cleavage of Eukaryotic Translation Initiation Factor Subunit 2alpha." *Biochem J* 342 (Pt 1) (1999): 65-70.

Schaecher, K., J. M. Goust, and N. L. Banik. "The Effects of Calpain Inhibition on Ikb Alpha Degradation after Activation of Pbmcs: Identification of the Calpain Cleavage Sites." *Neurochem Res* 29, no. 7 (2004): 1443-51.

Sebbagh, M., C. Renvoize, J. Hamelin, N. Riche, J. Bertoglio, and J. Breard. "Caspase-3-Mediated Cleavage of Rock I Induces Mlc Phosphorylation and Apoptotic Membrane Blebbing." *Nat Cell Biol* 3, no. 4 (2001): 346-52.

Sgorbissa, A., R. Benetti, S. Marzinotto, C. Schneider, and C. Brancolini. "Caspase-3 and Caspase-7 but Not Caspase-6 Cleave Gas2 in Vitro: Implications for Microfilament Reorganization During Apoptosis." *J Cell Sci* 112 (Pt 23) (1999): 4475-82.

Smith, L., L. Chen, M. E. Reyland, T. A. DeVries, R. V. Talanian, S. Omura, and J. B. Smith. "Activation of Atypical Protein Kinase C Zeta by Caspase Processing and Degradation by the Ubiquitin-Proteasome System." *J Biol Chem* 275, no. 51 (2000): 40620-7.

Song, Q., S. P. Lees-Miller, S. Kumar, Z. Zhang, D. W. Chan, G. C. Smith, S. P. Jackson, E. S. Alnemri, G. Litwack, K. K. Khanna, and M. F. Lavin. "DNA-Dependent Protein Kinase Catalytic Subunit: A Target for an Ice-Like Protease in Apoptosis." *Embo J* 15, no. 13 (1996): 3238-46.

Song, Q., T. Wei, S. Lees-Miller, E. Alnemri, D. Watters, and M. F. Lavin. "Resistance of Actin to Cleavage During Apoptosis." *Proc Natl Acad Sci U S A* 94, no. 1 (1997): 157-62.

Steinhusen, U., V. Badock, A. Bauer, J. Behrens, B. Wittman-Liebold, B. Dorken, and K. Bommert. "Apoptosis-Induced Cleavage of Beta-Catenin by Caspase-3 Results in Proteolytic Fragments with Reduced Transactivation Potential." *J Biol Chem* 275, no. 21 (2000): 16345-53.

Stennicke, H. R., M. Renatus, M. Meldal, and G. S. Salvesen. "Internally Quenched Fluorescent Peptide Substrates Disclose the Subsite Preferences of Human Caspases 1, 3, 6, 7 and 8." *Biochem J* 350 Pt 2 (2000): 563-8.

Taimen, P., and M. Kallajoki. "Numa and Nuclear Lamins Behave Differently in Fas-Mediated Apoptosis." *J Cell Sci* 116, no. Pt 3 (2003): 571-83.

Takahashi, M., H. Mukai, M. Toshimori, M. Miyamoto, and Y. Ono. "Proteolytic Activation of Pkn by Caspase-3 or Related Protease During Apoptosis." *Proc Natl Acad Sci U S A* 95, no. 20 (1998): 11566-71.

Takeda, Y., P. Caudell, G. Grady, G. Wang, A. Suwa, G. C. Sharp, W. S. Dynan, and J. A. Hardin. "Human Rna Helicase a Is a Lupus Autoantigen That Is Cleaved During Apoptosis." *J Immunol* 163, no. 11 (1999): 6269-74.

Tee, A. R., and C. G. Proud. "Caspase Cleavage of Initiation Factor 4e-Binding Protein 1 Yields a Dominant Inhibitor of Cap-Dependent Translation and Reveals a Novel Regulatory Motif." *Mol Cell Biol* 22, no. 6 (2002): 1674-83.

Thornberry, N. A., T. A. Rano, E. P. Peterson, D. M. Rasper, T. Timkey, M. Garcia-Calvo, V. M. Houtzager, P. A. Nordstrom, S. Roy, J. P. Vaillancourt, K. T. Chapman, and D. W. Nicholson. "A Combinatorial Approach Defines Specificities of Members of the Caspase Family and Granzyme B. Functional Relationships Established for Key Mediators of Apoptosis." *J Biol Chem* 272, no. 29 (1997): 17907-11.

Tikhomirov, O., and G. Carpenter. "Caspase-Dependent Cleavage of Erbb-2 by Geldanamycin and Staurosporin." *J Biol Chem* 276, no. 36 (2001): 33675-80.

Torres, J., J. Rodriguez, M. P. Myers, M. Valiente, J. D. Graves, N. K. Tonks, and R. Pulido. "Phosphorylation-Regulated Cleavage of the Tumor Suppressor Pten by Caspase-3: Implications for the Control of Protein Stability and Pten-Protein Interactions." *J Biol Chem* 278, no. 33 (2003): 30652-60.

Tozser, J., P. Bagossi, G. Zahuczky, S. I. Specht, E. Majerova, and T. D. Copeland. "Effect of Caspase Cleavage-Site Phosphorylation on Proteolysis." *Biochem J* 372, no. Pt 1 (2003): 137-43.

Tu, S., and R. A. Cerione. "Cdc42 Is a Substrate for Caspases and Influences Fas-Induced Apoptosis." *J Biol Chem* 276, no. 22 (2001): 19656-63.

Tulasne, D., J. Deheuninck, F. C. Lourenco, F. Lamballe, Z. Ji, C. Leroy, E. Puchois, A. Moumen, F. Maina, P. Mehlen, and V. Fafeur. "Proapoptotic Function of the Met Tyrosine Kinase Receptor through Caspase Cleavage." *Mol Cell Biol* 24, no. 23 (2004): 10328-39.

Utz, P. J., M. Hottelet, T. M. Le, S. J. Kim, M. E. Geiger, W. J. van Venrooij, and P. Anderson. "The 72-Kda Component of Signal Recognition Particle Is Cleaved During Apoptosis." *J Biol Chem* 273, no. 52 (1998): 35362-70.

Vartanyan, A. A., E. V. Stepanova, A. Y. Baryshnikov, and M. R. Lichinitser. "Involvement of Apoptosis in the Formation of Vasculogenic Mimicry in the Malignant Neoplasms." *Dokl Biol Sci* 402 (2005): 217-20.

Vito, P., T. Ghayur, and L. D'Adamio. "Generation of Anti-Apoptotic Presenilin-2 Polypeptides by Alternative Transcription, Proteolysis, and Caspase-3 Cleavage." *J Biol Chem* 272, no. 45 (1997): 28315-20.

Walter, B. N., Z. Huang, R. Jakobi, P. T. Tuazon, E. S. Alnemri, G. Litwack, and J. A. Traugh. "Cleavage and Activation of P21-Activated Protein Kinase Gamma-Pak by Cpp32 (Caspase 3). Effects of Autophosphorylation on Activity." *J Biol Chem* 273, no. 44 (1998): 28733-9.

Wang, K. K., R. Posmantur, R. Nath, K. McGinnis, M. Whitton, R. V. Talanian, S. B. Glantz, and J. S. Morrow. "Simultaneous Degradation of Alphaii- and Betaii-Spectrin by Caspase 3 (Cpp32) in Apoptotic Cells." *J Biol Chem* 273, no. 35 (1998): 22490-7.

Webb, S. J., D. Nicholson, V. J. Bubb, and A. H. Wyllie. "Caspase-Mediated Cleavage of Apc Results in an Amino-Terminal Fragment with an Intact Armadillo Repeat Domain." *Faseb J* 13, no. 2 (1999): 339-46.

Wellington, C. L., L. M. Ellerby, A. S. Hackam, R. L. Margolis, M. A. Trifiro, R. Singaraja, K. McCutcheon, G. S. Salvesen, S. S. Propp, M. Bromm, K. J. Rowland, T. Zhang, D. Rasper, S. Roy, N. Thornberry, L. Pinsky, A. Kakizuka, C. A. Ross, D. W. Nicholson, D. E. Bredesen, and M. R. Hayden. "Caspase Cleavage of Gene Products Associated with Triplet Expansion Disorders Generates Truncated Fragments Containing the Polyglutamine Tract." *J Biol Chem* 273, no. 15 (1998): 9158-67.

Weng, C., Y. Li, D. Xu, Y. Shi, and H. Tang. "Specific Cleavage of Mcl-1 by Caspase-3 in Tumor Necrosis Factor-Related Apoptosis-Inducing Ligand (Trail)-Induced Apoptosis in Jurkat Leukemia T Cells." *J Biol Chem* 280, no. 11 (2005): 10491-500.

Werner, M. E., F. Chen, J. V. Moyano, F. Yehiely, J. C. Jones, and V. L. Cryns. "Caspase Proteolysis of the Integrin Beta4 Subunit Disrupts Hemidesmosome

Assembly, Promotes Apoptosis, and Inhibits Cell Migration." *J Biol Chem* 282, no. 8 (2007): 5560-9.

Wu, Y. H., S. F. Shih, and J. Y. Lin. "Ricin Triggers Apoptotic Morphological Changes through Caspase-3 Cleavage of Bat3." *J Biol Chem* 279, no. 18 (2004): 19264-75.

Wu, Y. M., C. L. Huang, H. J. Kung, and C. Y. Huang. "Proteolytic Activation of Etk/Bmx Tyrosine Kinase by Caspases." *J Biol Chem* 276, no. 21 (2001): 17672-8.

Xu, J., D. Liu, and Z. Songyang. "The Role of Asp-462 in Regulating Akt Activity." *J Biol Chem* 277, no. 38 (2002): 35561-6.

Yankee, T. M., K. E. Draves, M. K. Ewings, E. A. Clark, and J. D. Graves. "Cd95/Fas Induces Cleavage of the Grpl/Gads Adaptor and Desensitization of Antigen Receptor Signaling." *Proc Natl Acad Sci U S A* 98, no. 12 (2001): 6789-93.

Yim, H., Y. H. Jin, B. D. Park, H. J. Choi, and S. K. Lee. "Caspase-3-Mediated Cleavage of Cdc6 Induces Nuclear Localization of P49-Truncated Cdc6 and Apoptosis." *Mol Biol Cell* 14, no. 10 (2003): 4250-9.

Yoshimori, A., R. Takasawa, and S. Tanuma. "A Novel Method for Evaluation and Screening of Caspase Inhibitory Peptides by the Amino Acid Positional Fitness Score." *BMC Pharmacol* 4 (2004): 7.

Zhan, Q., S. Jin, B. Ng, J. Plisket, S. Shangary, A. Rathi, K. D. Brown, and R. Baskaran. "Caspase-3 Mediated Cleavage of Brca1 During Uv-Induced Apoptosis." *Oncogene* 21, no. 34 (2002): 5335-45.

Zhang, B., Y. Zhang, and E. Shacter. "Caspase 3-Mediated Inactivation of Rac Gtpases Promotes Drug-Induced Apoptosis in Human Lymphoma Cells." *Mol Cell Biol* 23, no. 16 (2003): 5716-25.

Zhang, Y., D. M. Center, D. M. Wu, W. W. Cruikshank, J. Yuan, D. W. Andrews, and H. Kornfeld. "Processing and Activation of Pro-Interleukin-16 by Caspase-3." *J Biol Chem* 273, no. 2 (1998): 1144-9.

Zihni, C., C. Mitsopoulos, I. A. Tavares, B. Baum, A. J. Ridley, and J. D. Morris. "Prostate-Derived Sterile 20-Like Kinase 1-{Alpha} Induces Apoptosis: Jnk- and Caspase-Dependent Nuclear Localization Is a Requirement for Membrane Blebbing." *J Biol Chem* 282, no. 9 (2007): 6484-93.

Zihni, C., C. Mitsopoulos, I. A. Tavares, A. J. Ridley, and J. D. Morris. "Prostate-Derived Sterile 20-Like Kinase 2 (Psk2) Regulates Apoptotic Morphology Via C-Jun N-Terminal Kinase and Rho Kinase-1." *J Biol Chem* 281, no. 11 (2006): 7317-23.

Zou, H., R. Yang, J. Hao, J. Wang, C. Sun, S. W. Fesik, J. C. Wu, K. J. Tomaselli, and R. C. Armstrong. "Regulation of the Apaf-1/Caspase-9 Apoptosome by Caspase-3 and Xiap." *J Biol Chem* 278, no. 10 (2003): 8091-8.

Appendices

Database

Matrix Establishing Group “MEG”

#	Name	S.W	Motif 1		Motif 2		Reference
1	Bcl-2	P10415	dagd	34			[Bellows et al. 2000]
2	Calpastatin	P20810	daid	233			[Kato et al. 2000)]
3	PDE4A	P27815	damd	72			[Huston et al. 2000]
4	Alpha-adducin	P35611	ddsd	633			[Water et al. 2000]
5	HPK1 (MEKKK 1)	Q92918	ddvd	385			[Chen et al. 1999]
6	Androgen receptor	P10275	dedd	155			[Ellerby et al. 1999]
7	PARG	Q86W56	deid	256			[Affar et al. 2001)]
8	PMCA4	P23634	deid	1080			[Paszty et al. 2002]
9	Mst 2	Q13188	deld	322			[Graves et al. 1998]
10	D4-GDI	Q9TU03	deld	18			[Na et al. 1996]
11	MST1	Q13043	demd	326			[Graves et al. 1998]
12	SPAK	Q9UEW8	demd	392			[Johnston et al. 2000]
13	DCAMKL1	O15075	dend	369			[Kruidering et al. 2001]
14	SREBP-2	Q12772	depd	468			[Wang et al. 1996]
15	PP2A subunit A	P30153	deqd	317			[Santoro et al. 1998]
16	RAP1	Q15276	desd	438			[Cosulich et al.1997]
17	MASK (MST4)	Q9P289	desd	305			[Dan et al. 2002]
#	Name	S.W	Motif 1		Motif 2		Reference
18	DNA Frag. Factor	O00273	detd	117	davd	224	[Inohara et al. 1998]
19	Alpha-II spectrin	Q13813	detd	1185	dsld	1478	[Wang et al. 1998]
20	ROCK1	Q13464	detd	1113			[Sebbagh et al. 2001]
21	eIF4B	P23588	detd	45			[Bushell et al. 2000]
22	eIF4G2 (DAP5)	P78344	detd	792			[Henis-Korenblit et al. 2000]
23	TIAM 1	Q13009	detd	993			[Qi et al. 2001]
24	PARP-1	P09874	devd	214			[Lazebnik et al. 1994]
25	PRKD	P78527	devd	2713			[Song,et al. 1996]
26	RFC140	P35251	devd	723			[Rheaume et al. 1997]
27	PKC-theta	Q04759	devd	354			[Datta et al. 1997]
28	Beta-II spectrin	Q01082	devd	1457			[Wang et al. 1998]
29	IP3R	Q14643	devd	1900			[Bhanumathy et al. 2006]
30	TNTc	P45379	dfdd	98			[Communal et al. 2002]
31	Calcineurin	Q08209	dgfd	385			[Mukerjee et al. 2000]
32	PDE6A	P16499	dfvd	166			[Frame et al. 2001]
33	U1 snRNP 70 kDa	P08621	dgpd	341			[Casciola-Rosen et al. 1996]
34	LYN	P07948	dgvd	17			[Luciano et al. 2001]
35	P21 waf1/ cip1	P38936	dhvd	112			[Gervais et al. 1998]
36	Livin	Q96CA5	dhvd	52			[Nachmias etal.2003]
37	GRP2 (GrpL)	O75791	dind	241			[Yankee et al. 2001]
38	PIP5K1A	Q99754	dipd	279			[Mejillano et al. 2001]
39	PRK2 (PKL2)	Q16513	ditd	117	devd	700	[Cryns et al. 1997]
#	Name	S.W	Motif 1		Motif 2		Reference
40	LAP2A	P42166	dkdd	108			[Buendia et al. 1999]

41	MLH1	P40692	dktd	418			[Chen et al. 2004]
42	eIF3(p35)	O75822	dlad	242			[Morley et al.2005]
43	PDE10A2	Q9ULW9	dlfd	315			[Frame et al. 2001]
44	eIF4G1	Q04637	drld	1176			[Bushel et al. 2000]
45	BRCA-1	P38398	dlld	1155			[Zhan et al. 2002]
46	HEF1	Q14511	dlvd	363	ddyd	630	[Law et al. 2000]
47	Vav1	P15498	dlyd	161			[Hofmann et al.2000]
48	PKC-delta I	Q05655	dmqd	330			[Persaud et al. 2005]
49	IL-18	Q14116	dmtd	71			[Akita et al. 1997]
50	APC	P25054	dnid	777			[Webb et al. 999]
51	Helicard	Q8R5F7	dntd	208			[Kovacsovics et al. 2002]
52	DNA polymerase ε	Q07864	dmed	1214			[Liu and Linn 2000]
53	Caspase 9	P55211	dqld	330			[Zou et al. 2003]
54	pp125FAK	Q05397	dqtd	772			[Gervais et al. 1998]
55	Gelsolin	P06396	dqtd	403			[Kothakota et al.1997]
56	IkB-a	P25963	drhd	31			[Schaecher et al. 004]
57	AP2A	P05549	drhd	19			[Nyormoi et al. 2001]
58	CDC42	P60953	dlrd	121			[Tu and Cerione 2001]
59	CD-IC	O14576	dsgd	116			[Lane et al. 2001]
60	DRPLA	P54259	dsld	109			[Ellerby et al.1999]
61	NuMA	Q14980	dsld	1727			[Taimen and Kallajoki 2003]
#	**Name**	**S.W**	**Motif 1**		**Motif 2**		**Reference**
62	RAD21	O60216	dspd	279			[Chen et al. 2002]
63	Calsenilin	Q9Y2W7	dssd	64			[Choi et al. 2001]
64	Huntingtin	P42858	dsvd	513			[Wellington et al. 1998]
65	Vimentin	P08670	dsvd	84			[Byun et al. 2001]
66	DHX9	Q08211	dtpd	96			[Takeda et al. 1999]
67	E-cadherin	P12830	dtrd	750			[Keller and Nigam 2003]
68	RasGAP	P20936	dtvd	455	degd	157	[Yang and Widmann 2001]
69	TCR-zeta chain	P20963	dtyd	154			[Gastman et l.1999]
70	Rad51-A	Q06609	dvld	187			[Flygare et al.2000]
71	P130 cas	P56945	dvpd	318	dspd	650	[Kook et al. 2000]
72	MDM2	Q00987	dvpd	361			[Pochampally et al. 1998]
73	MDM4 (MDMX)	O15151	dvpd	361			Gentiletti et al.2002]
74	Histone deacetylase4	P56524	dvtd	289			[Liu et al. 2004]
75	iPLA2	O60733	dvtd	183			[Atsumi et al. 000]
76	Desmoglein-3	P32926	dyad	781			[Weiske et al. 2001]
77	ATM	Q13315	dypd	863			[Smith et al.1999]
78	PLCG1	P19174	aepd	770			[Bae et al. 2000]
79	Mst3	Q9Y6E0	aetd	325			[Huang et al. 2002]
80	eIF-2-alpha	P05198	aevd	300			[Lee et al. 1997]
81	POM121	Q9Y2N3	aled	530			[Satoh et al. 1999]
82	PKN	Q16512	dfld	454			[Takahashi et al. 998]
83	GCLC	P48506	avvd	499			[Franklin et al. 2002]
#	**Name**	**S.W**	**Motif 1**		**Motif 2**		**Reference**

84	PKC-mu	Q15139	cqnd	378			[Haussermann et al. 1999]
85	AKT (PKB)	P31749	ecvd	462			[Xu et al. 2002]
86	Topoisomerase I	P11387	ddad	146	eeed	170	[Samejima et al. 1999]
87	FYN	P06241	eerd	18			[Ricci et al. 1999]
88	PKC zeta	Q05513	eetd	210	dgmd	239	[Smith et al. 2000]
89	Hip-55	Q9UJU6	ehid	361			[Chen et al. 2001]
90	Beta-actin	P60709	elpd	244			[Song et al. 1997]
91	c-IAP1 (BIRC2)	Q13490	enad	372			[Clem et al. 2001]
92	BAX-alpha	Q07812	fiqd	33			[Itoh et al. 2000]
93	TAU	P10636	gssd	420			[Rissman et al. 2004]
94	TRAF 1	Q13077	levd	163			[Leo et al. 2001]
95	Parkin	O60260	lhtd	126			[Kahns et al. 2002]
96	DCC	P43146	lsvd	1290			[Mehlen et al. 1998]
97	BAD	Q92934	pagd	29			[Condorell et al. 2001]
98	CaMK (IV)	Q16566	papd	176			[McGinnis et al. 1998]
99	PTEN*	P60484	qeid	301	dvsd	371	[Torres et al.2003]
100	SAF-A	Q00839	sald	100			[Kipp et al. 2000]
101	BimEL	O43521	secd	13			[Chen and Zhou 2004]
102	SRP72	O76094	seld	613			[Utz et al. 1998]
103	SREBP-1	P36956	sepd	410			[Wang et al. 1996]
#	**Name**	**S.W**	**Motif 1**		**Motif 2**		**Reference**
104	XIAP (IAP3)	P98170	sesd	242			[Deveraux et al. 1999]
105	ERBB2	P04626	setd	1125			[Tikhomirov and Carpenter 2001]
106	PAK2	Q13177	shvd	212			[Walter et al. 1998]
107	GRASP65*	Q9BQQ3	slld	316	sfld	371	[Lane et al. 2002]
108	Gas2	O43903	srvd	278			[Sgorbissa et al.1999]
109	PKCepsilon	Q02156	sspd	383	ddvd	451	[Basu et al. 2002]
110	IL-16	Q14005	sstd	510			[Zhang et al. 1998]
111	Apaf-1	O14727	svtd	271			[Bratton et al. 2001]
112	Beta-Catenin	P35222	dlmd	764	ypvd	751	[Steinhusen et al. 2000]
113	BLM	P54132	tevd	415			[Bischof et al.2001]
114	Lamin A/C	P02545	veid	230			[Rao et al. 1996]
115	Keratin 18	P05783	vevd	238	dald	396	[Caulin et al. 1997]
116	APP	P05067	vevd	739			[Gervais et al. 1999]
117	4E-BP1	Q13541	vlgd	25			[Tee and Proud 2002]
118	p21-Rac1	P63000	vvgd	11	vmvd	47	[Zhang et al. 2003]
119	PITSLRE	Q9UQ88	yvpd	394			[Beyaert et al. 1997]

* PTEN has another two motifs: nepd (375), and (384).

* GRASP65 has another motif: tlpd (389).

Test Substrates Group "TSG"

#	Name	S.W	Motif 1		Motif 2		Reference
1	CDC6	Q99741	sevd	442			[Yim et al. 2003]
2	CEACAM1	P13688	dqrd	465			[Houde et al. 2003]
3	Cten	Q8IZW8	dstd	570			[Lo et al. 2005]
4	BUB1B_HUMAN	O60566	dtcd	610			[Kim et al. 2005]
5	MET_HUMAN	P08581	esvd	1002			[Tulasne et al. 2004]
6	BAT3_HUMAN	P46379	deqd	1001			[Wu et al. 2004]
7	HDAC4_HUMAN	P56524	dvtd	289			[Liu et al. 2004]
8	MCL1	Q07820	eeld	127	tstd	157	[Weng et al. 2005]
9	DLG1	Q12959	qpvd	427			[Gregorc et al. 2005]
10	p27Kip1	P46527	dspd	139			[Eymin et al. 1999]
11	Nedd4	P46934	dqpd	279			[Harvey et al. 1998]
12	cPLA2	P47712	deld	523			[Luschen et al. 1998]
13	Gamma-ECS	P48506	avvd	499			[Franklin et al. 2002]
14	CD2L2	Q9UQ88	ypvd	394			[Beyaert et al. 1997]
15	GGTase I	P49354	vsld	59			[Kim et al. 2001)]
16	Nup153	P49790	ditd	349			[Buendia et al. 1999]
17	Presenilin 2	P49810	dsyd	329			[Vito et al. 1997]
18	APLP1	P51693	vevd	620			[Galvan et al. 2002]
19	ETK (BMX)	P51813	dfpd	242			[Wu et al. 2001]
#	Name	S.W	Motif 1		Motif 2		Reference
20	SSRP1	Q08945	dqhd	450			[Landais et al. 2006]
21	RB	P06400	dsid	349			[Katsuda et al. 2002]
22	Giantin	Q14789	dvtd	1946	dasd	1137	[Lowe et al. 2004]
23	integrin beta4	P16144	deld	1109			[Werner et al. 2007]
24	TAO1	Q7L7X3	dvsd	376			[Zihni et al. 2006]
25	TAO2	Q9UL54	dpgd	919			[Zihni et al. 2007]

Species database

#	Substrate	Motif	Human ID	Mouse ID	Other Species			
1	PKN	DFLD	Q16512	P70268				
2	Bcl-2	DAGD	P10415	P10417	Rat	P49950	Bovine	O02718
3	BAX-alpha	FIQD	Q07812	Q07813	Rat	Q63690	Bovine	O02703
4	c-IAP1	ENAD	Q13490	Q62210				
5	Beta-Catenin	(all)	P35222	Q02248	Rat	Q9WU82		
6	CD-IC	DSGD	O14576	O88485				
7	Beta-actin	ELPD	P60709	P60710	Rat	P60711		
8	TCR-zeta chain	DTYD	P20963	P24161	Rabbit	Q9TUF8	Pig	Q9XSJ9
9	GrpL	DIND	O75791	O89100				
10	PARG	MDVD	Q86W56	O88622	Rat	Q9QYM2		
11	4E-BP1	VLGD	Q13541	Q60876	Rat	Q62622		
#	Substrate	Motif	Human ID	Mouse ID	Other Species			
12	ERBB2	SETD	P04626		Rat	P06494		
13	eIF4G2	DETD	P78344	Q62448	Rabbit	P79398		
14	eIF4G1	DLLD	Q04637		Rabbit	P41110		
15	SRP72	SELD	O76094		Dog	P33731		
16	DCC	LSVD	P43146	P70211				
17	IL16	SSTD	Q14005	O54824				
18	IL18	DMTD	Q14116	P70380				
19	TnT2	DFDD	P45379	P50752	Rabbit	P09741	Rat	P50753
20	MDM4	DVPD	O15151	O35618				
21	LAP2A	DKDD	P42166	Q61033				
22	Vimentin	DSVD	P08670	P20152	Rat	P31000	Pig	P02543
23	Gelsolin	DQTD	P06396	P13020				
24	p27Kip1	ESQD	P46527	P46414				
25	Keratin 18	VEVD	P05783	P05784				
26	Lamin A	VEID	P02545	P48678	Rat	P48679		
27	Alpha-II spectrin	DSLD	Q13813	P16546	Rat	P16086		
28	TRAF1	LEVD	Q13077	P39428				
29	FYN	EERD	P06241	P39688				
30	LYN	DGVD	P07948	P25911				
31	AKT	ECVD	P31749	P31750	Rat	P47196		
32	DCAMKL1	DEND	O15075	Q9JLM8	Rat	O08875		
33	HPK1	DDVD	Q92918	P70218				
34	GGTase I	VSLD	P49354	Q61239	Rat	Q04631		
35	ROCK 1	DETD	Q13464	P70335				

#	Substrate	Motif	Human ID	Mouse ID	Other Species			
36	SPAK	DEMD	Q9UEW8	Q9Z1W9	Rat	O88506		
37	Calcineurin	DGFD	Q08209	P63328	Rat	P63329		
38	Mst 2	DELD	Q13188	Q9JI10				
39	PP2A subunit A	DEQD	P30153	Q76MZ3				
40	APLP1	VEVD	P51693	Q03157				
41	TIAM 1	DETD	Q13009	Q60610				
42	PDE6A	DFVD	P16499	P27664				
43	PDE10A2	DLFD	Q9ULW9		Rat	Q9QYJ6		
44	RasGAP	DTVD	P20936		Rat	P50904		
45	CDC42	DLRD	P60953	P60766				
46	Calsenilin	DSSD	Q9Y2W7	Q9QXT8	Rat	Q9JM47		
47	PLCG1	AEPD	P19174		Rat	P10686		
48	iPLA2	DVTD	O60733	P97819	Rat	P97570		
49	Vav1	DLYD	P15498	P27870	Rat	P54100		
50	APP	VEVD	P05067	P12023	Rat	P08592		
51	Presenilin 2	DSYD	P49810	Q61144				
52	RAP	DESD	Q15276	O35551	Rat	O35550		

Amino acids table

#	Amino Acid	3-letter code	1-letter code	Properties
1	Aspartate	Asp	D	Acidic, hydrophilic charged (-)
2	Glutamate	Glu	E	Acidic, hydrophilic charged (-)
3	Histidine	His	H	Basic, hydrophilic charged (+)
4	Lysine	Lys	K	Basic, hydrophilic charged (+)
5	Arginine	Arg	R	Basic, hydrophilic charged (+)

6	Asparagine	Asn	N	Polar, hydrophilic, neutral
7	Glutamine	Gln	Q	Polar, hydrophilic, neutral
8	Serine	Ser	S	Polar, hydrophilic, neutral
9	Threonine	Thr	T	Polar, hydrophilic, neutral
10	Tyrosine	Tyr	Y	Polar, hydrophilic, neutral
11	Alanine	Ala	A	Hydrophobic, neutral
12	Leucine	Leu	L	Hydrophobic, neutral
13	Proline	Pro	P	Hydrophobic, neutral
14	Methionine	Met	M	Hydrophobic, neutral
15	Glycine	Gly	G	Hydrophobic, neutral
16	Valine	Val	V	Hydrophobic, neutral
17	Isoleucine	Ile	I	Hydrophobic, neutral
18	Phenylalanine	Phe	F	Hydrophobic, neutral
19	Tryptophan	Trp	W	Hydrophobic, neutral
20	Cysteine	Cys	C	Hydrophobic, neutral

Substrates cleavage sites prediction using CAT3

Substrate	Protein ID	Motif	Position	Final Score
Desmocollin 3	Q14574	DEND	238	**82**
XRCC4	Q13426	DVTD	265	**78**
Filamin	P21333	DVVD	1500	**78**

SRPK-2	P78362	DEED	407	**64**
Desmoplakin	P15924	DVLD	1641	**60**
NF-kBP50	P19838	AHVD	713	**59**
SRF	P11831	SESD	254	**58**
Cbl-b	Q13191	DVFD	770	**55**
alpha-Actinin	P12814	DFRD	61	**54**
Wee-1	P30291	DEDD	451	**53**
Topo-II alpha	P11388	DDSD	1475	**52**
NS1	O60506	DERD	382	**52**
alpha-Actin	P68032	DSGD	159	**50**
Ran-GAP1	P46060	DAVD	495	**48**
MCM3	P25205	DLVD	227	**48**
P70-56k	P23443	DSPD	396	**44**
PDE5A1	O76074	DCSD	378	**44**
TXBP151	Q86VP1	DSED	693	**43**
tTG	P21980	DVVD	403	**43**
CALM	O60641	DIPD	266	**43**
Relish	Q94527	DLLD	938	**41**
Substrate	**Protein ID**	**Motif**	**Position**	**Final Score**
FKBP46	Q26486	VVVD	45	**41**
Cortactin	Q12860	DGGD	984	**39**
HSF	Q00613	DHLD	389	**38**
N-cadherin	P19022	DIGD	834	**38**
RAP-alpha	P10276	DRVD	349	**38**
gama-Catenin	Q86W21	DDLD	698	**37**
c-Rel	Q04864	DCRD	86	**35**
Emerin	P50402	DMYD	75	**35**
HS1	P14317	DRVD	206	**35**
Cbl	P22681	DGYD	775	**34**
hTAF II (80)	P49848	DNQD	446	**32**
SRPK-1	Q96SB4	DPND	232	**31**
P150	Q14203	DTAD	302	**31**
PABP4	Q13310	DTID	306	**31**
NAC-alpha	Q13765	TESD	30	**27**
MEK	Q02750	VEGD	281	**25**
PAI-2	P05120	GSVD	191	**24**
LBR	Q14739	PLID	288	**20**
PDC-E2	P10515	RVVD	591	**20**
CaMK-II alpha	Q9UQM7	SEAD	111	**17**
Plakophilin-1	Q13835	SEPD	158	**17**
Src	P12931	ASAD	44	**17**
alpha-Tubulin	Q71U36	IQPD	33	**16**

The codes

CAT3 code

```
#! /usr/bin/perl

use submodules;
use ScoreMatrices;

#ouput file = CAT3.txt
$output="CAT3";
unless(open(CAT3,">$output.txt")){print
"cannot open file\"$output\"to write to !!\n\n";exit;}

# input the protein file as text file, remove white spaces
print "Please enter the name of the proteins file:\n";
$filename=<STDIN>;
chomp $filename;

unless (open (SWISSID,$filename)){print "can not open the file
$filename\n";exit;}
@protein=<SWISSID>;
close SWISSID;

#remove the all rows that may contain notes about the protein and
store it in $notes
#the name of the protein will be taken out from these notes and
printed using the
#submodule FILENAME

my $notes='';
foreach $line(@protein){if($line=~/^>/)
{ $notes .=$line;next;}else{ $protein .=$line;}}
$proteinNAME=FILENAME($notes);
#print"$proteinNAME\n";
print"\n\nRESULTS ARE FOUND IN THE FILE : CAT3.TXT\n\n";
print"RESULTS ARE FOUND IN THE FILE : CAT3.TXT\n\n";
#Now we have the protein as Amino acids without any notes.
# All white spaces in any line will be removed
chomp $protein;
$protein=~ s/\s//g;
$Pl=length $protein;

#Counting the number of Aspartate 'D' in the protein and the site of
each 'D'
$counter=0;
@dPos=[];
@protein=split('',$protein);

   for($i=0;$i<scalar @protein;$i++)
```

```
     {
   if (@protein[$i] eq 'D' )
          {$dPos[$counter]=$i+1;
          $counter++;
       }

   $dSum= scalar @dPos;
   }
#print "\n\nThis protein contains $dSum aspartic acids in the
following positions:\n @dPos\n";

#Cutting the protein to strings according to the position of each
'D'. Max 14 A.A
# 5-XXXD-5
@strings=[];
    for($i=0;$i< scalar @dPos;$i++){
           $string='';$string1='';$string2='';

             if ($dPos[$i]<9){
               for( $j=$dPos[$i]-1;$j>0;$j--){
                   $p=$j-1;
                   $string1.=$protein[$p];}
      $string1= reverse $string1;
      $string2=substr($protein,$dPos[$i]-1,6);
      $string=$string1.$string2;
      $strings[$i]=$string;
           }
 elsif ($dPos[$i]+5> scalar @protein){
         for( $j=$dPos[$i]-9;$j<scalar @protein;$j++){
              $string.=$protein[$j];}
              $strings[$i]=$string;
           }

else{ $string=substr($protein,$dPos[$i]-9,14);
   $strings[$i]=$string;

         }
         }

#now we have two vectors one contains the positions of 'D'= @dPos
#and the other one contains the strings of each 'D'= @strings
#each string in the @strings vector is assigned to a score

$j=0;
@Strings=@strings;
@Ps_Pa='';
@maxScore='';
@minScore='';
@p_score='';
    foreach $s (@Strings)
       {
```

```
 chomp $s;
  @s=split('',$s);
  $slength= length $s;
  if (length $s ==14)
   {

     if ($s[5] eq 'D')
      {
    my @score= DXXDscore($s);
#max score=P9+P8+P7+P6+P5+P3+P2+P'1+P'2+P'3+P'4+P'5
#min score=P5+P3+P2+P'1+P'2
$minScore[$j]=int((($score[1]/5)+5)/13.4*100);
$maxScore[$j]=int((($score[0]/12)+5)/9.3*100);
$p_score[$j]=int(($minScore[$j]+$maxScore[$j])/2);
$Ps_Pa[$j]=$score[2];
   }
  else  {  my @score=XXXDscore($s);
#max score=P9+P8+P7+P6+P5+P4+P3+P2+P'1+P'2+P'3+P'4+P'5
#min score=P5+P4+P3+P2+P'1+P'2
$minScore[$j]=int(($score[1]/6)/21.5*100);
$maxScore[$j]=int(($score[0]/13)/14.9*100);
$p_score[$j]=int(($minScore[$j]+$maxScore[$j])/2);
$Ps_Pa[$j]=$score[2];

  }
 }

  else {
       $ss=$s;

for(my $i=0;$i<length $s;$i++)
   { if ( $s[$i] eq D )
    {   $l=(length $s)-$i-1;
      if($i==8 or $l==5)
      {
   while (length $s<14){
     my $X='-';
    if($i==8)
      { $s=$s.$X;}
elsif($l==5){$s=$X.$s}
 };
}

}

  }
  if($s[$i-2] eq D)
  {
    my @score= DXXDscore($s);
#average of sum(P5->P'2)/Highest possible sum.(=13.4 for P5->P'2 and
=9.3 for P9->P'5) to get the %
 $minScore[$j]=int((($score[1]/5)+5)/13.4*100);
```

```
 $maxScore[$j]=int((($score[0]/($length-2))+5)/14.9*100);
 $p_score[$j]=int(($minScore[$j]+$maxScore[$j])/2);
 $Ps_Pa[$j]=$score[2];
  }
else {  my @score=XXXDscore($s);
 $minScore[$j]=int(($score[1]/6)/21.5*100);
 $maxScore[$j]=int(($score[0]/($slength-1))/13*100);
  $p_score[$j]=int(($minScore[$j]+$maxScore[$j])/2);
   $Ps_Pa[$j]=$score[2];
             }
    }

   $j=$j+1;
 }

@generalScore='';
$h=0;
@motifs='';
@finalScore='';
foreach $s14 (@Strings)
   {
      $s14Length=length $strings[$h];
      @s14=split('',$s14);
        $motifs[$h]=$s14[5].$s14[6].$s14[7].$s14[8];
               $generalScore[$h]=int((generalScore($s14)/($s14Length-
1))/12*100);
        if($p_score[$h] >100){$p_score[$h]=100;}
        if($Ps_Pa[$h]>100){$Ps_Pa[$h]=100;}
                 $finalScore[$h]=int(($p_score[$h]+$generalScore[$h]+
$Ps_Pa[$h])/3);
# as the highest  score we got after running the algorithm was 91
# score is modified = finalScore/91*100;
        $finalScore[$h]=int($finalScore[$h]/91*100);
        if($finalScore[$h]>100){$finalScore[$h]=100;}
        $h=$h+1;
      }

#printing the final matrices in organized way

#print"\n\nMotif\t\t\tPosition\t\t Score\n";
#print"========\t\t=========\t\t============\n";

#print results to the output "CAT3.txt"
print CAT3 "   ANALAYSIS REPORT OF CAT3 PROGRAM\n";
print CAT3 "   ================================\n\n\n";
print CAT3 "PROTEIN INFORMATION:\n";
print CAT3 "--------------------\n";
print CAT3 "Protein Name: $proteinNAME\n";
print CAT3 "Prortein Length: $Pl\n";
print CAT3 "Number of Asparatic acids in the Protein $proteinNAME:
$dSum\n";
```

```
print   CAT3   "Positions   of   Asparatic   acids   in   the   protein
$proteinNAME: @dPos\n";

print CAT3 "\n\nCLEAVAGE PREDICTION ANALYSIS:\n";
print CAT3      "----------------------------\n";
print CAT3 "No.\t";
print CAT3 "Protein\t";
print CAT3 "Strings\t\t\t";
print CAT3 "Motif\t";
print CAT3 "Position\t";
print CAT3 "Score\n";
print
CAT3"================================================================\
n";

for(my $k=0;$k< scalar @dPos;$k++)
   {
  if($finalScore[$k]>30)
  {
#    print "$motifs[$k]\t\t\t$dPos[$k]\t\t\t\t\t$finalScore[$k]\n";
   }

$no=$k+1;
print CAT3   "$no\t";
print CAT3   "$proteinNAME\t";
print CAT3   "$strings[$k]\t\t";
print CAT3   "$motifs[$k]\t";
print CAT3   "$dPos[$k]\t\t";
print CAT3   "$finalScore[$k]\n";

}
$flag=0;

my @fscore=sort{$b<=>$a} @finalScore;

for(my $k=0;$k< scalar @dPos;$k++)
 {    if($fscore[0]==$finalScore[$k])
  {           $flag=$k;}
  }

$cutVal=$dPos[$flag];
$newprot='';
for($b=0;$b<length $protein;$b++){
if($b == $cutVal-1){
$newprot .="$protein[$b] >< ";}
else {$newprot .=$protein[$b];}}

print CAT3"\n\n\n";
print CAT3"SIGNIFICANT CLEAVAGE SITE\n";
print CAT3 "-------------------------\n";

#print "\n\n\n ";
```

```
print_sequence($newprot,50);
print CAT3"\n\nThe most probable cleavage site of the protein
$proteinNAME is\n";
print CAT3" $motifs[$flag] at position $dPos[$flag] : Score =
$fscore[0]\n";

print CAT3" \n\nCOMMENTS:\n";
print CAT3     "--------\n";
print CAT3"If the Score > 30,cleavage is True \n";
print CAT3"If the Score =< 30, cleavage is False\n";
print CAT3"If the Score <=0 , cleavage will never happen
mostly\n\n\n";

print CAT3"CAT3: a powerful bioinformatics tool for prediction of
Caspase-3 substrate cleavage site.
Developed by Muneef Ayyash and Yaqoub Ashhab.
Biotechnology Research and Training Unit (BioTRU)
Palestine Polytechnic University,
Hebron, Palestine
Biotech.ppu.edu\n";
$email='yashhab@ppu.edu';
print CAT3"For further correspondence: ($email).\n";

#print" \n\nThe most probable cleavage site of the protein
$proteinNAME is\n";
#print" $motifs[$flag] at position $dPos[$flag] : Score =
$fscore[0]\n";

close (CAT3);
$stop_the_screen=<STDIN>;

exit;
```

Sub-modules of the code

submodules.pm

```
sub MOTIF{
my($input)=@_;
my($motif)=$input;
chomp $motif;
$motif=~s/\s//g;
$motif=~ tr/abcdefghijklmnopqrstuvwxyz/ABCDEFGHIJKLMNOPQRSTUVWXYZ/;
return $motif;}


sub INPUTFILE{
my($input)=@_;
my($file)=$input;
$file=~s/\s//g;
chomp $file;
unless (open (SWISSID,$file)){print "cannot open the file
$file\n";exit;}
@pro=<SWISSID>;
close SWISSID;
return @pro;}

sub FILENAME{
my($input)=@_;
my($note)=$input;
$a='|';@name=();@c=();
$note=~ s/\s//g;
@note=split('',$note);
for($i=0;$i<scalar @note;$i++)
{ if($note[$i] eq $a)
{push(@c,$i);}}
for($j=$c[0];$j<$c[1]-1;$j++)
{push(@name,$note[$j+1]);}
$name=join('',@name);$name=~ s/\s//g; return $name; $notes='';}


sub AACOUNT{
my(@input)=@_;
my(@protein)=@input;


$protein=join('',@protein);
$protein=~ s/\s//g;
for ($i=0;$i<=9;$i++){$protein=~ s/$i//g;}
$protein=~tr/abcdefghijklmnopqrstuvwxyz/ABCDEFGHIJKLMNOPQRSTUVWXYZ/;

@PROTEIN=split('',$protein);
$count_A=0;$count_C=0;$count_D=0;$count_E=0;$count_F=0;
$count_G=0;$count_H=0;$count_I=0;$count_K=0;$count_L=0;
$count_M=0;$count_N=0;$count_P=0;$count_Q=0;$count_R=0;
```

```
$count_S=0;$count_T=0;$count_V=0;$count_W=0;$count_Y=0;$count_error=
0;

foreach $base (@PROTEIN){if ($base eq 'A'){++$count_A;}
      elsif($base eq 'C'){++$count_C;}
      elsif($base eq 'S'){++$count_S;}
      elsif($base eq 'D'){++$count_D;}
      elsif($base eq 'T'){++$count_T;}
      elsif($base eq 'E'){++$count_E;}
      elsif($base eq 'F'){++$count_F;}
      elsif($base eq 'V'){++$count_V;}
      elsif($base eq 'G'){++$count_G;}
      elsif($base eq 'W'){++$count_W;}
      elsif($base eq 'H'){++$count_H;}
      elsif($base eq 'I'){++$count_I;}
      elsif($base eq 'Y'){++$count_Y;}
      elsif($base eq 'K'){++$count_K;}
      elsif($base eq 'L'){++$count_L;}
      elsif($base eq 'M'){++$count_M;}
      elsif($base eq 'N'){++$count_N;}
      elsif($base eq 'P'){++$count_P;}
      elsif($base eq 'Q'){++$count_Q;}
      elsif($base eq 'R'){++$count_R;}
      else { ++$count_error;}
          }

$TOTAL= $count_A+$count_C+$count_D+$count_E+$count_F+$count_G+
$count_H+
$count_I+$count_K+$count_L+$count_M+$count_N+$count_P+$count_Q+
$count_R+$count_S+$count_T+$count_V+$count_W+$count_Y;
return $TOTAL;}

sub AAcategories{
my ($input)=@_;
my ($AAA)=$input;

if( $AAA=~ /D/) {return 'A'}
elsif($AAA=~ /E/i) {return 'A'}
elsif($AAA=~ /H/i) {return 'B'}
elsif($AAA=~ /K/i) {return 'B'}
elsif($AAA=~ /R/i) {return 'B'}
elsif($AAA=~ /N/i) {return 'P'}
elsif($AAA=~ /Q/i) {return 'P'}
elsif($AAA=~ /S/i) {return 'P'}
elsif($AAA=~ /T/i) {return 'P'}
elsif($AAA=~ /Y/i) {return 'P'}
elsif($AAA=~ /A/i) {return 'N'}
elsif($AAA=~ /L/i) {return 'N'}
elsif($AAA=~ /P/i) {return 'N'}
elsif($AAA=~ /M/i) {return 'N'}
elsif($AAA=~ /G/i) {return 'N'}
elsif($AAA=~ /V/i) {return 'N'}
```

```
elsif($AAA=~ /I/i) {return 'N'}


elsif($AAA=~ /F/i) {return 'N'}
elsif($AAA=~ /W/i) {return 'N'}
elsif($AAA=~ /C/i) {return 'N'}
else{print"*";}}

sub AAconverter{
my($input)=@_;
my($string)=$input;
my $newString='';
for ( $i=0;$i<length $string;$i++)
      { $AA=substr($string,$i,1);
        $newString .=AAcategories($AA);
       } return $newString; }


sub AAcounter{
my ($input)=@_;
my ($protline)=$input;

@PROTEIN=split('',$protline);
$count_A=0;$count_C=0;$count_D=0;$count_E=0;$count_F=0;$count_G=0;
$count_H=0;$count_I=0;$count_K=0;$count_L=0;$count_M=0;$count_N=0;
$count_P=0;$count_Q=0;$count_R=0;$count_S=0;$count_T=0;
$count_V=0;$count_W=0;$count_Y=0;$count_error=0;
foreach $base (@PROTEIN){if ($base eq 'A'){++$count_A;}
      elsif($base eq 'C'){++$count_C;}
      elsif($base eq 'S'){++$count_S;}
      elsif($base eq 'D'){++$count_D;}
      elsif($base eq 'T'){++$count_T;}
      elsif($base eq 'E'){++$count_E;}
      elsif($base eq 'F'){++$count_F;}
      elsif($base eq 'V'){++$count_V;}
      elsif($base eq 'G'){++$count_G;}
      elsif($base eq 'W'){++$count_W;}
      elsif($base eq 'H'){++$count_H;}
      elsif($base eq 'I'){++$count_I;}
      elsif($base eq 'Y'){++$count_Y;}
      elsif($base eq 'K'){++$count_K;}
      elsif($base eq 'L'){++$count_L;}
      elsif($base eq 'M'){++$count_M;}
      elsif($base eq 'N'){++$count_N;}
      elsif($base eq 'P'){++$count_P;}
      elsif($base eq 'Q'){++$count_Q;}
      elsif($base eq 'R'){++$count_R;}
      else { ++$count_error;}

return $count_A;return $count_C;return $count_D;return $count_E;
return $count_F;return $count_G;return $count_H;return $count_I;
return $count_K;return $count_L;return $count_M;return $count_N;
```

```
return $count_P;return $count_Q;return $count_R;return $count_S;
return $count_T;return $count_V;return $count_W;return $count_Y;
return $count_error;}}

sub print_sequence {

my($sequence, $length) = @_;
# Print sequence in lines of $length
for ( my $pos = 0 ; $pos < length($sequence) ; $pos +=
$length ) {
print substr($sequence, $pos, $length), "\n";


print CAT3T substr($sequence, $pos, $length), "\n";}
}
1
```

ScoreMatrices.pm

```
sub DXXDscore{
my($input)=@_;
my($s)=$input;
chomp $s;
$string=$s;
@string=split('',$string);

#Score Matrix for DXXD Probability difference between cleaved and
uncleaved =%P(i)cleaved-%P(i)uncleaved
%dxxdp9=('D'=>'1.3','E'=>'2.8','H'=>'0.6','K'=>'-2.2','R'=>'-
1.1','N'=>'-0.8','Q'=>'-2.8','S'=>'-1.7','T'=>'-1.8','Y'=>'-
2.8','A'=>'0.9','L'=>'1.5','P'=>'1.6','M'=>'-
0.4','G'=>'8.2','V'=>'3.4','I'=>'-4.8','F'=>'-
3.2','W'=>'1.6','C'=>'-0.1');
%dxxdp8=('D'=>'-5.1','E'=>'4.3','H'=>'0.2','K'=>'-1.1','R'=>'-
3.9','N'=>'3.3','Q'=>'-6.7','S'=>'2.6','T'=>'0.5','Y'=>'-
1.6','A'=>'-0.9','L'=>'-7.1','P'=>'6.8','M'=>'-
0.8','G'=>'8.9','V'=>'-
1.4','I'=>'0.7','F'=>'0.6','W'=>'0.2','C'=>'0.7');
%dxxdp7=('D'=>'-4.4','E'=>'3.8','H'=>'1.2','K'=>'-
4.6','R'=>'3.1','N'=>'-3.7','Q'=>'-
4.0','S'=>'3.4','T'=>'1.2','Y'=>'-1.3','A'=>'-
1.9','L'=>'5.9','P'=>'1.6','M'=>'0.7','G'=>'2.9','V'=>'3.4','I'=>'-
5.7','F'=>'-2.9','W'=>'0.5','C'=>'1.0');
%dxxdp6=('D'=>'-1.7','E'=>'-4.8','H'=>'-
2.1','K'=>'0.8','R'=>'0.3','N'=>'-
2.0','Q'=>'3.0','S'=>'6.8','T'=>'4.1','Y'=>'-1.4','A'=>'-
0.2','L'=>'-5.9','P'=>'8.0','M'=>'-1.5','G'=>'5.8','V'=>'-
4.1','I'=>'-5.8','F'=>'0.6','W'=>'1.0','C'=>'-0.7');
%dxxdp5=('D'=>'4.3','E'=>'7.0','H'=>'-2.1','K'=>'-2.2','R'=>'-
4.8','N'=>'1.2','Q'=>'2.7','S'=>'-0.7','T'=>'-2.6','Y'=>'-
```

```
0.4','A'=>'-1.0','L'=>'-2.2','P'=>'3.3','M'=>'1.1','G'=>'-
1.2','V'=>'-1.6','I'=>'-0.3','F'=>'-0.6','W'=>'0.4','C'=>'-0.1');
%dxxdp3=('D'=>'-6.8','E'=>'22.0','H'=>'-0.8','K'=>'-3.8','R'=>'-
2.1','N'=>'-2.3','Q'=>'0.2','S'=>'3.6','T'=>'2.2','Y'=>'-
0.5','A'=>'0.7','L'=>'-4.3','P'=>'-
5.1','M'=>'2.1','G'=>'1.4','V'=>'1.1','I'=>'-2.4','F'=>'-
1.8','W'=>'-2.1','C'=>'-1.2');
%dxxdp2=('D'=>'-9.3','E'=>'-10.9','H'=>'0.7','K'=>'-6.6','R'=>'-
2.9','N'=>'-2.0','Q'=>'0.7','S'=>'0.4','T'=>'12.6','Y'=>'-
1.8','A'=>'0.9','L'=>'2.8','P'=>'8.6','M'=>'2.8','G'=>'-
1.2','V'=>'12.6','I'=>'-1.0','F'=>'-2.9','W'=>'-2.4','C'=>'-0.9');
%dxxdpp1=('D'=>'-7.2','E'=>'-7.0','H'=>'-0.2','K'=>'-3.7','R'=>'-
2.6','N'=>'2.0','Q'=>'-4.2','S'=>'16.4','T'=>'-
1.2','Y'=>'1.7','A'=>'-1.3','L'=>'1.4','P'=>'-2.1','M'=>'-
3.0','G'=>'15.4','V'=>'-4.1','I'=>'-1.7','F'=>'-1.2','W'=>'-
1.8','C'=>'4.6');
%dxxdpp2=('D'=>'-6.7','E'=>'-5.5','H'=>'2.7','K'=>'4.4','R'=>'-
1.1','N'=>'-2.8','Q'=>'-2.3','S'=>'4.7','T'=>'-0.3','Y'=>'-
3.6','A'=>'6.3','L'=>'-0.7','P'=>'2.0','M'=>'-
0.5','G'=>'9.1','V'=>'1.2','I'=>'-1.2','F'=>'-4.2','W'=>'-
0.1','C'=>'-1.2');
%dxxdpp3=('D'=>'-4.6','E'=>'1.0','H'=>'0.4','K'=>'-5.0','R'=>'-
0.1','N'=>'-0.1','Q'=>'-5.1','S'=>'5.5','T'=>'3.7','Y'=>'-
0.5','A'=>'3.5','L'=>'1.9','P'=>'1.0','M'=>'-1.1','G'=>'1.8','V'=>'-
1.8','I'=>'0.6','F'=>'-3.4','W'=>'-1.8','C'=>'4.3');
%dxxdpp4=('D'=>'-7.1','E'=>'4.8','H'=>'0.9','K'=>'-2.5','R'=>'-
2.7','N'=>'1.4','Q'=>'-
3.4','S'=>'4.7','T'=>'2.2','Y'=>'0.3','A'=>'2.4','L'=>'2.9','P'=>'3.
5','M'=>'1.7','G'=>'1.5','V'=>'-4.1','I'=>'-0.3','F'=>'-3.1','W'=>'-
3.3','C'=>'0.4');
%dxxdpp5=('D'=>'3.3','E'=>'-0.1','H'=>'-
0.7','K'=>'0.8','R'=>'0.1','N'=>'0.5','Q'=>'-1.4','S'=>'2.0','T'=>'-
1.6','Y'=>'3.1','A'=>'0.1','L'=>'0.1','P'=>'-3.1','M'=>'3.2','G'=>'-
0.4','V'=>'0.4','I'=>'-2.7','F'=>'-3.6','W'=>'-0.6','C'=>'0.8');



$dxxdscore[0]=$dxxdp9{$string[0]}+$dxxdp8{$string[1]}+
$dxxdp7{$string[2]}+$dxxdp6{$string[3]}+$dxxdp5{$string[4]}+
$dxxdp3{$string[6]}+$dxxdp2{$string[7]}+$dxxdpp1{$string[9]}+
$dxxdpp2{$string[10]}+$dxxdpp3{$string[11]}+$dxxdpp4{$string[12]}+
$dxxdpp5{$string[13]};
$dxxdscore[1]=$dxxdp5{$string[4]}+$dxxdp3{$string[6]}+
$dxxdp2{$string[7]}+$dxxdpp1{$string[9]}+$dxxdpp2{$string[10]};

#Score Matrix for DxxD Probability P(i)

%dxxdP6=('-'=>'1','D'=>'0.09','E'=>'0.07','H'=>'0','K'=>'0.04','R'=>
'0.04','N'=>'0.02','Q'=>'0.07','S'=>'0.14','T'=>'0.08','Y'=>'0.02','
```

```
A'=>'0.05','L'=>'0.09','P'=>'0.11','M'=>'0','G'=>'0.09','V'=>'0.02',
'I'=>'0.01','F'=>'0.03','W'=>'0.02','C'=>'0.01');
%dxxdP5=('-'=>'1','D'=>'0.12','E'=>'0.15','H'=>'0.00','K'=>'0.04','R
'=>'0.03','N'=>'0.03','Q'=>'0.07','S'=>'0.08','T'=>'0.02','Y'=>'0.01
','A'=>'0.04','L'=>'0.08','P'=>'0.07','M'=>'0.04','G'=>'0.07','V'=>'
0.04','I'=>'0.03','F'=>'0.03','W'=>'0.02','C'=>'0.01');
%dxxdP3=('-'=>'1','D'=>'0.05','E'=>'0.29','H'=>'0.02','K'=>'0.02','R
'=>'0.03','N'=>'0.02','Q'=>'0.03','S'=>'0.10','T'=>'0.05','Y'=>'0.02
','A'=>'0.05','L'=>'0.08','P'=>'0.00','M'=>'0.03','G'=>'0.04','V'=>'
0.08','I'=>'0.03','F'=>'0.03','W'=>'0.00','C'=>'0.00');
%dxxdP2=('-'=>'1','D'=>'0.03','E'=>'0.01','H'=>'0.02','K'=>'0.00','R
'=>'0.02','N'=>'0.02','Q'=>'0.02','S'=>'0.05','T'=>'0.16','Y'=>'0.03
','A'=>'0.03','L'=>'0.12','P'=>'0.11','M'=>'0.05','G'=>'0.03','V'=>'
0.20','I'=>'0.04','F'=>'0.02','W'=>'0.00','C'=>'0.00');
%dxxdPp1=('-'=>'1','D'=>'0.03','E'=>'0.01','H'=>'0.02','K'=>'0.01','
R'=>'0.02','N'=>'0.08','Q'=>'0.00','S'=>'0.24','T'=>'0.03','Y'=>'0.0
4','A'=>'0.04','L'=>'0.08','P'=>'0.00','M'=>'0.00','G'=>'0.22','V'=>
'0.02','I'=>'0.05','F'=>'0.03','W'=>'0.00','C'=>'0.05');

#Probability of Amino acids in General as computed from the cleaved
proteins =119 proteins.
%AAPi=('-'=>'1','D'=>'0.06','E'=>'0.08','H'=>'0.02','K'=>'0.06','R'=
>'0.06','N'=>'0.04','Q'=>'0.05','S'=>'0.08','T'=>'0.05','Y'=>'0.03',
'A'=>'0.07','L'=>'0.10','P'=>'0.06','M'=>'0.02','G'=>'0.06','V'=>'0.
06','I'=>'0.04','F'=>'0.03','W'=>'0.01','C'=>'0.02');

#dxxdPs is multiplied by 1 to show that 1 replace the P(5) which is
always D in DxxD
$dxxdPa=$AAPi{$string[3]}*$AAPi{$string[4]}*$AAPi{$string[5]}*$AAPi{
$string[6]}*$AAPi{$string[7]}*$AAPi{$string[9]};


$dxxdPs=$dxxdP6{$string[3]}*$dxxdP5{$string[4]}*$dxxdP3{$string[6]}*
$dxxdP2{$string[7]}*$dxxdPp1{$string[9]}*1;
$dxxdscore[2]=int($dxxdPs/$dxxdPa);

return @dxxdscore;}

sub XXXDscore{
my($input)=@_;
my($s)=$input;
chomp $s;
$string=$s;
@string=split('',$string);

#Score Matrix for xxxD Probability difference between cleaved and
uncleaved =%P(i)cleaved-%P(i)uncleaved
%xxxdp9=('D'=>'-1.4','E'=>'-0.1','H'=>'3.9','K'=>'-2.6','R'=>'-
1.3','N'=>'0.3','Q'=>'2.0','S'=>'-1.2','T'=>'2.2','Y'=>'-
2.8','A'=>'5.0','L'=>'1.1','P'=>'-0.8','M'=>'2.2','G'=>'3.3','V'=>'-
4.1','I'=>'-1.9','F'=>'-3.3','W'=>'-0.8','C'=>'0.4');
```

```
%xxxdp8=('D'=>'0.6','E'=>'4.5','H'=>'2.2','K'=>'-2.7','R'=>'-
1.4','N'=>'-4.1','Q'=>'-5.3','S'=>'0.8','T'=>'6.3','Y'=>'-
0.4','A'=>'-1.9','L'=>'6.2','P'=>'3.7','M'=>'-0.1','G'=>'-
3.5','V'=>'1.0','I'=>'-2.1','F'=>'-3.2','W'=>'1.3','C'=>'-1.7');
%xxxdp7=('D'=>'3.1','E'=>'-
3.7','H'=>'0.0','K'=>'4.3','R'=>'3.4','N'=>'-1.7','Q'=>'-
4.4','S'=>'-3.6','T'=>'-0.7','Y'=>'-1.0','A'=>'0.6','L'=>'-
1.7','P'=>'1.8','M'=>'-0.1','G'=>'7.6','V'=>'-1.4','I'=>'-
2.8','F'=>'3.4','W'=>'-1.1','C'=>'-1.9');
%xxxdp6=('D'=>'-3.9','E'=>'-3.1','H'=>'-2.8','K'=>'-6.3','R'=>'-
1.3','N'=>'-1.1','Q'=>'2.2','S'=>'2.6','T'=>'-3.2','Y'=>'-
3.2','A'=>'6.7','L'=>'3.4','P'=>'1.7','M'=>'-
2.4','G'=>'16.4','V'=>'-1.2','I'=>'-2.1','F'=>'-3.9','W'=>'-
1.1','C'=>'2.6');
%xxxdp5=('D'=>'11.6','E'=>'-1.0','H'=>'-2.3','K'=>'-4.5','R'=>'-
3.5','N'=>'4.6','Q'=>'-3.8','S'=>'3.0','T'=>'6.0','Y'=>'-
0.1','A'=>'-4.9','L'=>'-2.5','P'=>'-1.0','M'=>'0.4','G'=>'-
3.4','V'=>'2.9','I'=>'4.2','F'=>'-3.4','W'=>'-0.9','C'=>'-1.5');
%xxxdp4=('D'=>'','E'=>'7.0','H'=>'-2.1','K'=>'-8.3','R'=>'-
5.7','N'=>'-2.2','Q'=>'-2.4','S'=>'19.7','T'=>'-
1.1','Y'=>'1.7','A'=>'4.4','L'=>'-3.9','P'=>'-1.2','M'=>'-
2.5','G'=>'-3.8','V'=>'7.5','I'=>'-4.5','F'=>'-2.0','W'=>'-
1.2','C'=>'0.7');
%xxxdp3=('D'=>'-6.9','E'=>'32.6','H'=>'3.9','K'=>'-5.5','R'=>'-
2.2','N'=>'-1.9','Q'=>'-2.1','S'=>'0.5','T'=>'-4.7','Y'=>'-
2.6','A'=>'-0.2','L'=>'0.0','P'=>'-2.9','M'=>'-0.1','G'=>'-
6.4','V'=>'2.8','I'=>'-2.5','F'=>'-1.2','W'=>'-1.0','C'=>'0.6');
%xxxdp2=('D'=>'-6.2','E'=>'-5.7','H'=>'-1.9','K'=>'-7.1','R'=>'-
4.5','N'=>'-1.0','Q'=>'-2.9','S'=>'-2.9','T'=>'8.6','Y'=>'-
2.5','A'=>'-3.9','L'=>'-1.0','P'=>'13.1','M'=>'-
2.1','G'=>'0.2','V'=>'21.3','I'=>'2.2','F'=>'-3.4','W'=>'-
1.2','C'=>'0.8');
%xxxdpp1=('D'=>'-6.3','E'=>'-8.8','H'=>'2.6','K'=>'-
3.2','R'=>'1.2','N'=>'0.7','Q'=>'-3.9','S'=>'14.2','T'=>'-
5.0','Y'=>'-1.0','A'=>'5.1','L'=>'-10.8','P'=>'-0.5','M'=>'-
1.9','G'=>'32.4','V'=>'-6.1','I'=>'-5.4','F'=>'-2.5','W'=>'-
1.4','C'=>'0.5');
%xxxdpp2=('D'=>'-2.6','E'=>'-4.6','H'=>'0.3','K'=>'0.6','R'=>'-
2.4','N'=>'-1.6','Q'=>'3.0','S'=>'1.4','T'=>'-
4.4','Y'=>'3.4','A'=>'4.7','L'=>'-3.1','P'=>'11.3','M'=>'-
2.5','G'=>'3.2','V'=>'-0.3','I'=>'-1.0','F'=>'-4.0','W'=>'-
1.8','C'=>'0.4');
%xxxdpp3=('D'=>'-3.1','E'=>'-6.1','H'=>'-2.2','K'=>'3.8','R'=>'-
0.1','N'=>'-2.2','Q'=>'0.2','S'=>'-3.4','T'=>'-0.3','Y'=>'-
2.6','A'=>'6.7','L'=>'0.0','P'=>'-0.2','M'=>'-
0.2','G'=>'3.4','V'=>'5.5','I'=>'2.2','F'=>'-1.2','W'=>'-
0.9','C'=>'0.7');
%xxxdpp4=('D'=>'-0.1','E'=>'-8.8','H'=>'-
2.5','K'=>'1.9','R'=>'1.0','N'=>'-
1.7','Q'=>'1.8','S'=>'10.1','T'=>'-
0.5','Y'=>'1.9','A'=>'3.7','L'=>'-3.1','P'=>'-
```

```
4.1','M'=>'4.5','G'=>'3.7','V'=>'-3.8','I'=>'0.1','F'=>'-
1.5','W'=>'-1.2','C'=>'-1.6');
%xxxdpp5=('D'=>'-2.0','E'=>'-3.4','H'=>'0.0','K'=>'-2.8','R'=>'-
3.8','N'=>'-1.4','Q'=>'6.6','S'=>'6.2','T'=>'-2.3','Y'=>'-
0.3','A'=>'7.3','L'=>'-6.1','P'=>'12.7','M'=>'-0.1','G'=>'-
1.3','V'=>'-4.1','I'=>'1.5','F'=>'-3.5','W'=>'-1.3','C'=>'-1.9');


@xxxdscore[0]=$xxxdp9{$string[0]}+$xxxdp8{$string[1]}+
$xxxdp7{$string[2]}+$xxxdp6{$string[3]}+$xxxdp5{$string[4]}+
$xxxdp4{$string[5]}+$xxxdp3{$string[6]}+$xxxdp2{$string[7]}+
$xxxdpp1{$string[9]}+$xxxdpp2{$string[10]}+$xxxdpp3{$string[11]}+
$xxxdpp4{$string[12]}+$xxxdpp5{$string[13]};
@xxxdscore[1]=$xxxdp5{$string[4]}+$xxxdp4{$string[5]}+
$xxxdp3{$string[6]}+$xxxdp2{$string[7]}+$xxxdpp1{$string[9]}+
$xxxdpp2{$string[10]};

#Score Matrix for xxxD Probability P(i)
%xxxdP6=('-'=>'1','D'=>'0.02','E'=>'0.04','H'=>'0.00','K'=>'0.00','R
'=>'0.04','N'=>'0.02','Q'=>'0.07','S'=>'0.11','T'=>'0.02','Y'=>'0.00
','A'=>'0.13','L'=>'0.13','P'=>'0.07','M'=>'0.00','G'=>'0.22','V'=>'
0.04','I'=>'0.02','F'=>'0.00','W'=>'0.00','C'=>'0.04');
%xxxdP5=('-'=>'1','D'=>'0.18','E'=>'0.07','H'=>'0.00','K'=>'0.02','R
'=>'0.02','N'=>'0.09','Q'=>'0.02','S'=>'0.11','T'=>'0.11','Y'=>'0.02
','A'=>'0.02','L'=>'0.07','P'=>'0.04','M'=>'0.02','G'=>'0.02','V'=>'
0.09','I'=>'0.09','F'=>'0.00','W'=>'0.00','C'=>'0.00');
%xxxdP4=('-'=>'1','D'=>'0.00','E'=>'0.16','H'=>'0.00','K'=>'0.00','R
'=>'0.00','N'=>'0.02','Q'=>'0.02','S'=>'0.29','T'=>'0.04','Y'=>'0.04
','A'=>'0.11','L'=>'0.07','P'=>'0.04','M'=>'0.00','G'=>'0.02','V'=>'
0.13','I'=>'0.00','F'=>'0.02','W'=>'0.00','C'=>'0.02');
%xxxdP3=('-'=>'1','D'=>'0.00','E'=>'0.40','H'=>'0.07','K'=>'0.00','R
'=>'0.02','N'=>'0.02','Q'=>'0.02','S'=>'0.09','T'=>'0.00','Y'=>'0.00
','A'=>'0.07','L'=>'0.11','P'=>'0.02','M'=>'0.02','G'=>'0.00','V'=>'
0.09','I'=>'0.02','F'=>'0.02','W'=>'0.00','C'=>'0.02');
%xxxdP2=('-'=>'1','D'=>'0.00','E'=>'0.04','H'=>'0.00','K'=>'0.00','R
'=>'0.02','N'=>'0.02','Q'=>'0.02','S'=>'0.04','T'=>'0.13','Y'=>'0.00
','A'=>'0.02','L'=>'0.09','P'=>'0.18','M'=>'0.00','G'=>'0.07','V'=>'
0.27','I'=>'0.07','F'=>'0.00','W'=>'0.00','C'=>'0.02');
%xxxdPp1=('-'=>'1','D'=>'0.00','E'=>'0.00','H'=>'0.04','K'=>'0.02','
R'=>'0.07','N'=>'0.04','Q'=>'0.00','S'=>'0.22','T'=>'0.00','Y'=>'0.0
2','A'=>'0.11','L'=>'0.00','P'=>'0.04','M'=>'0.00','G'=>'0.38','V'=>
'0.00','I'=>'0.00','F'=>'0.02','W'=>'0.00','C'=>'0.02');


#Probability of Amino acids in General as computed from the cleaved
proteins =159 proteins.
%AAPi=('-'=>'1','D'=>'0.06','E'=>'0.08','H'=>'0.02','K'=>'0.06','R'=
>'0.06','N'=>'0.04','Q'=>'0.05','S'=>'0.08','T'=>'0.05','Y'=>'0.03',
'A'=>'0.07','L'=>'0.10','P'=>'0.06','M'=>'0.02','G'=>'0.06','V'=>'0.
06','I'=>'0.04','F'=>'0.03','W'=>'0.01','C'=>'0.02');
```

```
$xxxdPa=$AAPi{$string[3]}*$AAPi{$string[4]}*$AAPi{$string[5]}*$AAPi{
$string[6]}*$AAPi{$string[7]}*$AAPi{$string[9]};


$xxxdPs=$xxxdP6{$string[3]}*$xxxdP5{$string[4]}*$xxxdP4{$string[5]}*
$xxxdP3{$string[6]}*$xxxdP2{$string[7]}*$xxxdPp1{$string[9]};


if($xxxdPa>0){
$xxxdscore[2]=int($xxxdPs/$xxxdPa);}else{$xxxdscore[2]="n/a";}

return @xxxdscore;}

sub generalScore{
my($input)=@_;
my($s)=$input;
chomp $s;
$string=$s;
@string=split('',$string);


#A.A probability difference between cleaved and uncleaved in the
positions P9->P'5
%P9=('D'=>'1.4','E'=>'2.0','H'=>'1.6','K'=>'-2.6','R'=>'-
0.6','N'=>'-1.1','Q'=>'-1.7','S'=>'0.1','T'=>'-0.1','Y'=>'-
2.1','A'=>'2.0','L'=>'0.4','P'=>'0.0','M'=>'0.0','G'=>'5.5','V'=>'0.
4','I'=>'-3.4','F'=>'-2.0','W'=>'0.6','C'=>'-0.3');



%P8=('D'=>'-1.8','E'=>'3.7','H'=>'0.1','K'=>'-2.0','R'=>'-
2.3','N'=>'0.3','Q'=>'-4.7','S'=>'0.2','T'=>'1.9','Y'=>'-
1.2','A'=>'-0.5','L'=>'-0.8','P'=>'5.1','M'=>'-
0.1','G'=>'4.6','V'=>'-1.9','I'=>'0.0','F'=>'-
1.0','W'=>'0.5','C'=>'-0.3');
%P7=('D'=>'0.6','E'=>'0.7','H'=>'0.7','K'=>'-0.3','R'=>'3.3','N'=>'-
2.5','Q'=>'-3.7','S'=>'1.5','T'=>'0.7','Y'=>'-1.7','A'=>'-
1.0','L'=>'2.0','P'=>'1.1','M'=>'-0.1','G'=>'3.9','V'=>'-
0.5','I'=>'-4.4','F'=>'0.3','W'=>'-0.3','C'=>'-0.4');
%P6=('D'=>'0.2','E'=>'-1.9','H'=>'-2.8','K'=>'-3.2','R'=>'-
1.9','N'=>'-1.1','Q'=>'2.2','S'=>'4.8','T'=>'0.5','Y'=>'-
1.7','A'=>'1.5','L'=>'0.1','P'=>'4.8','M'=>'-2.3','G'=>'7.6','V'=>'-
2.8','I'=>'-3.0','F'=>'-1.7','W'=>'0.4','C'=>'0.3');
%P5=('D'=>'7.6','E'=>'4.8','H'=>'-2.3','K'=>'-3.0','R'=>'-
3.0','N'=>'1.0','Q'=>'-0.7','S'=>'0.7','T'=>'0.0','Y'=>'-
0.8','A'=>'-3.3','L'=>'-1.8','P'=>'0.5','M'=>'1.8','G'=>'-
0.6','V'=>'-0.1','I'=>'0.5','F'=>'-1.2','W'=>'0.5','C'=>'-0.7');
%P4=('D'=>'60.7','E'=>'-2.9','H'=>'-1.9','K'=>'-7.8','R'=>'-
5.4','N'=>'-3.5','Q'=>'-3.6','S'=>'1.0','T'=>'-3.8','Y'=>'-
1.1','A'=>'-2.6','L'=>'-7.7','P'=>'-3.8','M'=>'-2.3','G'=>'-
4.9','V'=>'-1.1','I'=>'-4.3','F'=>'-3.2','W'=>'-1.1','C'=>'-0.7');
```

```
%P3=('D'=>'-3.5','E'=>'25.0','H'=>'0.9','K'=>'-4.0','R'=>'-
1.6','N'=>'-1.9','Q'=>'-1.3','S'=>'1.3','T'=>'-0.9','Y'=>'-
1.2','A'=>'-0.8','L'=>'-2.4','P'=>'-4.4','M'=>'0.7','G'=>'-
3.3','V'=>'2.0','I'=>'-1.9','F'=>'-0.6','W'=>'-1.1','C'=>'-0.9');
%P2=('D'=>'-4.4','E'=>'-8.0','H'=>'-0.4','K'=>'-7.1','R'=>'-
4.4','N'=>'-1.1','Q'=>'-2.7','S'=>'-2.1','T'=>'10.8','Y'=>'-
0.4','A'=>'-3.0','L'=>'1.1','P'=>'8.7','M'=>'1.5','G'=>'-
1.9','V'=>'16.6','I'=>'0.6','F'=>'-2.0','W'=>'-1.2','C'=>'-0.7');
%Pp1=('D'=>'-4.4','E'=>'-8.0','H'=>'1.1','K'=>'-3.9','R'=>'-
1.7','N'=>'2.7','Q'=>'-3.9','S'=>'15.5','T'=>'-
2.7','Y'=>'0.5','A'=>'0.6','L'=>'-5.4','P'=>'-3.3','M'=>'-
1.9','G'=>'21.8','V'=>'-4.7','I'=>'-1.8','F'=>'-1.8','W'=>'-
1.5','C'=>'2.7');
%Pp2=('D'=>'-2.8','E'=>'-4.7','H'=>'1.1','K'=>'3.4','R'=>'-
2.4','N'=>'-2.4','Q'=>'0.0','S'=>'2.9','T'=>'-2.2','Y'=>'-
1.0','A'=>'4.0','L'=>'-0.3','P'=>'6.0','M'=>'-
1.1','G'=>'5.5','V'=>'1.8','I'=>'-1.7','F'=>'-4.0','W'=>'-
1.0','C'=>'-1.0');
%Pp3=('D'=>'-3.2','E'=>'-1.7','H'=>'-0.7','K'=>'-
2.1','R'=>'0.6','N'=>'-0.7','Q'=>'-
2.9','S'=>'1.8','T'=>'2.6','Y'=>'-
1.2','A'=>'3.1','L'=>'0.1','P'=>'2.5','M'=>'-
0.3','G'=>'1.9','V'=>'0.3','I'=>'0.1','F'=>'-2.0','W'=>'-
1.0','C'=>'2.9');
%Pp4=('D'=>'-3.0','E'=>'0.1','H'=>'-0.3','K'=>'-1.1','R'=>'-
1.2','N'=>'-0.2','Q'=>'-
1.9','S'=>'5.7','T'=>'2.3','Y'=>'1.1','A'=>'0.9','L'=>'-
0.1','P'=>'1.0','M'=>'3.0','G'=>'2.1','V'=>'-3.8','I'=>'-
0.6','F'=>'-2.3','W'=>'-1.4','C'=>'-0.2');
%Pp5=('D'=>'1.6','E'=>' 0.8','H'=>'-0.7','K'=>'-0.6','R'=>'-
1.6','N'=>'0.0','Q'=>'0.7','S'=>'4.5','T'=>'-
0.9','Y'=>'1.9','A'=>'2.1','L'=>'-3.9','P'=>'6.0','M'=>'-
0.1','G'=>'-1.3','V'=>'-1.8','I'=>'-0.8','F'=>'-3.5','W'=>'-
1.3','C'=>'-1.0');

$gscore=$P9{$string[0]}+$P8{$string[1]}+$P7{$string[2]}+
$P6{$string[3]}+$P5{$string[4]}+$P4{$string[5]}+$P3{$string[6]}+
$P2{$string[7]}+$Pp1{$string[9]}+$Pp2{$string[10]}+$Pp3{$string[11]}
+$Pp4{$string[12]}+$Pp5{$string[13]};
return $gscore;}1
```

Other Perl files:

supporting files that were used in the database analysis

species.pl

```
#!/usr/bin/perl
use submodules;

#ouput file
```

```
$outputfile="outfile";
unless(open(OUTFILE,">$outputfile.xls")){print"cannot  open  file  \
"$outputfile\ "to write to !!\n\n";exit;}

#print results to the outputfile "outfile.xls"
print OUTFILE "S.P NUM.\t";
print OUTFILE "PROTEIN LENGTH\t";
print OUTFILE "CLEAVAGE SITE\t";

print OUTFILE "Amino acids before C.S(up to 50)\t";
print OUTFILE "A.A properties\t";
print OUTFILE "%A1\t";
print OUTFILE "%B1\t";
print OUTFILE "%P1\t";
print OUTFILE "%N1-Phobic\t";
print OUTFILE "%H1-Philic\t";

print OUTFILE "Amino acids after C.S(up to 50)\t";
print OUTFILE "A.A properties\t";
print OUTFILE "%A2\t";
print OUTFILE "%B2\t";
print OUTFILE "%P2\t";
print OUTFILE "%N2-Phobic\t";
print OUTFILE "%H2-Philic\t";



print OUTFILE "TOTAL Amino acids(up to 104)\t";
print OUTFILE "TOTAL A.A properties\t";
print OUTFILE "%Atotal\t";
print OUTFILE "%Btotal\t";
print OUTFILE "%Ptotal\t";
print OUTFILE "%N-Phobic\t";
print OUTFILE "%H-Philic\n";

#input the protein file, remove fasta format and intialize @protein;
and #print the protein file
print "please enter the file name of the SWISSPROT IDs:\n";
$mainFile=<STDIN>;
chomp $mainFile;
unless  (open  (SWISSID,$mainFile)){print  "can  not  open  the  file
$mainFile\n";exit;}
@fileName=<SWISSID>;
close SWISSID;


print "please enter the file name of the motifs:\n";
$motiFile=<STDIN>;
chomp $motiFile;
unless  (open  (MOTIFS,$motiFile)){print  "can  not  open  the  file
$motiFile\n";exit;}
@motifName=<MOTIFS>;
```

```
close MOTIFS;

my $z=0;


do {

$a=$fileName[$z];chomp $a; $a=~ s/\s//g;
my $notes='';
my @protein=INPUTFILE($a);
my $protein='';
foreach            $line(@protein){if($line=~/^>/){            $notes
.=$line;next;}else{ $protein .=$line;}}

$proteinFILENAME=FILENAME($notes);
print"$proteinFILENAME\n";

#remove white spaces
$protein=~ s/\s//g;

#enter the motif to cutoff andintialize $motif
my $motif=$motifName[$z];
#print "please enter the motif for $proteinFILENAME :\n";
#$motif=<STDIN>;
$motif=MOTIF($motif);

$counter=0;

my $x=length $motif;
for(my $m=$0;$m<length $protein;$m+=$x)
{$repeat= substr($protein,$m,$x);
              if($repeat eq $motif){++$counter;}}
my @pro=();
if ($protein=~ /$motif/)
   {


if($counter>1){print"
\n*****************************ATTENTION**************************
**************\n";
print " the motif $motif  exists in this protein  $counter times the
first one is chosen!! \n\n";
                              print                                "
*****************************ATTENTION****************************
************\n";
                              }
      $protein=~ s/$motif/*/;
      @pro=split('',$protein);
      $aacount=AACOUNT(@pro);
      $proteinlength=$aacount+length $motif;
      print "AMINO ACIDS =$proteinlength\n"; }
      else {print "the motif $motif is not present\n";exit;}
```

```
#counting the string before and after the cleavage motif
#separating the two strings; and print the 50 A.A before and after
the motif
$counter=0;
 for($i=0;$i<scalar @pro;$i++)
      {if ($pro[$i] eq '*'){$counter1=$counter;$counter=0;}
          else{$counter++;}}
$counter2=$counter;

@S1=();@S2=();@A=();@B=();my  $AA1='';my  $AA2='';my  $AAmotif='';my
$totalAA=''; my $totalAAcateg='';
for($i=0;$i<$counter1;$i++){push(@S1,$pro[$i])}
for ($i=$counter1+1;$i<scalar @pro;$i++){push(@S2,$pro[$i])}
@revS1=reverse @S1;

for($i=0;$i<50;$i++){push(@A,$revS1[$i]);@revA=reverse @A;}
for($i=0;$i<50;$i++){push(@B,$S2[$i])}
print"\n";print @revA;print " $motif ";print @B;print"\n";

$first50=join('',@revA);
$after50=join('',@B);
$AA1=AAconverter($first50);
$AA2=AAconverter($after50);
$AAmotif=AAconverter($motif);

print $AA1;print" $AAmotif ";print $AA2;print"\n";
$totalAA=$first50.$motif.$after50;
$totalAAcateg=$AA1.$AAmotif.$AA2;

my @val1; my @val2;
%h1=AAcategoryCOUNT($AA1);
@keys1=keys %h1;
@val1=values %h1;
for ($k=0;$k<5;$k++){
 print  "$keys1[$k]\t$val1[$k]\n"; }

%h2=AAcategoryCOUNT($AA2);
@keys2=keys %h2;
@val2=values %h2;
for ($k=0;$k<5;$k++){
  print  "$keys2[$k]\t$val2[$k]\n"; }

my  $countA1;my  $countB1;my  $countP1;my  $countN1;my  $countH1;my
$total1;


my  $countA2;my  $countB2;my  $countP2;my  $countN2;my  $countH2;my
$total2;
my   $countA;my   $countB;my   $countP;my   $countN;my   $countH;my
$totalCOUNT;
```

```
my $Aper; my $Bper; my $Pper; my $Nper; my $Hper;
my $A1; my $B1; my $P1; my $N1; my $H1;
my $A2; my $B2; my $P2; my $N2; my $H2;
$countA1= $val1[0];
$countB1= $val1[1];
$countP1= $val1[2];
$countN1= $val1[3];
$countH1= $val1[4];
$countA2=  $val2[0];
$countB2=  $val2[1];
$countP2=  $val2[2];
$countN2=  $val2[3];
$countH2=  $val2[4];
$countA= $val1[0]+ $val2[0];
$countB= $val1[1]+ $val2[1];
$countP= $val1[2]+ $val2[2];
$countN= $val1[3]+ $val2[3];
$countH= $val1[4]+ $val2[4];
$total1=$val1[0]+$val1[1]+$val1[2]+$val1[3];
$total2=$val2[0]+$val2[1]+$val2[2]+$val2[3];
$totalCOUNT=$countA+$countB+$countP+$countN;

$A1=$countA1/$total1*100;
$B1=$countB1/$total1*100;
$P1=$countP1/$total1*100;
$N1=$countN1/$total1*100;
$H1=$countH1/$total1*100;
$A2=$countA2/$total2*100;
$B2=$countB2/$total2*100;
$P2=$countP2/$total2*100;
$N2=$countN2/$total2*100;
$H2=$countH2/$total2*100;
$Aper=$countA/$totalCOUNT*100;
$Bper=$countB/$totalCOUNT*100;
$Pper=$countP/$totalCOUNT*100;
$Nper=$countN/$totalCOUNT*100;
$Hper=$countH/$totalCOUNT*100;

$z=$z+1;

print"____________________________________________________________\n
";

#print the database to OUTFILE.XLS

print OUTFILE   "$proteinFILENAME\t";
print OUTFILE   "$proteinlength\t";
print OUTFILE   "$motif\t";
print OUTFILE   "$first50\t";
print OUTFILE   "$AA1\t";
print OUTFILE   "$A1\t";
print OUTFILE   "$B1\t";
```

```
print OUTFILE   "$P1\t";
print OUTFILE   "$N1\t";
print OUTFILE   "$H1\t";
print OUTFILE   "$after50\t";
print OUTFILE   "$AA2\t";
print OUTFILE   "$A2\t";


print OUTFILE   "$B2\t";
print OUTFILE   "$P2\t";
print OUTFILE   "$N2\t";
print OUTFILE   "$H2\t";
print OUTFILE   "$totalAA\t";
print OUTFILE   "$totalAAcateg\t";
print OUTFILE   "$Aper\t";
print OUTFILE   "$Bper\t";
print OUTFILE   "$Pper\t";
print OUTFILE   "$Nper\t";
print OUTFILE   "$Hper\n";

  }

#until($k==2);
 until(@fileName[$z]=~ /^\s*$/ );

close (OUTFILE);

$f=<STDIN>;
exit;
```

aminoacids.pl

```
#! /user/bin/perl

use submodules

#ouput file
$outputfile="AM";
unless(open(AM,">$outputfile.xls")){print    "cannot    open    file\
"$outputfile\"to write to !!\n\n";exit;}

#print results to the outputfile "am.xls"
```

```
print AM "D\t";
print AM "E\t";
print AM "H\t";
print AM "K\t";
print AM "R\t";
print AM "N\t";
print AM "Q\t";
print AM "S\t";
print AM "T\t";
print AM "Y\t";
print AM "A\t";
print AM "L\t";
print AM "P\t";
print AM "M\t";
print AM "G\t";
print AM "V\t";
print AM "I\t";
print AM "F\t";
print AM "W\t";
print AM "C\n";

print "please enter the file name to count!:\n";
$file=<STDIN>;
chomp $file;
unless (open (PROTEIN,$file)){print "can not open the file
$file\n";exit;}
@file=<PROTEIN>;
close PROTEIN;

my @protein=INPUTFILE($file);

foreach $line(@protein){

$line=~ s/\s//g;

@PROTLINE=split('',$line);
$count_A=0;$count_C=0;$count_D=0;$count_E=0;$count_F=0;$count_G=0;
$count_H=0;$count_I=0;$count_K=0;$count_L=0;$count_M=0;$count_N=0;
$count_P=0;$count_Q=0;$count_R=0;$count_S=0;$count_T=0;
$count_V=0;$count_W=0;$count_Y=0;$count_error=0;

foreach $base (@PROTLINE)
{if ($base eq 'A'){++$count_A;}
 elsif($base eq 'C'){++$count_C;}
elsif($base eq 'S'){++$count_S;}
elsif($base eq 'D'){++$count_D;}
elsif($base eq 'T'){++$count_T;}
elsif($base eq 'E'){++$count_E;}
elsif($base eq 'F'){++$count_F;}
elsif($base eq 'V'){++$count_V;}
elsif($base eq 'G'){++$count_G;}
```

```
elsif($base eq 'W'){++$count_W;}
elsif($base eq 'H'){++$count_H;}
elsif($base eq 'I'){++$count_I;}
elsif($base eq 'Y'){++$count_Y;}
elsif($base eq 'K'){++$count_K;}
elsif($base eq 'L'){++$count_L;}
elsif($base eq 'M'){++$count_M;}
elsif($base eq 'N'){++$count_N;}
elsif($base eq 'P'){++$count_P;}
elsif($base eq 'Q'){++$count_Q;}
elsif($base eq 'R'){++$count_R;}
else { ++$count_error;} }

print AM "$count_D\t";
print AM "$count_E\t";
print AM "$count_H\t";
print AM "$count_K\t";
print AM "$count_R\t";
print AM "$count_N\t";
print AM "$count_Q\t";
print AM "$count_S\t";
print AM "$count_T\t";
print AM "$count_Y\t";
print AM "$count_A\t";
print AM "$count_L\t";
print AM "$count_P\t";
print AM "$count_M\t";
print AM "$count_G\t";
print AM "$count_V\t";
print AM "$count_I\t";


print AM "$count_F\t";
print AM "$count_W\t";
print AM "$count_C\n";

print "  $count_D\n";
      }

print "\n$count_error\n";

close (AM);
exit;
```

chemicalgroups.pl

```
#!/usr/bin/perl
use submodules;
```

```
#ouput file
$outputfile="outfile";
unless(open(OUTFILE,">$outputfile.xls")){print      "cannot      open
file\"$outputfile\"to write to !!\n\n";exit;}

#print results to the outputfile "outfile.xls"
print OUTFILE "S.P NUM.\t";
print OUTFILE "PROTEIN LENGTH\t";
print OUTFILE "CLEAVAGE SITE\t";

print OUTFILE "Amino acids before C.S(up to 5)\t";
print OUTFILE "A.A properties\t";
print OUTFILE "#A1\t";
print OUTFILE "#B1\t";
print OUTFILE "#P1\t";
print OUTFILE "#N1\t";
print OUTFILE "#H1\t";

print OUTFILE "Amino acids after C.S(up to 5)\t";
print OUTFILE "A.A properties\t";
print OUTFILE "#A2\t";
print OUTFILE "#B2\t";
print OUTFILE "#P2\t";
print OUTFILE "#N2\t";
print OUTFILE "#H\n";

#print OUTFILE "TOTAL Amino acids(up to 14)\t";
#print OUTFILE "TOTAL A.A properties\t";
#print OUTFILE "#Atotal\t";
#print OUTFILE "#Btotal\t";
#print OUTFILE "#Ptotal\t";
#print OUTFILE "#N\t";
#print OUTFILE "#H-Philic\n";

#input the protein file, remove fasta format and intialize @protein;
and print the protein file
print "please enter the file name of the SWISSPROT IDs:\n";
$mainFile=<STDIN>;
chomp $mainFile;


unless  (open  (SWISSID,$mainFile)){print  "can  not  open  the  file
$mainFile\n";exit;}
@fileName=<SWISSID>;
close SWISSID;


print "please enter the file name of the motifs:\n";
$motiFile=<STDIN>;
chomp $motiFile;
unless  (open  (MOTIFS,$motiFile)){print  "can  not  open  the  file
$motiFile\n";exit;}
```

```
@motifName=<MOTIFS>;
close MOTIFS;

my $z=0;


do {

$a=$fileName[$z];chomp $a; $a=~ s/\s//g;
my $notes='';
my @protein=INPUTFILE($a);
my $protein='';
foreach            $line(@protein){if($line=~/^>/){            $notes
.=$line;next;}else{ $protein .=$line;}}

$proteinFILENAME=FILENAME($notes);
print"$proteinFILENAME\n";

#remove white spaces
$protein=~ s/\s//g;

#enter the motif to cutoff andintialize $motif
my $motif=$motifName[$z];
#print "please enter the motif for $proteinFILENAME :\n";
#$motif=<STDIN>;
$motif=MOTIF($motif);

$counter=0;

my $x=length $motif;
for(my $m=$0;$m<length $protein;$m+=$x)
{$repeat= substr($protein,$m,$x);
  if($repeat eq $motif){++$counter;}}
my @pro=();
if ($protein=~ /$motif/)
   {
if($counter>1){print"
\n*****************************ATTENTION**************************
**************\n";
 print " the motif $motif  exists in this protein  $counter times
the first one is chosen!! \n\n";
                              print                                 "
****************************ATTENTION***************************
************\n";
                              }
      $protein=~ s/$motif/*/;
      @pro=split('',$protein);
      $aacount=AACOUNT(@pro);
      $proteinlength=$aacount+length $motif;}


#     print "AMINO ACIDS =$proteinlength\n"; }
```

```
      else {print "the motif $motif is not present\n";exit;}


#counting the string before and after the cleavage motif
#separating the two strings; and print the 5 A.A before and after
the #motif
$counter=0;
 for($i=0;$i<scalar @pro;$i++)
      {if ($pro[$i] eq '*'){$counter1=$counter;$counter=0;}
          else{$counter++;}}
$counter2=$counter;

@S1=();@S2=();@A=();@B=();
my    $AA1='';my    $AA2='';my    $AAmotif='';my    $totalAA='';    my
$totalAAcateg='';

for($i=0;$i<$counter1;$i++){push(@S1,$pro[$i])}
for ($i=$counter1+1;$i<scalar @pro;$i++){push(@S2,$pro[$i])}
@revS1=reverse @S1;

for($i=0;$i<5;$i++){push(@A,$revS1[$i]);@revA=reverse @A;}
for($i=0;$i<5;$i++){push(@B,$S2[$i])}
#print"\n";print @revA;print " $motif ";print @B;print"\n";

$first5=join('',@revA);
$after5=join('',@B);
$AA1=AAconverter($first5);
$AA2=AAconverter($after5);
$AAmotif=AAconverter($motif);

#print $AA1;print" $AAmotif ";print $AA2;print"\n";
$totalAA=$first5.$motif.$after5;
$totalAAcateg=$AA1.$AAmotif.$AA2;

my @val1; my @val2;
%h1=AAcategoryCOUNT($AA1);
@keys1=keys %h1;
@val1=values %h1;
for ($k=0;$k<5;$k++){}
# print  "$keys1[$k]\t$val1[$k]\n"; }

%h2=AAcategoryCOUNT($AA2);
@keys2=keys %h2;
@val2=values %h2;
for ($k=0;$k<5;$k++){}
#  print  "$keys2[$k]\t$val2[$k]\n"; }

my  $countA1;my  $countB1;my  $countP1;my  $countN1;my  $countH1;my
$total1;
my  $countA2;my  $countB2;my  $countP2;my  $countN2;my  $countH2;my
$total2;
```

```
my   $countA;my   $countB;my   $countP;my   $countN;my   $countH;my
$totalCOUNT;
my $Aper; my $Bper; my $Pper; my $Nper; my $Hper;
my $A1; my $B1; my $P1; my $N1; my $H1;
my $A2; my $B2; my $P2; my $N2; my $H2;
$countA1= $val1[0];
$countB1= $val1[1];
$countP1= $val1[2];
$countN1= $val1[3];
$countH1= $val1[4];
$countA2=  $val2[0];



$countB2=  $val2[1];
$countP2=  $val2[2];
$countN2=  $val2[3];
$countH2=  $val2[4];
$countA= $val1[0]+ $val2[0];
$countB= $val1[1]+ $val2[1];
$countP= $val1[2]+ $val2[2];
$countN= $val1[3]+ $val2[3];
$countH= $val1[4]+ $val2[4];
$total1=$val1[0]+$val1[1]+$val1[2]+$val1[3];
$total2=$val2[0]+$val2[1]+$val2[2]+$val2[3];
$totalCOUNT=$countA+$countB+$countP+$countN;

$A1=$countA1;
$B1=$countB1;
$P1=$countP1;
$N1=$countN1;
$H1=$countH1;
$A2=$countA2;
$B2=$countB2;
$P2=$countP2;
$N2=$countN2;
$H2=$countH2;
$Aper=$countA;
$Bper=$countB;
$Pper=$countP;
$Nper=$countN;
$Hper=$countH;

$z=$z+1;

#print"_____________________________________________________\n";

#print the database to OUTFILE.XLS

print OUTFILE   "$proteinFILENAME\t";
print OUTFILE   "$proteinlength\t";
print OUTFILE   "$motif\t";
```

```
print OUTFILE   "$first5\t";
print OUTFILE   "$AA1\t";
print OUTFILE   "$A1\t";
print OUTFILE   "$B1\t";
print OUTFILE   "$P1\t";
print OUTFILE   "$N1\t";
print OUTFILE   "$H1\t";
print OUTFILE   "$after5\t";
print OUTFILE   "$AA2\t";
print OUTFILE   "$A2\t";
print OUTFILE   "$B2\t";
print OUTFILE   "$P2\t";
print OUTFILE   "$N2\t";
print OUTFILE   "$H2\n";
#print OUTFILE   "$totalAA\t";
#print OUTFILE   "$totalAAcateg\t";
#print OUTFILE   "$Aper\t";
#print OUTFILE   "$Bper\t";
#print OUTFILE   "$Pper\t";
#print OUTFILE   "$Nper\t";


#print OUTFILE   "$Hper\n";

  }


 until(@fileName[$z]=~ /^\s*$/ );

close (OUTFILE);


exit;
```

transpose.pl

```
#!/usr/bin/perl


use submodules;
#ouput HTML file (unc.html)
$outputfile="unc";
unless(open(UNC,">$outputfile.html")){print "cannot open file\
"$outputfile\"to write to !!\n\n";exit;}
```

```
print "please enter the TEXT file :\n";
$file=<STDIN>;
chomp $file;
unless (open (UNCP,$file)){print "can not open the file
$file\n";exit;}
@uncp=<UNCP>;
close UNCP;

$l='';$newstring='';
for($i=0;$i<15;$i++){
$y=scalar @uncp;
 print "$y\n";
$c=0;
# for($j=0;$j<scalar @uncp;$j++)
   foreach $line(@uncp){
$line=~ s/\s//g;
@line=split('',$line);

$l=shift @line;
$transLine.=$l;
$Line=join('',@line);
@newstring[$c]=$Line;
$c++;}

print"$transLine\n";

print UNC"$transLine\n";
@final[$i]=$transLine;
@uncp=@newstring;
$transLine='';}
#print "@final\n\n";

close (UNC);
exite;
```

List of Figures

List of Tables

Abstract

Caspases are responsible for all the morphological and biochemical changes that end in apoptosis. Their apoptotic function is done in a cascade, which includes the cleavage of many different substrates. Some of these substrates are converted from inactive form to active form through this cleavage while others are considered as controllers and inducers for reactions and processes included in apoptosis. In general, caspases shared a remarkable feature in their cleavage process which is the specificity to Aspartic acid (D) in the P1 subsite. All caspases substrates are cleaved after the amino acid “D”. Caspase-3 (interleukin-1beta converting enzyme/CED-3) is the main executer caspase that is responsible for the cleavage of many key proteins. Up-to-date; caspase-3 has more than 150 known substrates experimentally.

A tool for predicting a substrate cleavage site/s becomes a need for most of researchers who work in apoptotic and cancer field and other related fields. The few available bioinformatics tools have very low accuracy.

The present study introduces a new bioinformatics tool to predict the cleavage site of caspase-3 substrates. CAT3 “Caspase-3 Tool” is specific only to caspase-3. This specificity makes CAT3 a powerful tool with a higher accuracy compared to other available related tools. This tool comes to predict the undetermined cleavage sites of many proteins defined as caspase-3 substrates. Also it can predict other substrates that still not considered as a caspase-3 substrate. CAT3 successfully predicts 23 out of 27 cleavage sites (about 85.2%) of randomly chosen substrates that their cleavage sites are experimentally determined.