A HYBRID NEURAL NETWORK – HIDDEN MARKOV MODEL – FUZZY LOGIC METHOD FOR PROTEIN CLASSIFICATION

by

MARTIN CHEW WOOI KEAT

Thesis submitted in fulfillment of the
requirements for the degree
of Doctor of Philosophy

AUGUST 2007

ACKNOWLEDGEMENTS

To Dr. Rosni Abdullah and Dr. Rosalina Abdul Salam, thank you for everything.

TABLE OF CONTENTS

CHAPTER THREE : LITERATURE REVIEW

CHAPTER FOUR : PERCEPTRON-BASED HYBRID PROTEIN CLASSIFIER

CHAPTER FIVE : WEIGHTLESS NETWORK BASED HYBRID CLASSIFIER

CHAPTER SIX : MULTI-PERCEPTRON NETWORK BASED HYBRID CLASSIFIER

CHAPTER SEVEN : VALIDATION

CHAPTER EIGHT : CONCLUSION

LIST OF TABLES

LIST OF FIGURES

KAEDAH HIBRID RANGKAIAN NEURAL – MODEL "HIDDEN MARKOV" - LOGIK KABUR UNTUK KLASIFIKASI PROTEIN

ABSTRAK

Tujuan tesis ini adalah untuk menyiasat cara-cara menyokong hasrat negara untuk membangunkan industri bioteknologi tempatan. Bioteknologi adalah berdasarkan ciptaan protein baru untuk tujuan industri atau perubatan. Klasifikasi protein merupakan langkah pertama dalam proses ciptaan bahan protein baru. Maka, kami menyiasat kaedah-kaedah pengkomputeran yang digunakan untuk mengklasifikasikan protein. Kelebihan and kelemahan pelbagai kaedah klasifikasi protein telah dikaji. Daripada kajian ini, kami menggabungkan rangkaian neural, Model "Hidden Markov" dan logik kabur untuk membentuk satu sistem klasifikasi hibrid. Kaedah hibrid ini mengambil kira kelebihan-kelebihan kaedah-kaedah sebelumnya tetapi bukan kelemahan-kelemahannya. Kami juga menyiasat bagaimana prestasi dan ketepatan kaedah klasifikasi protein ini dapat dipertingkatkan. Kaedah pemprosesan selari dan rangkaian tanpa pemberat telah disiasat untuk meningkatkan prestasi, manakala kaedah rangkaian multiperseptron dan logik kabur telah digunakan untuk meningkatkan ketepatan. Ketepatan keputusan yang diperoleh dari hibrid kami didapati setaraf dengan ketepatan keputusan kaedah-kaedah sebelumnya. Hibrid kami juga berkebolehan mengendalikan saiz data yang lebih besar untuk memenuhi keperluan komersial. Keputusan kami setaraf dengan keputusan kaedah-kaedah sebelumnya kerana kami mengelak daripada menggunakan teknik penjajaran berbilang jujukan and juga kerana kami menggunakan logik kabur untuk pembetulan ralat. Hibrid kami berkebolehan mengendalikan saiz data yang besar kerana kami menggunakan perseptron yang mudah dan juga rangkaian tanpa pemberat yang hanya memerlukan satu lintasan pada data latihan.

A HYBRID NEURAL NETWORK – HIDDEN MARKOV MODEL – FUZZY LOGIC METHOD FOR PROTEIN CLASSIFICATION

ABSTRACT

The purpose of this thesis is to investigate how to support the national drive to have a biotechnology industry. Biotechnology is based on the creation of new proteins for either industrial or medicinal purposes. Protein classification is the first step in the creation of new protein materials or drugs. Therefore, we investigated the computational methods used to classify proteins. The strengths and weaknesses of various protein classification methods were analyzed. A hybrid classification method based on a combination of neural networks, Hidden Markov Models and fuzzy logic was developed. This hybrid leveraged on the strengths of previous methods, while avoiding their pitfalls. We investigated ways to further improve this hybrid in terms of performance and accuracy. To improve performance, we investigated the use of parallel processing and weightless networks. To improve accuracy, we investigated the use of multi-perceptron networks and also fuzzy logic. Our method obtained results comparably with previous methods, and is also scalable in order to be able to meet commercial requirements. We obtained comparable results because we avoided the fundamental flaw that we identified in previous protein classification methods, which is the multiple sequence alignment flaw, and also because we applied fuzzy logic in a novel manner for error correction purposes. Our method was scalable because we relied on simple perceptrons and also weightless networks that require only one pass at training data.

CHAPTER ONE
INTRODUCTION

1.0 Definition

Fredj Tekaia at the Pasteur Institute of France offered this definition of bioinformatics: "*The mathematical, statistical and computing methods that aim to solve biological problems using DNA and amino acid sequences and related information*". One of the biological problems to solve is *protein classification.* The protein classification problem can be stated formally as follows. Given an unlabeled protein sequence S and a known superfamily F. We want to determine whether or not S belongs to F. (We refer to F as the target class and the set of sequences not in F as the non-target class). In general, a superfamily is a group of proteins that share similarity in structure and function. If the unlabeled sequence S is detected to belong to F, then one can infer the structure and function of S. This process is important in many aspects of computational biology. For example, in drug discovery, if sequence S is obtained from some disease X and it is determined that S belongs to the superfamily F, then one may try a combination of the existing drugs for F to treat the disease X. Furthermore, given an unlabeled sequence S, how do we verify that its function is new (i.e. yet to be discovered). If S is found to be incompatible with any known protein families, we propose that S is potentially the first member of a new protein family (i.e. a new type of protein).

1.1 Nature of Protein Sequences

Protein sequences which share a similar function are grouped into a family. Same family sequences are called homologous sequences. However, the similarity of the same family sequences is buried within a high level of "noise". For example, homology has been detected among sequences with as little as 10% sequence identity. Furthermore, sequences with up to 25% sequence identity (which would be

assumed to be homologous), has been found to be non-homologous (Rost, 1999b). In addition to the problem of noise, there is also the problem of scale. The Pfam (http://www.sanger.ac.uk/Software/Pfam/) protein family database had 7973 families as of August 2005. Each family could contain a few dozen to a few hundred sequences.

1.2 General Concept of Protein Classifiers.

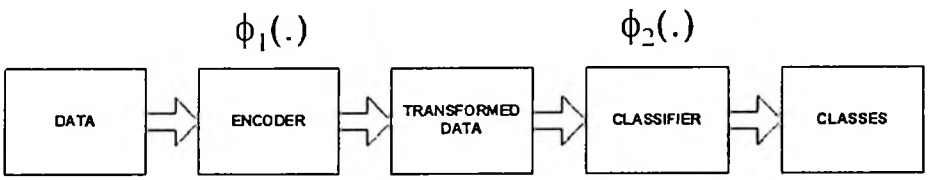

Figure 1.1: General Concept of Protein Classifiers

In general, protein classifiers operate as per the model specified in Figure 1.1. Data (i.e. protein sequences) are processed by one or more encoders. An encoder is essentially a function to transform data from one form of representation to another. This transformation is necessary in order to extract embedded features in the untransformed data, as well as to generate a representation of the data in a format acceptable by the classifier. The transformed data (i.e. the encoded data) is then processed by the classifier. The classifier is also essentially a function to separate the transformed data into two or more classes.

Protein classification systems have to cope with explicit and implicit similarities. Implicit similarities are similarities which have to be brought to surface with the application of some form of transformation or projection (i.e. encoding method). The "signals" generated from the transformation or projection are then analyzed. As such, a protein classification system is essentially a signal analyzer.

A typical protein classification approach would involve identifying a criterion to distinguish one protein family from another. For example, attempting to discover a particular motif (i.e. recurring sub-sequence within a protein sequence) common across members of a given protein family, and using that motif as the "fingerprint" to identify yet-to-be-discovered members of that family (Eddy, 2004). Such an approach would be effective as long as at least one motif exists in a given protein family, and that this motif is unique to that family. If no motif exists, or the chosen motif appears in numerous other protein families as well, then this protein classification solution could be rendered ineffective.

Knowledge about the nature of proteins as a whole, or about a particular protein, evolves over a period of time. For example, a 10% sequence similarity threshold may have been set as the boundary for membership in a protein family. However, a discovery that there exists non-homologous sequences with 25% sequence similarity would render the 10% sequence identify threshold outdated.

As such, any protein classification system must be able to incorporate new discoveries into its future decision making process. Any new discoveries must also be propagated throughout past classifications in order to determine its implications. Furthermore, the protein classification system must also be able to handle the noise and scale of the protein classification problem. We designed the objectives of our thesis in order to support these requirements.

1.3 Thesis Objective and Constraints

Our objective is to develop a protein classification system which is:

1. modular
2. accurate
3. scalable

4. *novel but comparatively simpler*

It has to be modular in order to be able to support various and future encoding schemes that incorporate current and new discoveries about protein relationships. It has to have an accuracy rate comparable or near-comparable to existing classifiers to be able to handle the noise associated with protein classification. It has to be scalable in order to be able to handle the scale of the protein classification problem. *Last but not least, our system must deliver these benefits while being novel and comparatively simpler than existing protein classification systems.*

Our first constraint concerns the principles of protein relationships. The actual principles governing protein relationships are OUTSIDE of our scope. For purposes of experimentation, we will use a toy protein sequence encoding scheme called 2-gram encoding to stand-in for actual principles of protein relationships. More details on 2-gram encoding will be provided in subsequent chapters. Protein interactions are governed by principles of biochemistry. Toy encoding methods such as 2-gram encoding were NOT derived from the principles of biochemistry, but from the principles of statistics. However, 2-gram encoding is sufficient to analyze protein sequences with a *particular degree* of explicit similarity and a *particular aspect* of implicit similarity. Actual principles of protein relationships has to be discovered by molecular biologists, and communicated to the bioinformatician to be represented as protein sequence encoding methods. Our protein classification system may serve as a foundation on which to build or explore various protein sequence encoding methods which act as classification solutions. An analogy to the relationship between our protein classification system and the field of protein classification is the relationship between the spreadsheet concept and the field of accountancy. The spreadsheet is the system to build or explore for various accounting solutions. The spreadsheet itself is not the accounting solution, but the foundation designed to support any accounting solution.

Furthermore, the current state of knowledge concerning protein interactions are based on current, known principles of protein family relationships. In the event of the discovery of a new principle of relationship, existing knowledge has to be systematically revised in order to deduce any new implications as a result of the newly discovered principle.

Our second constraint concerns the experimental data. There are thousands of protein families, but for the purpose of this thesis, we limit our sampling to the families used by prior researchers covered in our literature review, in addition to a sample of randomly chosen families not covered by the aforementioned researchers. This is to determine the effectiveness of our system against the systems covered in our literature review.

1.4 Thesis Research Methodology

We began by doing a generic background study on proteins. We focused on:

- the relationship between genetics and proteins
- the concept of protein evolution

We focused on these two areas because proteins are 'generated' by genes and their functionality is believed to be governed by molecular evolution.

From this, we proceeded to a generic study on the earliest method used to analyze protein functionality – the method of multiple sequence alignment. Multiple sequence alignment was pioneered by Dr. Margaret Oakley Dayhoff (1925-1983). She is credited today as a founder of the field of Bioinformatics (Dayhoff, 2007). The key to multiple sequence alignment is the scoring matrix used to perform the alignment. We looked into:

- how a scoring matrix is derived
- what are the weaknesses of the scoring matrix approach

We focused on these two areas because if the very method used to derive a scoring matrix is also the root cause of its weaknesses, then multiple sequence alignment has a fundamental flaw.

The alternative to multiple sequence alignment is artificial intelligence methods. We listed down the artificial intelligence methods used in bioinformatics and did a literature review to provide an introduction to the method in the context of bioinformatics, as well as to ascertain each method's strengths and weaknesses. This was required to enable us to identify the more promising artificial intelligence methods that will be used as a basis to achieve our thesis objective. Once the more promising artificial intelligence methods have been identified, our literature review continued with hybridization strategies for the methods we identified. Hybridization is used to compensate the weakness of one method with the strength of another. Once we have identified our hybridization strategy to support our thesis objective, we proceeded to design our system.

The details of our system design and the rationale behind the design decisions are given in subsequent chapters of this thesis. To validate the soundness of our design, we tested it against protein families used in prior literature. To validate that we have met our thesis objective, we conducted a qualitative and quantitative analysis on the performance of our system.

1.5 Thesis Roadmap and Outline

Figure 1.2 gives an overview of this thesis. The topics in the shaded boxes (refer to Figure 1.2) will be the focus of this thesis.

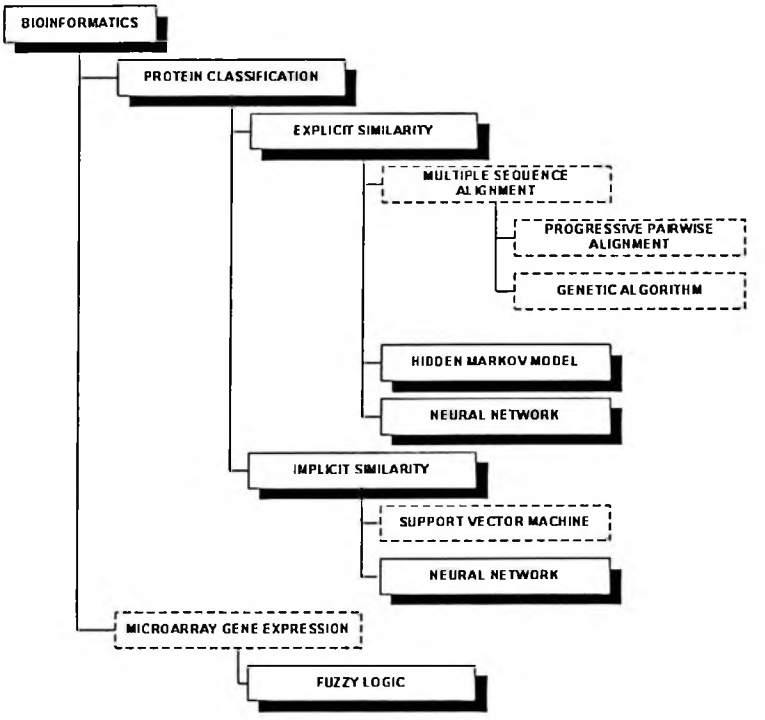

Figure 1.2: "Roadmap" of the Thesis

The thesis begins with the definition of bioinformatics. This sets the boundary for the research. The scope is then narrowed down to protein classification, its definition and purpose being given. The nature of protein sequences, and the classification problems arising from this nature is discussed. The objectives of this thesis were then formulated in order to cope with the problems.

Chapter 2 provides a background to the field of bioinformatics. Protein sequences are generated based on patterns found in DNA sequences. Proteins are grouped into families, each family corresponding to a particular structure or function. Protein sequences of the same family need not, and does not share 100% identity. One possible cause of this divergence is evolution. The protein structure of a parent specie changes as that parent evolves into other species. Multiple sequence alignment was devised in order to re-build the relationship tree among protein sequences of the

same family, as well as to detect potential new members of that family. However, this conjecture was challenged when the relationship among aligned sequences did not correspond to the physio-morphological relationship observed among the host species of those sequences. One potential cause of this disconnect could be because the fitness of a multiple sequence alignment is determined with respect to statistically derived scoring tables. These scoring tables determine the points scored for a match, and the points deducted for a mismatch. Sample alignments were used to derive the scoring tables, and these samples were based on explicit similarity among the sequences. Implicit similarities among sequences in a family were ignored. Therefore, alternative methods of protein sequence analysis and classification are required. We decided to use of artificial intelligence methods because these methods are especially suitable for handling noisy and multi-dimensional data.

Chapter 3 provides a review of prior applications of artificial intelligence techniques in the field of protein classification. We reviewed 5 methods: fuzzy logic, genetic algorithms, Hidden Markov Models, Neural Networks, and Support Vector Machines. Fuzzy logic was never applied in the field of protein classification before. Genetic algorithms were used for multiple protein sequence alignment. Hidden Markov Models were used to model and detect motifs (i.e. recurring sub-sequences) in protein sequences. Neural networks were used in conjunction with regular expression functions to abstract the explicit and implicit similarities of a particular protein family. Support vector machines were also used to classify proteins, but required more implementation effort compared to neural networks in return for similar benefits in terms of accuracy. As a result, we dismissed genetic algorithms and support vector machines, and focused on Hidden Markov Models and neural networks. We also decided to explore the application of fuzzy logic for protein classification, in conjunction with Hidden Markov Models and neural networks. We explored various hybridization strategies of the three methods (i.e. fuzzy logic, Hidden Markov Models and neural

networks) in order to address the shortcomings of the prior protein classification methods reviewed.

Chapter 4 provides a description of our **first contribution**: a perceptron array based hybrid protein classifier. Hidden Markov Models were used to bias the network if any motifs are detected, and fuzzy logic is used to post-process the results from the array for better classification accuracy. We have experimental results on 8 protein families, together with additional experimental results on *Bacillus Subtilis* proteins. We also attempted to improve the performance of the perceptron array with parallel processing. Chapter 5 provides a description of our **second contribution**: a weightless network array based hybrid protein classifier. A weightless network is typically used for image processing, but we adapted it for the purpose of protein classification. A weightless network only requires one pass at a training dataset, unlike a perceptron, which has to iterate at a training dataset until a minima is reached. Fuzzy logic was used to post-process the results of both arrays to obtain better accuracy. Chapter 6 provides a description of our **third contribution**: a multi-perceptron array based hybrid protein classifier. Perceptrons are only capable of linear separation. Multi-perceptrons are used to abstract non-linear domains, in place of backpropagation networks. This is because, unlike backpropagation networks, multi-perceptron network are not blackboxes. We adapted the multi-perceptron concept for the purpose of protein classification. This is especially important when tweaks have to be applied to a network stuck in local minima.

Chapter 7 discusses the validation of the contributions of our proposed system. The concluding chapter describes how our protein classification system could be expanded to the field of DNA classification (i.e. gene finding). In the Appendix, a list of papers published is provided. In the following chapter, we proceed to a background on bioinformatics.

CHAPTER TWO
BACKGROUND

2.0 Introduction

This chapter describes the relationship between genetics and proteins, especially on how genetics determine the generation of proteins. The nature of proteins are then elaborated on, especially the concept of protein homology (proteins with similar structure or function). In order to analyze protein homology, multiple sequence alignment is widely used. Multiple sequence alignment tools rely on scoring schemes based on the concept of protein evolution. However, the assertion that protein homology is due to protein evolution is challenged when certain conclusions of multiple sequence alignment do not correspond to fossil or morphological conclusion. As a result, alternative approaches to protein homology analysis are deemed necessary, especially approaches capable of handling implicit (i.e. "hidden") similarities among protein sequences.

2.1 From Atoms to Genetics

The basic building blocks of chemistry are the atomic elements. An example of an atomic element is hydrogen, which is the lightest of all the elements. Another example of an atomic element is oxygen, which is 16 times heavier than hydrogen.

These atomic elements are joined together in various combinations and permutations to form molecules. A particular combination (in terms of the composition of the various atomic elements) will create a particular type of molecule. A particular *permutation of that combination* (in terms of the spatial arrangement of those atomic elements) will create a variation of that particular type of molecule.

Different molecules could further combine with other molecules to create macromolecules. A nucleotide is an example of a type of macromolecule. A nucleotide is a combination of three other molecule types (Toole, et. al. 1991):

- "phosphoric acid" molecule type (made of elements of hydrogen, oxygen & phosphor)
- "carbohydrate" molecule type (made of elements of hydrogen, oxygen & carbon)
- "organic" molecule type (made of elements of hydrogen, oxygen, carbon & nitrogen)

A genetic strand is a chain of macromolecules of the nucleotide type. There are four possible permutations of nucleotide that could make up the chain. The four permutations are named as (Toole, et. al. 1991):

- Adenine
- Cytosine
- Guanine
- Thymine

The four nucleotide permutations are represented by their first alphabets (A, C, G and T). A nucleotide chain (i.e. a genetic strand) is represented by a sequence of alphabets, with the valid alphabets being A, C, G and T. An example of a genetic strand representation is AAACCGGTTT.

Each species has its own unique genetic strand length. For example, the human species has a genetic strand length of at least 3 billion. An Escherichia Coli

bacterium has a sequence size of 4.6 million (Vaisman, 2002). Furthermore, each individual human being has his own unique variation in the composition of the nucleotide chain that makes up the genetic strand. That unique genetic strand variety is present in each cell of that particular individual's body.

A basic cell is a self-contained pool of various molecules and macromolecules. The genetic strand resides in the pool, and acts a master pattern list, from which the construction of macromolecules is based on. Only certain sections of the genetic strand are "active" during certain times. Each nucleotide in the "active" section of the genetic strand attracts a complimentary nucleotide from the pool of molecules. The attracted complimentary nucleotides are chained together. A complimentary strand, mirroring the "active" section of the master strand, is thus generated, and released into the pool. "Active" sections of the genetic strand are called **genes** (Toole, et. al. 1991).

The generated "messenger" strand acts like an assembly line, from which molecules are pulled from the pool, and chained into a macromolecule. Three "messenger" strand nucleotides (i.e. a triplex) are needed to pull one molecule. This triplex is called a **codon**. A sequence of codons represents the information required to build a macromolecule chain. The macromolecule chains being constructed by the "messenger" strand are collectively known as proteins (Toole, et. al. 1991).

Therefore, given a protein sequence, it is possible to deduce the codons, "messenger" strand, and eventually the "active" section of the genetic strand that generated that protein sequence. Please refer to Figure 2.1.

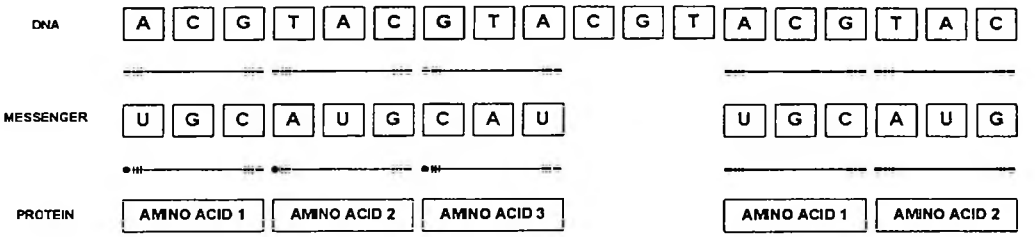

Figure 2.1: From DNA to RNA to Protein.

The complete collection of an organism's various genes is called the organism's **genome**, whereas the complete collection of an organism's various proteins is called the organism's **proteome**. We will discuss about proteins in the next section.

2.2 From Genetics to Proteins

A protein is a chain of amino acid type molecules. An amino acid type molecule is made of elements of hydrogen, oxygen, carbon and nitrogen. In the human system, there are up to 20 different variations of amino acid, with each variation being represented by an alphabet. An example of a protein sequence is AVVKVPLKKF. The genetic strand determines an organism's characteristics by determining what proteins are produced (Toole, et. al. 1991).

A protein macromolecule chain takes a three dimensional shape. Different proteins assume different three dimensional shapes. When a protein assumes a three dimensional shape, it becomes an analogy of a three dimensional jigsaw puzzle piece. Proteins interact with one another as jigsaw puzzle pieces interact with one another. Certain puzzle pieces act to lock on to others – in other words, one piece attaches (i.e. docks) itself on to another piece. When docking occurs between 2 or more proteins, the overall shape is changed, thus disabling the function of the proteins (Shatsky, et. al. 2004).

Proteins are grouped into families. A protein family is a collection of protein sequences that share a similar a three dimensional structure or function. Such protein sequences are said to be **homologous**. There are currently at least 6000 protein families (Bateman, et. al. 2004), and at least 150,000 protein sequences (Huang, et. al. 2000). The top 20 families contain over 2500 sequences (Bateman, et. al. 2004).

The family relationship of a protein is determined after the three dimensional structure of a protein is determine using X-ray analysis. Proteins of the same family tend to have near-similar sequences. *However, it is also possible for sequences with insignificant sequence similarities to belong to the same family* (Pandit, et. al. 2002). Homology can be observed among sequences with as little as 25% similarity across their lengths (Pearson, 2001).

Since it is possible for protein sequences with insignificant similarities to belong to the same family, this makes it difficult to predict secondary and tertiary structure from the primary structure. Protein structure is divided into four categories (Protein structure, 2005):

- Primary : the amino acid sequence of the protein.
- Secondary : divided into "alpha" and "beta"
    - Alpha : sections of the sequence assuming a helix shape.
    - Beta : sections of the sequence in parallel with one another.
- Tertiary : the overall 3-D shape of the protein.
- Quaternary : the structure from the union of more than one protein molecule.

2.3 Protein Evolution

Homologous proteins are assumed to have diverged from a common evolutionary ancestor. For example, if a protein sequence in the yeast fungi is homologous to another protein sequence in the Escherichia Coli bacterium, then it is assumed that those two sequences must have also existed in the primordial organism that gave rise to fungi and bacteria (Pearson, 2001).

Homology among protein sequences from different species is called **orthologous** homology. Protein sequences which are homologous in terms of structure but differs in function is said to exhibit **paralogous** homology.

The concept of homology as a result of evolutionary divergence forms the basis for multiple sequence alignment, in the sense that, homologous protein sequences from different species are deemed to be **meaningfully** aligned, *if the alignment accurately reflects the evolutionary distance of those species.* For example, when aligned with one another, a given protein from a particular mammal should show more similarity with a homologous protein from another mammal, compared to a homologous protein from a reptile (Pearson, 2001).

In Bioinformatics, the argument of homology due to evolution is used as the basis for making deductions or inferences about unknown proteins based on known proteins (Birney, 2001).

2.4 Multiple Sequence Alignment

Alignment means the positioning of one sequence relative to another. The goal of multiple sequence alignment is to highlight what are the *areas of similarity across* three or more given protein sequences. These areas of similarity are called **hot-spots**

(Camp, et. al., 1998). The underlying premise is that given a collection of sequences that share the same molecular structure or function, the areas of similarity across those sequences are possibly responsible for the characteristics of those sequences. An alignment example is given in Figure 2.2:

```
A A B B C D D D E
        B C D D E
      A B B C D E
```

Fig. 2.2: Multiple Alignment Example

Multiple alignment tools aim to generate *optimal alignments*. The alignment example in Figure 2.2 is not optimal. A better alignment is given in Figure 2.3:

```
A A B B C D D D E
      B C D D E
  A B B C D E
```

Fig. 2.3: Optimal Multiple Alignment Example

The fitness of an alignment (i.e. the quality of a particular arrangement over another) is judged based on a given scoring method. For example, a match could be scored +1, while is mismatch is scored -1. However, a different scoring method might score a match as +2, and a mismatch as 0. This would redefine which alignments are optimal, and which are not. The scoring method used by a particular multiple alignment tool is called a **substitution matrix**. The substitution matrix is the foundation on which a given multiple alignment tool determines what is optimal (Pearson, 2001).

2.5 Substitution Matrix

A multiple alignment tool has two components:

- Algorithm for aligning sequences
- Substitution matrix on which the alignment scoring is based.

Substitution matrices are calculated by examining sample blocks of related protein sequences that differ by no more than a given percentage. The protein sequences are related according to their chemical properties, such as the electrostatic charge and the ability to dissolve in water (Pairwise sequence alignment, 2005). The chemical properties of the various amino acids determine their relative replaceability with one another. Figure 2.4 gives a toy example of a substitution matrix with three residues (A, B and C). Figure 2.5 shows the chemical properties of the various amino acids.

	A	B	C
A	+4		
B	+2	+2	
C	-4	-1	+1

Fig. 2.4: Substitution Matrix Example

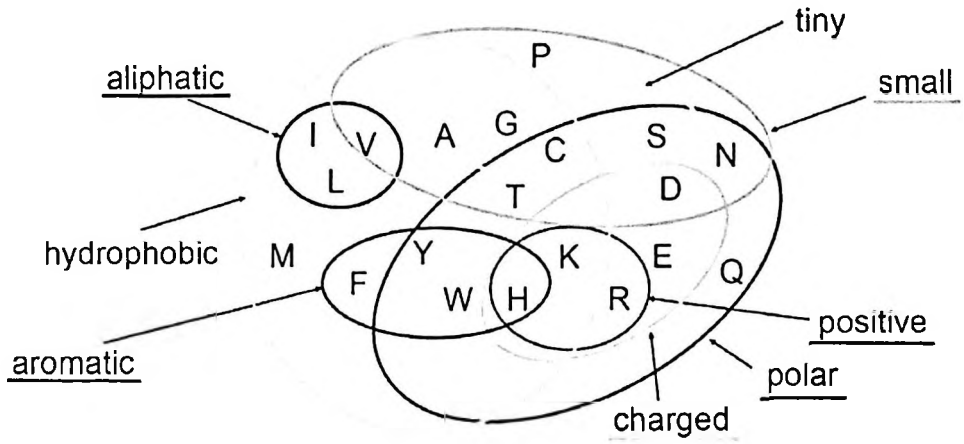

Fig. 2.5: Properties of the Various Amino Acids

If the substitution matrix example given in Figure 2.4 is used to score an alignment, matching an "A" with another "A" scores 4 points, whereas matching a "B" with another "B" scores 2 points. Such differences in scoring weight are due to the predominance of one residue over another in the sample block being examined. Furthermore, mismatching an "A" to a "B" scores 2 points, whereas mismatching an "A" to a "C" gives a penalty of negative 4. Such a situation arises when the sampling block indicates that an "A" would most likely evolve into a "B", but unlikely to evolve into a "C".

2.6 Examples of Multiple Sequence Alignment Tools

There are two types of alignment methods: **global alignment** and **local alignment**. Global alignment means that the sequences are aligned from left to right, with gaps being added in a left to right direction. Local alignment means the common region of highest similarity is first identified between 2 sequences, and then gaps are inserted to build the alignment outwards from that region (Pairwise sequence alignment, 2005).

An example of how gapping could improve an alignment is given in Figures 2.6a and 2.6b:

A A A A B B B C C C C D D D D D D

E E E A B B B C C D D E E E E E E

Fig. 2.6a: Ungapped (6 matches)

A A A A B B B C C C C D D D D D D

E E E A B B B C C - - D D E E E E E E

Figure 2.6b: Gapped (8 matches)

However, if too many artificial gaps are inserted, the overall score of the alignment could be reduced (Altschul, et. al. 1997).

An example of an alignment tool based on the global alignment method is ClustalW (Narayanan et. al. 1999). The approach used by ClustalW to align multiple sequences is also known as the **pairwise progressive** method. For example, if there are four sequences (S1, S2, S3, and S4) to be aligned:

- Step 1: Align S1 to S2 (to obtain S1:S2)
- Step 2: Align S3 to S4 (to obtain S3:S4)
- Step 3: Align S1:S2 to S3:S4 or S3:S4 to S1:S2 (select the optimal)

There could be other permutations to the example given above. Instead of starting by aligning S1 with S2, we could also start by aligning S1 with S3, etc. This problem of permutations makes the problem of multiple sequence alignment **exponential** in nature (Pairwise sequence alignment, 2005).

An example of a multiple sequence alignment tool based on the local alignment method is BLAST (Basic Local Alignment Tool) (Pearson, et. al., 2001). BLAST is used to perform local alignment on a given set of sequences. The substitution matrix used by BLAST is BLOSUM (Block Substitution Matrix). There are several varieties of BLOSUM (e.g. BLOSUM45, BLOSUM52, BLOSUM60, BLOSUM80 and BLOSUM90). BLOSUM45 means that the matrix was built by examining related sequences with at most 45% similarity. BLAST users decide which matrix to use based on their pre-conception of the nature of the sequences they are aligning. For sequences deemed related but highly diverged (i.e. very little similarity), a matrix with a low number is used. For sequences deemed related and still unevolved, a matrix with a high number is used (What is BLAST, 2005).

2.7 Problems with the Concept of Multiple Sequence Alignment

When we align multiple sequences, we are implicitly assuming an evolutionary relationship among the sequences (Pearson, 2001). However, there are instances where conclusions from the field of molecular evolution either do not correspond with the conclusions from the study of fossils, or do not yield logical results. For example:

- Analysis of the cytochrome c protein (needed for cell respiration) of different species showed the chicken to be related more closely to the penguin, than to ducks and pigeons. The turtle (a reptile) was also shown to be more closely related to birds, rather than the rattlesnake (another reptile) (Ayala, 1978).
- The molecular evolution of the cytochrome c protein of 26 angiosperm (flower producing) plant species yields multiple minimal trees. A minimal tree is a tree of minimum overall length which connects a given set of points. All of the minimal trees are highly incongruent with each other, and none is congruent with any actual biological tree (Synanen, et. al., 1989)

- Molecular analysis indicates that modern birds and mammals diverged at least 100 million years ago. However, fossil records indicate a branching time of only 50-70 million years ago (Benton, 1999).
- Molecular analysis also indicates that guinea pigs are not a part of the rodent (mice, rats, etc.) family, even though they are morphologically related (Sullivan, et. al. 1997).
- Molecular analysis suggests the African elephant is more closely related to the mammoth, whereas morphological analysis suggests the Asian elephant is a closer relative to the mammoth, compared to its African cousin (Thomas, et. al., 2000).

Furthermore, with the use of gaps to "improve" alignments, hot-spots may be inadvertently created where there are none, especially for sequences which are homologous but have very insignificant sequence similarity. These problems arise mainly because substitution matrices are based on probability theory, rather than biological theory (Eddy, 2004).

2.8 Explicit Similarity and Implicit Similarity

Multiple sequence alignment tools process protein sequences *as they are.* In other words, no transformation is applied to the sequences in order to generate another form of representation before the processing begins. An example of this is character-to-character matching. Therefore, when we refer to sequence similarity in the context of multiple sequence alignment, we are generally referring to explicit similarity.

Besides explicit similarity, sequences could also have implicit similarities. For example:

- Source: T H E Q U I C K B R O W N F O X
- Target : D E R K W I K B R A U N F A U X

The "pronunciation" of the target is very similar to that of the source, even though the "spelling" is not well-formed (i.e. "noisy"). This is analogous to having proteins that differ in terms of amino acid sequencing, but performs the same function due to similarities in three-dimensional structure.

Sequences with insignificant explicit similarity may have significant implicit similarities. In order to bring to surface implicit similarities, raw sequences require a transformation in order to generate an alternative representation, which will make explicit what was previously implicit. Different transformations will generate different representations. Different representations will bring to surface different implicit similarities.

2.9 Protein Classification using Multiple Sequence Alignment

The goal of multiple sequence alignment is to *discover regions of explicit similarity among protein sequences known to be of the same family.* Once a collection of sequences known to be of the same family is multiply aligned, this alignment can be used as a benchmark to evaluate unknwon sequences. An unknown sequence is aligned against that particular multiple alignment and the score calculated. A high score would indicate that the unknown sequence most likely belongs to that particular family. However, multiple sequence alignment has difficulty coping with implicit similarities. Alternatively, Artificial Intelligence methods may be more suitable for protein classification.

2.10 Artificial Intelligence Methods

Artificial intelligence methods are systems in computer science which are able to be trained to extract buried patterns from or derive conclusions on a particular mass of data. Artificial intelligence methods differ from conventional algorithmic methods in the sense that conventional algorithmic methods are step-by-step instructions on how to solve a given problem, whereas artificial intelligence methods tend to discover on their own the solution to the problem. In order to be able to arrive at the solution on their own, Artificial Intelligence methods has built-in mechanisms that enable them to handle "noise". The ability to handle noise enables Artificial Intelligence methods to handle implicit similarities.

The main artificial intelligence methods used in bioinformatics are:-

- Fuzzy Logic
- Genetic Algorithms
- Hidden Markov Models (HMM)
- Neural Networks (Connectionist Networks)
- Support Vector machines (SVM)

2.11 Fuzzy Logic

Fuzzy logic is an alternative to traditional notions of logic. The central notion of fuzzy logic is that truth (i.e. set membership) values are indicated by a range, typically between 0.0 to 1.0, with the minimum value of the range representing absolute falseness, and the maximum representing absolute truth.

For example, take the statement: Jane is old. If Jane is 55 years old, we might assign the statement above the fuzzy value of 0.55. At this juncture, it is important to distinguish between fuzzy logic and probability. Both may operate over the same

numeric range. However, using the probabilistic approach, the value 0.55 yields the statement: "There is a 55% chance that Jane is old". Whereas, the fuzzy approach yields the statement: "Jane's degree of membership in the set of old people is 0.55". In other words, the probabilistic approach still implicitly carries the notion of old-or-young, whereas, with the fuzzy approach, Jane may be considered to be both old and young, the difference being in the degree of membership. The sets in a fuzzy logic system are also susceptible to typical set operations, such as intersections, unions, exclusions, etc.

2.12 Genetic Algorithms

Genetic Algorithm programs start with a population of randomly generated proposed solutions. The goal is to evolve this set of randomly proposed solutions to generate an optimum, or near-optimum solution. For the purpose of evolution, a **fitness function** is required. A fitness function is a method used to evaluate the optimality of a particular solution. Comparison of various possible solutions is then possible, and culling is done based on the intent to eliminate a pre-determined percentage of the less optimal solutions in the population pool. Once the lesser solutions are eliminated, the better solutions are then jointly used to generate additional proposed solutions, by mutating the representations of those better solutions. This process is repeated until a global minimum or maximum is reached.

The SAGA (Sequence Alignment by Genetic Algorithm) project (Notredame, 1996) attempted to use genetic algorithm for the purpose of multiple alignment. SAGA's pseudocode is given below:

INITIALISATION

1. create Generation 0.

EVALUATION