International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 24 Issue 2 – FEBRUARY 2017.

Neural Network Based Protein Sequence Analysis System

J. SHEELA JASMINE[#1]

[#] *Assistant Professor and Head of the Department, Department of Information Technology,*
Annai Vailankanni Arts and Science College, Thanjavur, India

***Abstract*— In this project the protein sequence is analyzed and displayed in a table according to the presence of alpha, beta, or random in the sequence. The final table is then converted into a 2d structure according to the sample sequence. A basic principle of molecular biology is protein sequence determines protein structure knowledge of a protein's amino acid sequence, called the primary structure, and makes it possible to predict more complex levels of that protein's structure. Indeed, protein structure is organized in a hierarchy. At the most basic level is the primary structure. The next and most important level, secondary structure is mainly composed of alpha helices and beta strands, which are formed from local sequences of amino acids. Even more complex are the tertiary and quaternary structures, each of which is based on the elements of the structure preceding it in the hierarchy. Knowing a protein's secondary structure helps to determine the structural properties of the protein. Several methods have been developed to determine secondary structure, with varying accuracy. One method involves analyzing the X-ray direction patterns of crystallized proteins. While X-ray direction is rather time consuming, it is extremely accurate. Another method, structure homology, or threading, utilizes an amino acid sequence with a known secondary structure as a model to predict the secondary structure of another similar sequence.**

***Index Terms*— Molecular biology, Quaternary structures, Time-consuming, Threading**

I. INTRODUCTION

A basic principle of molecular biology is that protein sequence determines protein structure knowledge of a protein's amino acid sequence, called the primary structure, and makes it possible to predict more complex levels of that protein's structure. Indeed, protein structure are organized in a hierarchy. At the most basic level is the primary structure. The next and most important level, the secondary structure is mainly composed of alpha helices and beta strands, which are formed from local sequences of amino acids. Even more complex are the tertiary and quaternary structures, each of which is based on the elements of the structure preceding it in the hierarchy. Knowing a protein's secondary structure helps to determine the structural properties of the protein. Several methods have been developed to determine the secondary structure, with varying accuracy. One method involves analyzing the X-ray di.raction patterns of crystallized proteins. While X-ray di.raction is rather time-consuming, it is extremely accurate. Another method, structure homology, or threading, utilizes an amino acid sequence with a known secondary structure as a model to predict the secondary structure of another similar sequence. In this project the protein sequence is analyzed and displayed in a table according to the presence of alpha, beta, or random in the sequence. The final table is then converted into a 2d structure according to the sample sequence.

Prediction of structure began in the 1960s when the first protein crystal amino acids could be used to predict helices in myoglobin and haemoglobin. Schiffer & Edmunson(1967) developed the helical wheel both to predict helical potential and, if a helix is present, to indicate the presence of a hydrophobic region. Ptitsyn (1969) studied secondary structurs of seven globular proteins and found certain amino acids were partitioned differently between helical and non-helical sections; his conclusions agreed with prothero. More details of the early history in this field are given in Fasman (1989b).

The most commonly used is the secondary structure predication methods. One approach of the method is the statistical methods. Alternative approaches also applied to the problem of predicting secondary and tertiary structure, which include *neural networks and molecular modeling*

II. PROTEIN IN BIOINFORMATICS

Bioinformatics if fast becoming an oft-uttered buzzword these days. Bioinformatics in nothing but good, sound, regular biology appropriately dresses so that it can fit into a computer. Bioinformatics is about searching biological databases, comparing sequences, looking at protein structures, and more generally, asking biological questions with a computer. We'd like to take a more selective approach by focusing on those aspects of protein sequences where bioinformatics analyses can be most useful. The following list describes the topic relevant to proteins, which plays important role in the field of bioinformatics.

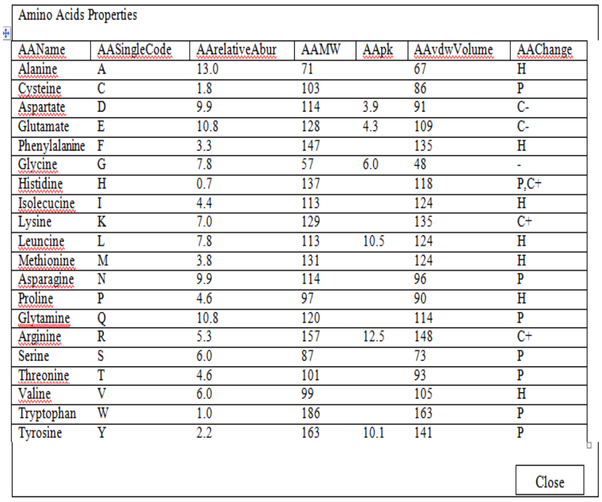
- Retrieving protein sequences
- Computing amino acid composition
- Predicting elements of secondary
- Visualizing protein structures
- Classification of protein into families
- Evolutionary relationships between proteins

III. Analyzing Protein Sequences

Moreover, all the protein molecules are made up of the same building blocks, called amino acid. Amino acids are already quite complex organic molecules, made up of carbon, hydrogen, oxygen, nitrogen and sulphur atoms. So the overall recipe for the protein is something like C^{1200} H2400 O600 N300 S100. The early days of biochemistry were devoted to finding out better way to represent proteins, preferably in terms of a formula that would explain their biological properties.

Biochemists realized over time that proteins were huge molecules(macromolecules) made up of large numbers of amino acids(typically from 100 to 500), picked out from a selection of 20 "flavours" with names such as alanine, glycine, tyrosine and so on. The table below provides the list of the twenty building blocks, with their full names, three letter codes, and one-letter codes(*IUPA*C code, after the International Union of Pure and Applied Chemistry committee that designed it).

TABLE 1



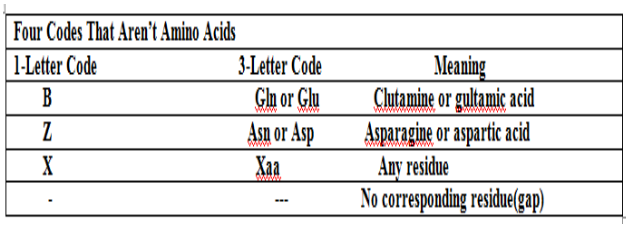
Amino Acids Properties

AAName	AASingleCode	AArelativeAbur	AAMW	AApk	AAvdwVolume	AAChange
Alanine	A	13.0	71		67	H
Cysteine	C	1.8	103		86	P
Aspartate	D	9.9	114	3.9	91	C-
Glutamate	E	10.8	128	4.3	109	C-
Phenylalanine	F	3.3	147		135	H
Glycine	G	7.8	57	6.0	48	-
Histidine	H	0.7	137		118	P,C+
Isolecucine	I	4.4	113		124	H
Lysine	K	7.0	129		135	C+
Leuncine	L	7.8	113	10.5	124	H
Methionine	M	3.8	131		124	H
Asparagine	N	9.9	114		96	P
Proline	P	4.6	97		90	H
Glytamine	Q	10.8	120		114	P
Arginine	R	5.3	157	12.5	148	C+
Serine	S	6.0	87		73	P
Threonine	T	4.6	101		93	P
Valine	V	6.0	99		105	H
Tryptophan	W	1.0	186		163	P
Tyrosine	Y	2.2	163	10.1	141	P

IV. Additional Amino Acid Codes

When working with or analyzing programs, some unusual letters may need to pop now and then in the protein sequences. These letters don't designate actual amino acids, but are used to denote various level of ambiguity- or a total lack of information-about certain positions in the sequence.

TABLE 2



Four Codes That Aren't Amino Acids

1-Letter Code	3-Letter Code	Meaning
B	Gln or Glu	Clutamine or gultamic acid
Z	Asn or Asp	Asparagine or aspartic acid
X	Xaa	Any residue
-	---	No corresponding residue(gap)

A. History of Sequence Analysis

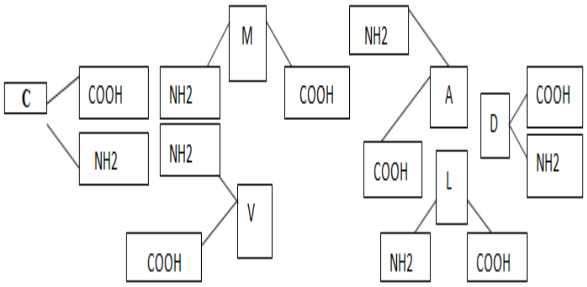
Besides earning Alfred sanger his first Nobel prize, the sequencing of insulin inaugurated the modern era of molecular biology. In early sixties, protein sequences accumulated slowly – perhaps a blessing in disguise, given that computers capable of analyzing them were not yet invented !

In this precomputer era, sequences were assembled, analyzed, and compared by (manually) writing them on pieces of paper, taping then slide by slide on walls, and/or moving them around for optimal alignment or *pattern matching.* As soon as computers became available, the first computational biologists started to enter the manual algorithms into the available machines. This was new because nobody before that had manipulated and analyzed these kinds of molecular *texts.* Most methods had to be invented from scratch and in, the process, a new area research- the analysis of protein sequences using computers – was generated. *This is the genesis of bioinformatics.*

V. Reading Protein Sequences N To C

The twenty amino acids found in properties have different bodies but all have the same pair of hooks – NH2 and COOH – that are used to from the s0-called *peptidic bond* between successive residues in the sequence. The figure below shows the free individual amino acids floating about, displaying their hooks.

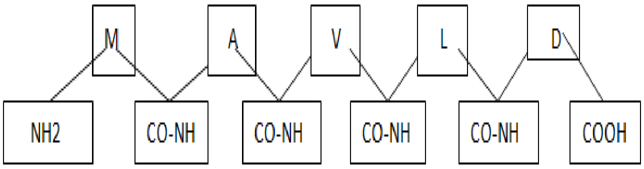
Figure 1



Free amino acids floating around

The protein molecule itself is then made chemically linking a free NH2 group with a COOH group with a COOH group, forming the peptide bond CO-NH. The figure below show a schematic picture of the resulting chain.

Figure 2



Amino acids chained together to constitute a protein molecule

As a result of this chaining process, the protein molecule is going to be left with an unused NH2 at one end and an unused COOH at the other end. These extremities are respectively called the *N-terminus and C- terminus* of the protein chain. This is important to know because the sequence of a protein is always defined as the succession of its constituent amino acids listed from the N-terminus to the C-terminus.

The sequence of the above protein is as follows:

MAVLD = Met – Ala – val – Leu – Asp = Methionine - Alanine – Valine Leucine – Aspartic

VI. Working With protein Structures

A protein molecule, once made, is not a chain-like, highly flexible object, rather a compact well bundled ball of string. The protein chain is affected by other influences, such as the electric charges carried by some of the amino acids, or their capacity to fit with their immediate neighbours.

Sequence αStructure αFunction

Playing with protein structure models and computer screen is, of course, much easier than carrying around a thousand – piece 3d plastic puzzle. As a consequence, an increasing proportion of the bioinformatics pie is now devoted to the development of tools to navigate between sequences and 3d structures. *This specialized area is called structural bioinformatics*. Thanks to many free resources on the internet, it is not difficult to display some beautiful protein pictures on any of the computer, and start playing with them like in video games.

VII. Accuracy of Predictions

To compare various methods of protein structure prediction requires a gold standard of structure. Structures derived from X-ray or NMR measurements are usually the standards for tertiary structure predication. Many published tertiary structures include secondary structure assignments, but these can be incomplete and subjective. Thus when secondary structure assignments, but these can be incomplete and subjective. Thus when secondary structure is predicted, it is often preferable to reassign secondary structure from tertiary coordinates. Accuracy of tertiary structure predication is usually measured by comparing the coordinates for correct and predicated structures using the *root mean square (R.M.S)* deviation, let x_i stand for a set of atomic coordinates for one atomic coordinates for one atom in a (possibly known) structure, and y_i for corresponding atom in a second (possibly known) structure. One can mathematically transform the set of y_i coordinates Yi such that the sum of the squares of the distance deviations

$$\sum \qquad x_1 \qquad | \quad - \qquad y_1 \qquad\qquad (1)$$

is a minimum. Then the R.M.S deviation is defined as

AR = √

Where *N* is the total number of atoms in the structure

A more frequently used measure, *singly residue accuracy*, is the number of residues correctly predicted to contain a structure divided by the number of residues that do contain that structure. To determine overall accuracy, this can be summed over the number of different structures or states predicated, usually either three (helix, strand, other) or four (helix, strand, turn, coil). Since turns and surface (random coil) loops are frequently interchanged in homologous proteins, three state accuracy is arguably the better measure. Also four-state values can easily be converted into three state ones.

Three-state single residue accurary (Q_3) is :

$$Q_3 = \frac{p + p + p \quad (5)}{N}$$

Where n is the total number of predicted residues and P_α is the number of correctly predicted secondary structures of type a. Q_3 values of from 0.5 to 0.7 (50 – 70% accuracy).

Lower limit for expected Q_3 accuracy is as low as 70%, an accuracy range achieved by several recent predictions. When compared secondary structures of proteins with the same tertiary fold, they found an average three-states single residue accuracy of 88.4%, with a standard deviation of 9%. All the above accuracy discussions assume a predictive method is being used as its developers intended. While the Chou- Fasman secondary structure prediction method can be carried out manually, it and most other methods are usually implemented in computer programs.

VIII. NEURAL NET AND BIOLOGICAL MODEL

A. *Neural Nets*

A neural net is an artificial representation of the human brain that tries to simulate its learning process. The team artificial means that neural nets are implemented in computer programs that is able to handle the large number of necessary calculations during the learning process. Neural networks are a form of multiprocessor computer system , with:

- Simple processing elements
- A high degree of interconnection-memory and processing elements collocated
- Simple scalar messages
- Adaptive interaction between elements
- Self-organization during learning

B. *The Biological Model*

The human brain consists of a large number(more than a billion) of neural cells that process information. Each cell works like a simple processor and only the massive interaction between all cells and their parallel processing makes the brain's abilities possible.

C. *FEED-FORWARD NETS*

Feed-forward nets are the most well-known and widely used class of neural network. The popularity of feed-forward networks derives from the fact they have been applied successfully to wide range of information processing tasks in such diverse fields as speech recognition, financial prediction, image compression, medical diagnosis and protein structure prediction; new applications are being discovered all the time.

IX. PROTEIN STRUCTURE

Every protein molecule has a characteristic three –dimensional shape, or conformation. Fibrous proteins, such as collagen and keratin, consists of polypeptide chains arranged in roughly parallel fashion along a single linear axis, thus forming tough, usually water-insoluble, fibers or sheets. Globular proteins, e.g., many of the known enzymes, show a tightly folded structural geometry approximating the shape of an ellipsoid (or) sphere. Because the physiological activity of most proteins is closely linked to their three dimensional architecture, specific terms are used to refer to different aspects of protein structure. The term primary structure denotes the precise linear sequence of amino acids that constitutes the polypeptide chain of the protein molecule.

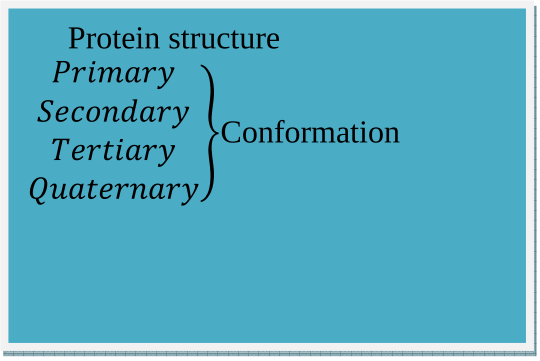
Automated techniques for amino-acid sequencing have made possible the determination of the primary structure of hundreds of proteins. The physical interaction of sequential amino-acid subunits results in a so-called secondary structure, which often can be either be a twisting of the polypeptide chain approximating a linear helix (configuration), or a zigzag pattern (-configuration). Most globular protein molecular are easily crystallized and have been examined by X-ray diffraction, a technique that allows the visualization of the precise three dimensional positioning of atoms in relation to each other in a crystal.

The tertiary structure of several molecules has been determined from X-ray diffraction analysis. Two or more polypeptide chains that behave in many ways as a single structural and functional entity are said to exhibit quaternary structure. The separate chains are not linked through covalent chemical bonds but by weak forces association.

The precise three-dimensional structure of a protein molecule is refereed to as its native state and appears, in almost all cases, to be required for proper biological function (especially for the enzymes). If the tertiary or quaternary structure of a protein is altered, e.g., by such physical factors as extremes of temperature, changes in pH, or variations in salt concentration, the protein is said to be denatured; it usually exhibits reduction or loss of biological activity.

Protein Structure

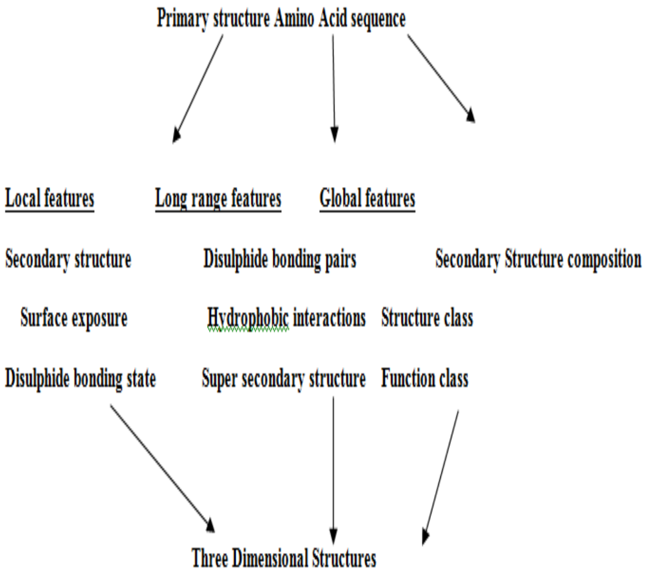
Figure 3



X. Primary Structure Analysis

The primary structure is the amino acid sequence of the protein. When primary structure is analyzed the potential interactions between amino acids are ignored. A basic principle of molecular biology is that protein sequence determines protein structure knowledge of a protein's amino acid sequence, called the primary structure, and makes it possible to predict more complex levels of that protein's structure. Indeed, protein structures are organized in a hierarchy. At the most basic level is the primary structure. Primary structure is fairly straightforward and refers to the number and sequence of amino acids in the protein or polypeptide chain. The primary structure or amino acids are attached to one another ultimately is dependent on the genetic code from DNA. This primary structure. However, the primary structure alone cannot carry out the functions of protein. The amino acids bonded to one another in a line don't make a protein function anymore than plants lined up along a roadside or in a nursery make a garden. Determination of primary structure is an essential step in the characterization of a protein.

Figure 4



XI. Secondary Structure

The next level of protein structure generally refers to the amount of structural regularity or shape that the polypeptide chain adopts. Proteins generally have a mix of different kinds of secondary structure. They are rarely, all of a particular type. One part of a particular protein makes have an alpha helix, whereas another part can be folded back in the beta sheet arrangement. Whatever type of secondary structure is held in place by *hydrogen bonding* interactions between the amino group of the amino acid residue and the carbonyl group of another amino acid residue.

The Secondary structure is the way a small part, spatially near in the linear sequence of a protein folds up into:

1. **Helices:** Where residues seemed to be following the shape of a spring. The most common are the so called ***alpha-helices.***
2. **Extended or Beta strands:** Where residues are in line and successive residues turn their back to each other.

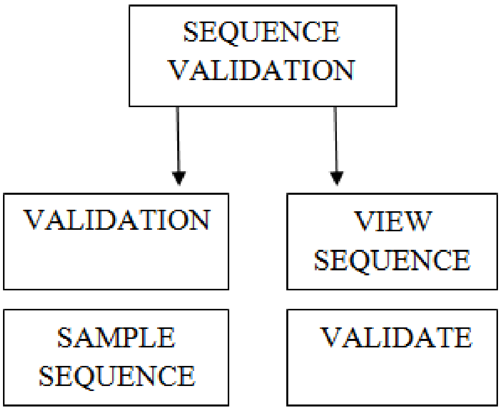
Random coils: when the amino-acid chain is neither helical or extended.

XII. Primary to Secondary Structure

Biologists refer to the sequence of a protein as its primary structure as opposed to the 3-D tertiary structure that is the final shape of the protein. An intermediary level of structure exists that is known as the Secondary structures. When the first crystallographers started looking at protein structures, they discovered (and predicted) that there was a hierarchy in the way that amino acid sequences fold onto themselves to become a biologically active molecule. Amino acids first look to their immediate neighbours in the sequence to form regions of regular, periodical conformations.

After this is done, the chain collapses further by folding those preshaped regions on to each other(or onto unstructured regions),leading to the final structure.

Validation

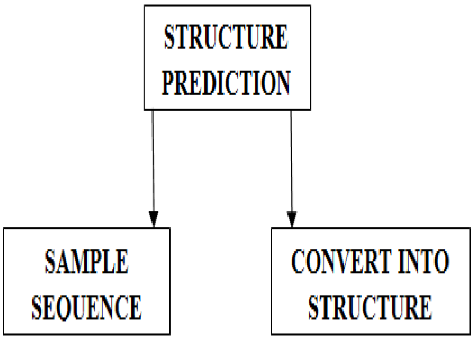


1. After the percentage and charts are computed then the sequence validation occurs.

2. A new sequence is entered in sequence validation. When the sequence entered does not match the existing sequence, the new is updated in the database.

3. When the already existing sequence is to retrieve, it can be obtained from the view sequence option.

Prediction



1. After the sequence is validated, then the structure is computed.

2. For structure formation, the sequence is got from the already existing database (i.e., form the sample), or inputting sequence directly can form it.

3. The final structure is obtained by converting the sequence entered in the sample sequence part.

XIII. CONCLUSION

It provides a broad overview and assessment of strengths and weakness of the various methods that may be used in the evaluation of proteins. The proteins secondary structure can be calculated (calculation of presence of alpha, beta and random). The proteins 2-D structure can be formed at the utmost accuracy. When the secondary structure is accurately predicated, tertiary structure can be formed which is the final shape of the protein. The top performance in this field is still very in adequate. In most cases the 3-D structures of the complex of two interacting proteins cannot yet be computationally derived from their individual structures.

REFERENCES

[1] Jackel M. Zurada, “Introduction to Artificial Neural Systems”, Jaico Publishing

[2] Bioinformatics – A Beginner’s Guide”, Jean-Michel Claverie, Wiley Publishing Inc.

[3] Protein sequences
www.ncbi.nml.nih.gov
www.expasv.ch

[4] Protein domains
www.ebi.ac.uk

[5] Protein Structures
www.rcsb.org/pdb
www.neurotechnologies.com
www.ncbi.org