CHAPTER III
SEQUENCE ARITHMETIC

Proteins are classified by various methods; one technique is based on motifs, the functional or structural aspect of a protein. Motif is a consecutive sequence of characters that occur frequently in a protein. The occurrence of a specific motif is one of the significant features [Falquet, 2002].

Motif based analysis needs experts intervention in protein classification. Since the main objective of the study is to develop machine-learning tool with minimal human interference and the power of motifs motivated to contemplate the analysis based on occurrence of characters in the present work.

Protein is a sequence of 22 characters (20 labels for amino acids and two characters (B, Z) for substitution of amino acids). A protein essentially need not contain all amino acids. The absence of amino acids can be deliberated as a feature of that protein. This chapter expounds a systematic way of dealing with sequences by introducing various types of sets based on the occurrence of amino acids in a given protein (or) set of proteins.

Sets, operations and representation aspects of the sets and inferences associated with them have been referred as Sequence Arithmetic (SA) in the present work.

Sequence definitions along with their characteristics, the generation of different sets have been discussed in Section 3.1. Creation of knowledge base has been described in Section 3.2. Demonstration of Sequence Arithmetic has been illustrated in Section 3.3. The knowledge extraction process is presented in Section 3.4. Design and implementation issues of Sequence Arithmetic have been discussed in Section 3.5. Demonstration has been carried out on G-Protein Coupled Receptors and Myoglobin datasets. The reduction in search space has been elucidated in Section 3.6. The complexity analysis of the module is discussed in Section 3.7, followed by conclusions in Section 3.8.
3.1 SEQUENCE ARITHMETIC

Sequence types and basic string definitions are discussed in this section.

3.1.1 Sequence

Sequence has been defined in many ways depending upon the instance, location and domain. Some of the definitions of sequences with examples are given as follows:

- A serial arrangement in which things follow in logical order or a recurrent pattern.
  
  ➔ The sequence of names was alphabetical.

- Following of things one after another in time.
  
  ➔ The doctor saw patients in a sequence.

- Film consisting of a succession of related shots that develop a given subject in a movie.

- Succession: the action of following in a particular order.
  
  ➔ He played the trumps in sequence.

- In mathematics, a sequence is a list of objects (or events) arranged in a ‘linear fashion’, such that the order of the members is well defined and significant.

- The linear arrangement of building blocks in biological macromolecules like DNA, RNA, protein and polysaccharides. DNA and RNA macromolecules are linear polymers of nucleotides. Proteins are linear polymers of amino acids. Polysaccharides are linear and branched polymers of monosaccharides (sugars).

  In the present research, protein is a sequence of varying length with permutations of 22 characters (20 amino acids and 2 substitution characters).

3.1.2 String

Let $\Sigma$ be any finite set of alphabets. A finite sequence of zero or more elements chosen from $\Sigma$ is a string over $\Sigma$. The length of the string ‘$n$’ is denoted by $|n|$. The set of strings of length ‘$k$’ is denoted by $\Sigma^k$.

The well-formed formulae representation of a string as defined [Tremblay, 1987] [Mott, 1999] are:
1. \( \Sigma^0 = \{\Lambda\} \); it is null string.
2. \( \Sigma^{k+1} = \{wa \mid w \in \Sigma^k \text{ and } a \in \Sigma\} \) for \( k \geq 0 \).
3. \( \Sigma^* = \bigcup_{k=0}^{n} \Sigma^k \), denoting the set of all strings over \( \Sigma \).
4. \( \Sigma^+ = \bigcup_{k=1}^{n} \Sigma^k \), denoting the set of all non-null strings over \( \Sigma \).

Protein is a sequence consisting of 22 characters (20 amino acids and 2 substitution characters).
\( \Sigma = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\} \cup \{B, Z\} \) (Table 2.1).

Each and every sequence obtained from the definition may not be a protein though characters confine to the set of character ‘\( \Sigma \)’. The following characteristics of proteins have been obtained:

i. Length of sequence.
ii. Location / Space preserving information.
iii. Frequency tabulation of characters (pattern representation).
iv. Statistical information based on Run length encoding.

Frequency tabulation of a character ascertains the information of occurrences of a character in the sequence. The characters, which have non-trivial frequency in a frequency table of a given sequence ‘s’ constitute A-Set of the sequence and is denoted as \( A(s) \). Similarly the character that has trivial frequency (i.e., ‘0’) in a frequency table is called D-Set of the sequence and written as \( D(s) \).

\( D_\Sigma(s) = \Sigma - A(s) \).

Similarly A-Set, D-Set for a given set of sequences ‘\( S \)’ are denoted by a

1. \( A(S) = \bigcup_{s \in S} A(s) \).
2. \( D(S) = \Sigma - A(S) \).

In addition to these two sets, Prime set (denoted as \( P(S) \)) has been defined which is intersection of A-Set of each sequence of \( S \).

\( P(S) = \bigcap_{s \in S} A(s) \)

By considering a set of proteins the above concepts have been demonstrated in the following section.
3.1.3 Protein Sequences

Proteins are long chains of twenty different amino acids and two substitutions characters i.e., $\Sigma = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\} \cup \{B, Z\}$ (Table 2.1), that serve as building blocks for proteins. Each amino acid has a specific chemical structure, which contains a carbon backbone similar to all other amino acids, and a residue. The residue varies from one amino acid to another amino acid. The length of a protein chain can range from 50 to 1000-5000 amino acids (350 on average). Proteins are the molecules with diverse chemical properties.

The proteins [Mount, 2001]

- form functional and structural machinery of each cell in every organism.
- are involved for giving biological signals.
- act as catalyst for metabolic activities of the body.
- have many levels of structure viz.,
  i. Primary structure, which is the 1-D sequence.
  ii. Secondary structure (2-D), which is a composition of the regular substructures that the protein polymer forms due to streric and hydrogen bond interactions.
  iii. Tertiary structure (3-D), which is the most complex protein structure, composed of multiple chains.

Proteins are essentially grouped according to their secondary structure but not by functional families. The focus in protein classification has been on finding proteins that have similar chemical architectures. The significant property of a protein is the length and composition of amino acid chain. The series can be obtained automatically from the gene that encodes for the protein. The amino acids composition of a protein determines the tertiary structure, which is unique [Mount, 2001].

A protein sequence is a sequential arrangement of the amino acids (with permutation). Sequence Arithmetic for protein classification aims at finding the information with respect to essential amino acids, which occurs in a family. This amino acids information has been used to map with the sequences that are already present in the protein database-Brookhaven National Laboratory [Brookhaven NL].
Various Sequence Arithmetic operations have been used to reduce the search space, find the nearest neighbors within a family and to which family the sequence belongs.

### 3.1.4 Mathematical Representation of Set

Let ‘s’ be any sequence and ‘S’ be set of ‘n’ sequences, then the mathematical representations are as given below:

\[
A(s) = \{x / x \text{ is valid character in the sequence 's'}\}
\]

\[
A(S) = \bigcup_{s \in S} A(s)
\]

\[
A(s) \subseteq A(S)
\]

\[
D(S) = \Sigma - A(S)
\]

\[
P(S) = \bigcap_{s \in S} A(s)
\]

The knowledge base has been built by finding the \(\Sigma\), A-Set, D-Set, and P-Set for each family of sequences collected from the sample data of sequence Information System. They have been used to determine the family of an unknown sequence using expert system techniques namely forward chaining and backward chaining [Ramadevi, 2004b].

### 3.1.5 Alphabet Set (A-Set)

**A-Set of a sequence:** \(A(s)\): It consists of all the symbols, which appear in the sequence.

**A-Set of a class:** \(A(C)\): It is the union of A-Set of all the sequences, which belong to a particular class.

**A-Set of a family:** \(A(F)\): It is the union of A-Set of all the class that belongs to a particular family. In some cases, it is equal to Universal Set (\(\Sigma\)).

**Example 3.1:**

Consider the sequences of a class belonging to Myoglobin family

**Sequence 1**

<table>
<thead>
<tr>
<th>s1</th>
<th>GLSDGEWELVLKTWGBKVEDIPGHEFVLVRLFTGHPTELEKFDKFKHLKTE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GEMKASEDLKKQGVTVLTAAGGLKGGHHEAIQPLAQSHATKHKIPKLEF</td>
</tr>
<tr>
<td></td>
<td>ISDAIIHLQSKHAPAEFGADAQGAMKKALELFRNDIAAKKELGFQG</td>
</tr>
</tbody>
</table>

\[A(s_1) = \{A, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W\}\]
Sequence 2(s2)
GLSDGEWQLVLHVWGKVEADLAGHGQEVLIRLFKGHPETLEKFNKFHKIKSE
DEMKASEDLKKHGVTVLTLGGLKKGGHHEAEIKPLAQSHATKHKIPIKLE
FISEAIHVLQSKHPGFADAGAMNKALELFKDIAYAKKEKLGFQG

A(s2) = \{A,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y\}

Let C = \{s1, s2\}

A-Set for the class is the union of the A-Set of all the sequences s_i present in the class. The accuracy of the set depends on the number of samples taken.

A(C) = A(s1) \cup A(s2)

\cup \{A,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y\}\}


A-Set for a family: It is the union of A-Set of all the classes that belong to a particular family. A-Set for Myoglobin family is as follows:


3.1.6 Difference Set (D-Set)

D-Set for a sequence: It consists of all the symbols that are not present in the sequence.

D-Set for the Sequence 1 and Sequence 2 are as follows:

D(s1) = \{B,C,Y,Z\}

D(s2) = \{B,C,Z\}

D-Set for a class: It is the intersection of D-Set of all the sequences, which belong to a particular class.

D(C) = \{\{B,C,Y,Z\} \cap \{B,C,Z\}\}

= \{B,C,Z\}

D-Set for a family: It is the intersection of D-Set of all the classes that belong to a particular family.

D(Myoglobin) = \{\}\ (Appendix A)
3.1.7 Prime Set (P-Set)

P-Set for a class-P(C): It is the intersection of A-Set of all the sequences, which belong to a particular class.

P-Set for a family- P(F): It is the intersection of A-Set of all the sequences, which belong to a particular family.

Example 3.2:
Consider the Sequence 1 and Sequence 2

\[
\]

P(Myoglobin) = {} /* it is null set */

The A-Set, D-Set and P-Set for the Myoglobin family and G-Protein Coupled Receptors have been given Appendix A

3.2 SEQUENCE ARITHMETIC KNOWLEDGE DATABASE

The classification pattern of proteins as discussed in Chapter II comprises:

- Proteins are classified into families.
- Families into classes.
- Classes into subclasses and so on.

A unique identification number (Fid) has been given to a family. Attaching a Family ID(Fid) to the sequence helps to acquire the knowledge for classification of a new sequence. The information regarding A-Set, D-Set and P-Set of the family have been generated and stored in the Sequence Arithmetic Knowledge Database.

These sets have been utilized in framing the Sequence Arithmetic Rules, which are discussed in Section 3.5. These generated rules have been stored in Sequence Arithmetic Knowledge Database and are used for the identification of unknown sequence. It has been possible to identify an effective inference engine model that works on this knowledge base for the classification of unknown sequences.
3.3 DEMONSTRATION OF SEQUENCE ARITHMETIC

The input to the Sequence Arithmetic module is the unidentified sequence ‘y’. A(y) and D(y) of the unknown sequence are computed. Knowledge has been extracted by mapping these sets with the information present in the Sequence Arithmetic Knowledge Database by following the rules (Section 3.4.1). The output of the Sequence Arithmetic module is the family to which the sequence belongs. The family information < F > consists of A-Set, P-Set and D-Set for the corresponding family along with the Fid. The output of the Sequence Arithmetic module will be the A(y), D(y), A(F), D(F) and P(F). The Sequence Arithmetic module is shown in Figure 3.1.

![Figure 3.1: Sequence Arithmetic module](image)

The internal processes are shown in Figure 3.2. Sequence Arithmetic Knowledge Database consists of data related to various families. When unknown sequence ‘y’ is given as input to Sequence Arithmetic, the map process is used to map the unknown sequence information with the different families’ information in the Sequence Arithmetic Knowledge Database.

![Figure 3.2: Internal processes of the Sequence Arithmetic module](image)

49
3.3.1 Sequence Arithmetic Datasets

For a given set of sequences, which belong to a class, sets have been generated as discussed in Section 3.1.4. These sets play an important role in the classification / identification of a novel sequence. G-Protein Coupled Receptors dataset and Myoglobin dataset have been used for generation of different sets. A sample of 1041 out of 3896 G-Protein Coupled Receptors family consisting of five classes (Class A, B, C, D and E) has been taken for demonstration [Horn, 1998]. Sequence Arithmetic has also been demonstrated on sample set of Myoglobin sequences consisting of 399 out of 727 available sequences. These sequences are categorized into 36 classes.

3.4 KNOWLEDGE EXTRACTION FROM SEQUENCE ARITHMETIC REPRESENTATION FOR PROTEIN FAMILIES

The input for this process is sequence information. The following information about the sequence is required for the knowledge extraction:

1. Sequence.
2. Sequence Family ID.

The necessary steps in the process of extraction of knowledge base for the sequences are as follows:

1. Extract the distinct family IDs from the sequence samples.
2. Compute A-Set for each sequence.
3. For each sequence ‘s’ within a class and each class ‘C’ within a family ‘F’ apply the following rules:
   i)  $A(C) = \bigcup_{s \in C} A(s)$.
   ii) $A(F) = \bigcup_{C \in F} A(C)$.
4. Compute D-Set based on rules below:
   i)  $D(C) = \Sigma -A(C)$.
   ii) $D(F) = \Sigma -A(F)$.
5. Compute the P-Set of the family.

The intersection of A-Set of all the sequence of a family generates the P-Set of a family.
i) \( P(C) = \bigcap_{s \in C} A(s) \).

ii) \( P(F) = \bigcap_{C \in F} A(C) \).

### 3.4.1 Inference Engine for Identification of Unknown Sequence

The knowledge base has been built by finding the \( \Sigma \), A-Set, D-Set, and P-Set for each family of sequences collecting the sample data from sequence Information System. They have been used to determine the family of unknown sequences using expert system techniques namely forward chaining and backward chaining.

**Lemma 3.1:** If \( s \in F \), then \( D(F) \subseteq D(s) \).

**Proof:** Let \( s \in F \), then \( A(s) \subseteq A(F) \).

\[
D_\Sigma(s) = \Sigma - A(s) \quad \text{and} \quad D_\Sigma(F) = \Sigma - A(F).
\]

If \( A(s) \subseteq A(F) \) then \( D(F) \subseteq D(s) \) (Proved).

**Lemma 3.2:** If \( s \in F \), then \( P(F) \subseteq A(s) \).

**Proof:** Let \( s \in F \), then \( A(s) \subseteq A(F) \).

\[
P(F) = \bigcap A(s)
\]

If \( A(s) \subseteq A(F) \) then \( P(F) \subseteq A(s) \) (Proved).

**Corollary 1:** If \( D(F) \not\subset D(s) \) then \( s \) will not be a sequence in family \( F \).

**Corollary 2** If \( P(F) \not\subset A(s) \) then \( s \) will not be possible sequence in family \( F \).

The above said lemmas and corollary helps in developing algorithms and improving the efficiency of the algorithms for classifying a new sequence. The following rules are used for reducing complexity of classification of unidentified sequences into one of the available family.

**Rule 1:** If \( D(F) \not\subset D(s) \), then do not search in family \( F \).

**Rule 2:** If \( P(F) \not\subset A(s) \), then do not search in family \( F \).

**Rule 3:** If \( D(F) \subseteq D(s) \), \( F \) is possible search space.

**Rule 4:** If \( P(F) \subseteq A(s) \), \( F \) may be potential family where ‘\( s \)’ can be.
The rules for not searching definitely are Rule 1 and Rule 2, whereas Rule 3 and Rule 4 still have ambiguity. In order to reduce the ambiguity, exhaustive knowledge has to be extracted to arrive at better decision rules. The same are subject of interest in Chapters IV, V and VI.

The inference engine follows Algorithm FIdentifier for the classification of unknown protein sequence ‘y’.

Algorithm 3.1
Algorithm FIdentifier (y)
Input: Sequence ‘y’.
Output: Family F.
1. Compute the A(y).
2. Use the following conditions for deciding whether the sequence belong to a family or not using Sequence Arithmetic Knowledge Database. (Appendix A)
   a. \( A(y) - A(F) = \emptyset \)
   b. \( P(F) - A(y) = \emptyset \)
   c. \( D(F) \cap A(y) = \emptyset \)
   If any of the above conditions are not satisfied, it implies that the sequence does not belong to that family.
3. Step two will give a set of family / families closely related to the sequence ‘y’.

Algorithm 3.1 will give a set of protein families, which are very close to the unknown sequence ‘y’. Minute operations can be performed for further classification for finding exact family. The algorithm may return more than one family for some sequences, therefore, it is advised to rank these identified families by taking the closeness of \( A(s) \) to \( P(F) \) and \( D(s) \) to \( D(F) \). The methods developed in the following chapters need to be followed for all identified families in rank order.

3.5 SEQUENCE ARITHMETIC MODULE
The architecture of the model expert system and different methods used for the knowledge extraction are discussed in this section.

3.5.1 Architecture of Model Expert System
The architecture of a Model Expert System is shown in Figure 3.3. All the sequences have been stored in the Sequence Database, in the format as in Table 3.1.
using a VB module. Here the Sequence Database can be updated whenever new sequence is identified. The different interface methods generated in this Framework are shown in Table 3.2 and the output screens are as in Figures 3.4, 3.5, 3.6 and 3.7.

![Architecture of Expert System](image)

**Figure. 3.3: Architecture of Expert System**

**Table 3.1: Format for sequence stored in database**

<table>
<thead>
<tr>
<th>S.No</th>
<th>Sequence ID</th>
<th>Fid</th>
<th>Sequence</th>
</tr>
</thead>
</table>
Figure 3.4: Extraction of sequence – minimum number of sequences are considered.

Figure 3.5: Generation of Alphabet Sets for all sequences in the database
Figure 3.6: Generation of A-Sets for a given sequence

Figure 3.7: Testing of A-Set for a given sequence and displaying the families that are closer to it
Sequence Arithmetic has been used to identify a family, which has been found to be faster than other methods. The arithmetic is totally dependent on known protein samples for the classification. The accuracy of this method depends on the size of the protein sequence information system and accuracy of the data in the sampling system. The limitation of the method is that it cannot give assurance for finding exact protein family. But it gives probable set of family / families to which the protein sequence belongs, depending on the behavior of the sample protein sequence of the family.

### 3.5.2 Sets Generation for G-Protein Coupled Receptors Family

The G-Protein Coupled Receptors family has been taken as a sample database; they are divided into 5 classes (class A, B, C, D and E). By considering the sequences, which belong to Class A of G-Protein Coupled Receptors dataset and the following sets, are generated:

\[ D(A) = \{B,Z\} \]
Similarly the outputs for the remaining classes is as shown in Table 3.3

Table 3.3: A-Set, P-Set and D-Set for GPCR database

<table>
<thead>
<tr>
<th>Class</th>
<th>A-Set</th>
<th>D-Set</th>
<th>P-Set</th>
</tr>
</thead>
</table>

Given an unknown sequence ‘y’, generate the A(y), D(y) and then apply Rules 1-4 as given in Section 3.4.1. These A(y) and D(y) have been mapped with the rules in Sequence Arithmetic Knowledge Database and family information has been obtained.

If the D(y) is {B,Z} then the sequence ‘y’ belongs to GPCR family.


3.5.3 Testing and Result Analysis

Consider the unknown Sequence ‘y’

GLSWELVLKTKVEADIPGHGEGFLVRLFTGHPELEKFDKFKHLKTEGEM
KASEDLKKQGVTVLALGGILKKKGHEAEIQPLAQSHATKHKPICIPIKYLEFISDAIIHVLQSKHPAEFGADAQGAMFRNDIAAKYKELGFQG

D(y) = {B,Z}
A(y) ⊃ P (GPCR) and A(y) ⊃ P(Myoglobin).
Therefore ‘y’ belongs to two families.

Now compute A(y) - P(GPCR) = { } and
\( A(y) - P(\text{Myoglobin}) = \{A,C,D,E,F,G,H,I,J,K,L,M,N,P,Q,R,S,T,V,W,Y\} \)

The identified families are ranked in increasing order of number of elements in the resultant set. The families are searched in rank order, so the sequence ‘y’ is searched in GPCR first and it belongs to GPCR family (Appendix A). Thus the search is confined to only GPCR family.

### 3.6 REDUCTION OF SEARCH SPACE

If an unknown sequence ‘y’ is given then BLAST finds similar sequences by comparing ‘y’ with all the sequences in the database. In Sequence Arithmetic module the family information has been obtained by computing \( A(y) \), \( D(y) \) and subsequently applying the Sequence Arithmetic Rules present in the Sequence Arithmetic Knowledge Database. The first level of horizontal fragmentation of domain search space is achieved.

A sample dataset of Myoglobin family and GPCR family containing 1440 sequences have been considered for demonstration of SA module. A random selection of 60 sequences has been made and used for testing. Sequence Arithmetic has been applied on them. 32 sequences were classified into one family (Myoglobin) and 28 sequences has multi-valued output. If a new sequence is given, it is classified into one of the two families, thus on an average 60% of the database has to be searched. Thus the reduction of domain search is about 40% by applying Sequence Arithmetic module (Appendix D).

### 3.7 COMPLEXITY

The complexity of the Sequence Arithmetic module depends on A-Set, D-Set and P-Set and storing the data in Sequence Arithmetic Knowledge Database.

Let \( T_A \) be the complexity for finding the A-Set. For initial ‘n’ sequences, a single scan for A-Set is to be performed. \( T_A = O(nm) \), where ‘m’ is maximum length of any sequence. If the number of classes within a family is ‘C’ then \( T_{A(C)} = O(22n)(\text{as number of characters is 22}) \). For a family \( T_{SA} = O(nm) + O(3n*22) = O(nm) \).

Let ‘b’ be the complexity of searching for a sequence in a database with ‘n’ sequences without using Sequence Arithmetic.
Family information is obtained by applying Sequence Arithmetic module. If \( n_1 \) sequences are present in a family, then \( n_1 < n \).

With Sequence Arithmetic it will be \( T_{SA} + O(n_1b) \) which will be less than or equal to \( O(nb) \). Therefore the complexity of Sequence Arithmetic is \( O(n_1b) + O(nm) \).

### 3.8 SUMMARY

The search space has been considerably reduced both for Myoglobin and G-Protein Coupled Receptors. The horizontal fragmentation of the dataset has been achieved by using Sequence Arithmetic, thus reducing the complexity of the search. The generation of A-Set and D-Set for the unknown sequences, and applying the rules reduced the domain search space to either of the families. The application of the Sequence Arithmetic Rules to Myoglobin and GPCR families has shown about 40% reduction in the search space (Appendix D).

The result of applying Sequence Arithmetic is that family information is generated and at present the search is confined only to it. The inability to identify the class within the family is conquered by proceeding to the Reduct based Decision Tree module (Chapter IV) and Neighborhood Associations module (Chapter V).