PROTEIN SECONDARY STRUCTURE PREDICTION USING ENSEMBLE NEURAL NETWORKS WITH LOCAL AND LONG-RANGE AMINO-ACID FEATURES

FAWAZ HAMEED HAZZAA MAHYOUB

UNIVERSITI SAINS MALAYSIA

2021

PROTEIN SECONDARY STRUCTURE PREDICTION USING ENSEMBLE NEURAL NETWORKS WITH LOCAL AND LONG-RANGE AMINO-ACID FEATURES

by

FAWAZ HAMEED HAZZAA MAHYOUB

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosphy

October 2021

ACKNOWLEDGEMENT

In the name of ALLAH the Most Gracious and Most Compassionate

My first and foremost duty is that I must be sincerely thankful to Almighty ALLAH for giving me the strength and ability to complete my research work. Without his blessings, this achievement would not have been possible.

I would like to express my deep appreciation and heartfelt gratitude to my supervisor Professor Dr. Rosni Abdullah, School of Computer Sciences in Universiti Sains Malaysia, for her wise counsel, caring, patience and insightful guidance throughout the entire period of my research.

My family deserves special thanks. Words cannot express how grateful I am to my beloved father, mother, wife, brothers and sisters for their unwavering love, support, patience and for all of the sacrifices that they have done for me. Special thanks to my colleagues and friends for their very useful contributions made through our mutual discussions, suggestions and their motivation.

My sincere appreciations are extended to Taiz University for its financial support. Finally, thanks and appreciations to all staff of school of Computer Sciences in Universiti Sains Malaysia, who have helped and supported me in many ways.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF SYMBOLS

α	alpha
β	beta
C_α	α-carbon
Å	angstrom
ω	omega
ϕ	phi
ψ	psi

LIST OF ABBREVIATIONS

1D	one-dimensional
3D	three-dimensional
BGRU	Bidirectional Gated Recurrent Units
BLAST	Basic Local Alignment Search Tool
BLSTM	Bidirectional Long Short-Term Memory
BRNNs	Bidirectional Recurrent Neural Networks
CAMEO	Continuous Automated Model EvaluatiOn
CASP	Critical Assessment of Structure Prediction
CNN	Convolutional Neural Networks
CRF	Conditional Random Field
DCRNN	Deep Convolutional and Recurrent Neural Network
DeepCNF	Deep Convolutional Neural Fields
DNA	Deoxyribonucleic Acid
DNN	Deep Neural Network
DSSP	Define Secondary Structure of Proteins
FFNN	Feed-Forward Neural Network
FVS	Feature Vector Space
GSN	Generative Stochastic Network
HMM	hidden Markov model

IDRs	Intrinsically Disordered Regions
lDDT	local Distance Difference Test
LSTM	Long Short-Term Memory
MSA	Multiple Sequence Alignment
NEFF	Number of EFFective homologous
NMR	Nuclear Magnetic Resonance
NN	Neural Network
PALSSE	Predictive Assignment of Linear Secondary Structure Elements
PCASSO	Protein C-Alpha Secondary Structure Output
PDB	Protein Data Bank
PP	physiochemical properties
PROSIGN	PROtein Structure assIGNment
P-SEA	Protein Secondary Element Assignment
PSI-BLAST	Position Specific Iterated Blast
PSSM	Position Specific Scoring Matrix
ReLU	Rectified Linear Unit
RBM	Restricted Boltzmann Machine
RNA	Ribonucleic Acid
RNNs	Recurrent Neural Networks
SPOT-1D	Sequence-based Prediction Online Tools for one dimensional structural features

SSAE	Stacked Spars Auto-Encoder
SS	secondary structure
STRIDE	secondary STRuctural IDEntification
UniRef	UniProt Reference Clusters

LIST OF APPENDICES

PERAMALAN STRUKTUR SEKUNDER PROTEIN MENGGUNAKAN RANGKAIAN NEURAL ENSEMBLE DENGAN CIRI AMINO ASID TEMPATAN DAN BERANTAIAN PANJANG

ABSTRAK

Meramal struktur protein daripada jujukan adalah masalah yang mencabar. Pencarian struktur sekunder protein adalah pendekatan yang berkesan untuk menghasilkan struktur protein yang lengkap. Interaksi baki amino asid tempatan dan berantaian panjang didalam protein adalah kunci penyumbang dalam membentuk struktur sekunder protein. Kerja-kerja terbaru telah tertumpu kepada pengumpulan interaksi amino asid tempatan dan berantaian panjang menggunakan pelbagai ramalan ciri-ciri struktur protein menggunakan penyatuan teknik-teknik pembelajaran mendalam. Walau bagaimanapun, pencarian ciri-ciri struktur ini sentiasa dikaitkan dengan pengkomputeran intensif. Tambahan pula, peramalan prestasi lebih bergantung kepada kualiti ciri-ciri data yang terhasil daripada evolusi protein yang berkaitan. Kajian ini mencadangkan kaedah ramalan struktur sekunder protein dengan menggabungkan Rangkaian Neural Suapan-Kehadapan (FFNN) dengan dwiarah rangkaian Memori Jangka-Pendek yang Panjang (LSTM) untuk mengumpul interaksi amino-asid tempatan dan berantaian-panjang. Untuk meningkatkan lagi ketepatan ramalan protein dengan beberapa evolusi protein yang berkaitan, tambahan ciri-ciri data berdasarkan sifat fizikokimia amino asid telah dicadangkan. Hasil emprikal bagi kaedah yang dicadangkan oleh kajian ini menunjukkan persaingan dalam ketepatan ramalan berbanding Perkakasan Ramalan berdasarkan-Jujukan dalam Talian bagi ciri struktur satu dimensi (SPOT-1D) dan PORTER5. Disamping itu, kaedah ini telah melebihi beberapa kaedah canggih yang terkenal dengan peningkatan 2-3 mata-peratusan. Selain itu, sifat fisiokimia asid ami-

no yang dicadangkan menunjukkan peningkatan ketepatan sebanyak 0.1 mata peratus daripada 82.9%.

PROTEIN SECONDARY STRUCTURE PREDICTION USING ENSEMBLE NEURAL NETWORKS WITH LOCAL AND LONG-RANGE AMINO-ACID FEATURES

ABSTRACT

Predicting protein structures from sequences is a challenging problem. Determining the secondary structures of the protein is an effective approach to infer the complete protein structure. The interactions of local and long-range amino-acid residues in proteins are key contributors in defining the protein secondary structures. Recent works have focused on capturing local and long-range amino-acid interactions using various predicted protein structural features via an ensemble of deep learning techniques. Nevertheless, determining these structural features is always associated with intensive computing. Moreover, their predictive performance is heavily relied on the quality of the data features resulting from evolutionarily related proteins. This study proposes a method for predicting protein secondary structure by incorporating Feed-Forward Neural Network (FFNN) with bidirectional Long Short-Term Memory (LSTM) networks to capture local and long-range amino-acid interactions. To further improve the prediction accuracy of proteins with few evolutionarily related proteins, additional data features based on the physicochemical properties of amino acids have been proposed. The empirical outcomes indicate that the proposed method in this study shows competitive prediction accuracy compared to Sequence-based Prediction Online Tools for one dimensional structural features (SPOT-1D) and PORTER5. In addition to that, the method outperformed several well-known cutting-edge methods by 2-3 percentage-point improvement. Moreover, the proposed amino-acid physiochemical properties showed an increased accuracy by 0.1 percentage points from 82.9%.

CHAPTER 1

INTRODUCTION

1.1 Introduction

From DNA repair to enzyme catalysis, proteins are the chief actors within the cell. They are not just the building blocks of cells; proteins function in all kinds of biological processes in cells. The function of a protein is determined entirely by its amino acid sequence, but the rules that govern how a protein chain of a given sequence folds up are poorly comprehended (Zvelebil & Baum, 2007). With the rapid expansion in the fields of proteomics, an enormous accumulation of protein sequences (more than 190 million) are deposited in protein data banks (MacDougall et al., 2020). Nonetheless, nearly 160 thousand known protein structures are now available (Goodsell et al., 2020). As a result, various studies on protein structure prediction have been conducted to bridge the gap between sequence and structure data via predicting protein structures using available protein sequences and proteins with known structures.

Acquiring knowledge about the structure of proteins from protein sequences is one of the challenges facing bioinformatics (Jiang et al., 2017). One ideal solution to predict the full structure of the protein is to decide its secondary structure elements. However, the secondary structure's pattern rules are not precisely known (Benítez et al., 2011). The secondary structure of the protein denotes the local conformations of a protein segment formed by hydrogen bonds (Voet & Voet, 2011). The formation of all secondary structure elements is guided by local and long-range interactions among amino-acid residues in the proteins (Bruce et al., 2015). The local amino-acid inter-

actions refer to the interactions between the nearest-neighbouring amino-acid residues that are in contact along the protein sequence, while the non-local or long-range amino-acid interactions denote the spatially mediated interactions between the amino-acid residues which are distant along the protein sequence (Ronner, 2017). Details on the biology of the secondary structure elements are presented in Chapter 2 of this thesis.

From the computer science perspective, the protein secondary structure prediction task can be defined as a sequence structural labelling problem: each element (amino-acid residue) of the protein sequence has to be assigned a label (secondary structure class). There are 20 different types of amino acids and eight possible secondary structure classes, composed of three elements of helices (H, G, I), two elements of β-sheets (B, E), and three elements of coils (T, S, L) (Kabsch & Sander, 1983). The prediction of protein secondary structures has been intensively investigated. A common approach to secondary structure prediction is to use traditional machine learning models that use sliding windows with various sizes (Yang et al., 2017; Heffernan et al., 2016; Ma et al., 2018), where a single secondary structure class is predicted by feeding the target amino-acid residue and its surrounding residues through the prediction model. This is then repeated for each amino-acid residue in the protein sequence. These methods are fully reliant on the input windows of adjacent amino-acid residues data, implying they are incapable of wholly learning the relationship between the amino-acid residues in the entire protein sequence. Consequently, the limitation of these methods is that they are unable to capture the long-range interactions among amino-acid residues in the protein sequence.

Several computational methods have been proposed to address the drawbacks of windowing through the use of whole-sequence learning via deep learning techniques (Hanson et al., 2019; Torrisi et al., 2019; Heffernan et al., 2017). These methods offer high prediction accuracy using the prediction of many protein structural features such as backbone angles, solvent accessibility, and contact numbers (Klausen et al., 2019; Hanson et al., 2019). Nevertheless, the prediction of these structural features is always associated with intensive computing and high-complexity processes (Juan et al., 2020); their predictive performance is heavily relied on the quality of the data features resulting from evolutionarily related proteins (Heinzinger et al., 2019). Furthermore, evolutionarily related proteins are not available for many proteins that would be targets of secondary structure prediction (Heinzinger et al., 2019). This circumstance invites more research in finding a prediction method that gives an accurate prediction result and, at the same time, reduces the computational complexity (Shapovalov et al., 2020).

In this study, we propose a method to predict protein secondary structure by incorporating the benefits of both the sliding windows approach and the bidirectional LSTM neural networks to improve the capture of local and long-range amino-acid interactions respectively. The proposed method will take into consideration the high-complexity processes involved in the state-of-the-art secondary structure prediction methods. Furthermore, we propose additional data features based on the physicochemical properties of amino acids to further enhance the prediction accuracy for proteins with few evolutionarily related proteins.

1.2 Motivation

Predicting protein secondary structure accurately is crucial for structural and functional characteristics of proteins. For instance, protein disorder (Hanson et al., 2017), solvent accessibility (Heffernan et al., 2017), protein function (Faraggi et al., 2011), and accurate three-dimensional (3D) structure modelling (Przybylski & Rost, 2004). The prediction of the secondary structure of proteins has an extensive background, starting by the early work on the secondary structures of the backbone of proteins (Pauling et al., 1951; Pauling & Corey, 1951a,b). Despite its long history, the theoretical limits of three-class secondary structure prediction (88–90%) (Rost, 2001) have not been reached yet. Even obtaining improvements of fractions of a percent has become difficult (Yaseen & Li, 2014). The underling challenges in predicting the secondary structure of proteins vary from the capture of local and long-range amino-acid interactions in the protein sequence. Sliding windows approach for protein secondary structure prediction can capture local and 'short to intermediate' long-range amino-acid interactions in protein sequences (Yang et al., 2017). Nonetheless, accounting the long-range amino-acid interactions in the prediction process is essential to improve the prediction performance (Hanson et al., 2017).

The rapid expansion in the field of proteomics has led to an enormous accumulation of protein sequence data (Jiang et al., 2017). This enables deep learning-based methods to complement and outperform other methods for predicting the secondary structure of proteins (Yang et al., 2018). Moreover, in many sequential tasks, learning efficiently in the presence of long-range dependencies is not possible due to the lack of information about which remote sequence positions interact (Ceroni et al., 2005). In

the secondary structure prediction task, dependencies are mainly local, but long-range interactions might significantly help towards accuracy improvements (Busia & Jaitly, 2017). Rost & Sander (2000) showed that the ambiguity in helix and strand predictions is commonly observed in regions that are stabilized by long-range amino-acid residue interactions. LSTM network, a deep learning technique, achieved significant success in various tasks that involve learning from sequences, including natural language processing (Sundermeyer et al., 2012), speech recognition (Amodei et al., 2016), as well as in predicting protein secondary structure (Heffernan et al., 2017). It offers considerable potential over other deep learning techniques in handling long-range dependencies (Hanson et al., 2019; Zhang et al., 2018). The aforementioned observed advantages motivated us to use the LSTM network for secondary structure prediction method proposed in this study.

In addition, the leading development since the inception of secondary structure prediction has been the use of evolutionary information obtained through multiple sequence alignment of evolutionarily related proteins to represent the protein sequences (Walsh et al., 2016). However, this representation is not applicable to all proteins (Heinzinger et al., 2019). Additionally, this representation does not directly allow the extraction of additional information about the proteins, such as hydrogen bonds in the proteins. Thus, prioritizing the development of data features that can enhance the prediction process over the contribution of only evolutionary information signifies a positive approach to improve the prediction accuracy (Hanson et al., 2019). Since protein folding is constrained by physicochemical interactions among amino acids in the protein sequence, the proposed method in this study will consider physicochemical properties of the amino acids in the input data features.

1.3 Problem Statement

Existing successful methods for predicting the secondary structure of proteins are based on various types of deep learning techniques or on standard FFNN that uses sliding windows of stretches of amino-acid residues (Yang et al., 2018; Torrisi et al., 2020). As another traditional machine learning technique used for predicting the secondary structure of proteins, standard FFNN can capture local and 'short to intermediate' long-range interactions among amino acids. Nevertheless, standard FFNN is limited to the window size chosen. Compared to standard FFNN, deep learning-based methods have the capability to capture local and long-range amino-acid interactions in protein sequences (Heffernan et al., 2017; Busia & Jaitly, 2017). However, deep learning-based methods suffer from intensive computing and high-complexity processes due to the use of data features based on various predicted protein structural features for capturing local and long-range amino-acid interactions. Most known protein structural feature prediction methods utilize the evolutionary information-based data features carved into the sequence profiles as their input (Hanson et al., 2019). Such a data feature representation form lacks information about physicochemical interactions inside proteins, such as hydrogen bonds, which could aid the predictor in identifying potential interactions between amino-acid residues and improve protein secondary structure prediction. As a result, when the protein sequence under consideration contains few known evolutionarily related protein sequences, these methods are likely to have lower prediction quality (Wang et al., 2016; Hanson et al., 2019).

In this study, we ask whether we can benefit from both FFNN and bidirectional LSTM neural networks to capture local and long-range amino-acid interactions to

improve the accuracy of secondary structure prediction, taking into consideration the high-complexity processes involved in the state-of-the-art secondary structure prediction methods. Furthermore, when proteins under consideration have sparse sequence profiles, their secondary structure prediction can be further improved by incorporating additional information into their input data features (Heinzinger et al., 2019) . Therefore, we also ask whether the combination of evolutionary information-based data features and data features based on physicochemical properties of amino acids can further enhance the prediction accuracy for proteins with no sufficiently known evolutionarily related protein sequences.

1.4 Research Objectives

This study aims to enhance the performance of secondary structure predictions by incorporating traditional FFNN and bidirectional LSTM neural networks. Emphasis is placed on the investigation of novel secondary structure prediction approach to capture both local and long-range amino-acid interactions to improve the prediction accuracy. The research objectives of this thesis are as follows:

1. To propose a predictive model that incorporates feed-forward neural networks and bidirectional long short-term memory neural networks for protein secondary structure prediction.

2. To propose additional data features based on physicochemical properties of amino acids in order to improve the prediction accuracy for proteins without sufficiently known evolutionarily related protein sequences.

1.5 Overview of Methodology

The outline of the main framework for this research is illustrated in Figure 1.1. Protein sequences and their secondary structure labels are retrieved from Protein Data Bank (PDB). In the preprocessing phase, these files will go through a cleaning process before they are stored in a local database. Referring to the protein sequence database (UniProt), the preprocessed data is transformed into a suitable representation which highlights the evolutionary information features of the considered protein sequences. The resulting evolutionary information-based data features will be utilized to develop the proposed method, which takes advantage of both traditional FFNNs for effectively capturing local amino-acid interactions and bidirectional LSTM networks to improve the capture of local and long-range amino-acid interactions for predicting the secondary structures of the proteins.

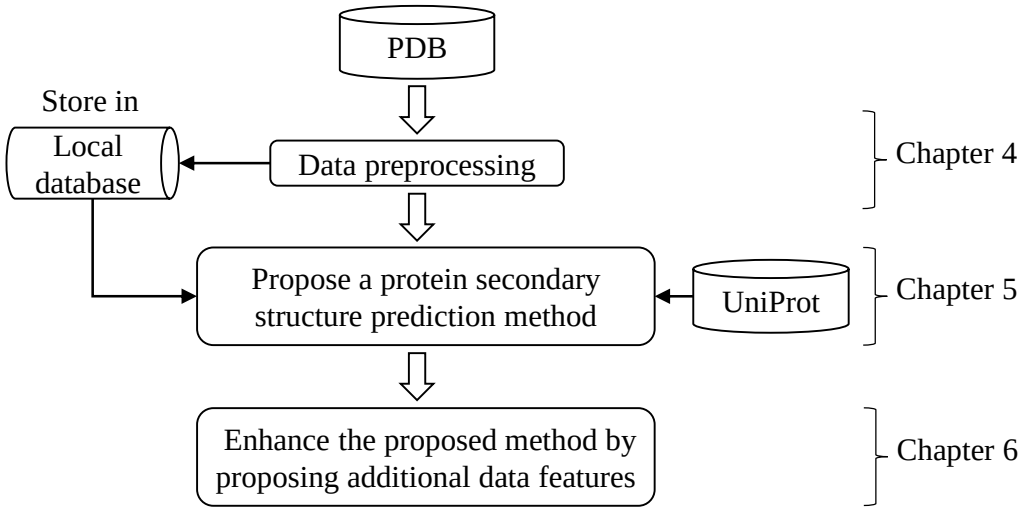


Figure 1.1: The outline of the main framework

Additional data features based on physicochemical properties of the amino acids contained in the protein sequence will be utilized to enhance the predictive performance of the proposed method, mainly when proteins under consideration have not sufficiently known evolutionarily related proteins in the protein sequence database. The proposed method will be evaluated based on performance metrics and comparisons with seven state-of-the-art methods developed for predicting the secondary structure of the proteins.

1.6 Research Contributions

This thesis has led to two main contributions, which are detailed as follows:

1. Developing a new method for protein secondary structure predictions. The developed method incorporated the benefits of both traditional FFNN to effectively capture local amino-acid interactions and bidirectional LSTM to improve the capture of local and long-range amino-acid interactions. In comparison to several existing methods, the proposed method requires fewer procedures to extract evolutionary information from related protein sequences, and uses a smaller number of tuning parameters. In addition, the proposed method managed to reduce erroneous predictions in the boundary region of the secondary structure segments. The proposed method could be employed to predict other proteins' one-dimensional (1D) structural features such as backbone torsion angles, residue accessible surface area, etc.

2. Proposing additional data features based on the physicochemical properties of amino acids in the protein sequence. The quantitative performance comparison

with and without the proposed physicochemical features shows that the proposed method has improved predictive performance not only for proteins without sufficiently known evolutionarily related proteins, but also for proteins with sufficient evolutionary information. The proposed data features could be applied to enhance the prediction performance of other proteins' 1D structural features such as backbone torsion angles, residue accessible surface area, etc.

1.7 Scope of the Research

This study focuses on predicting protein secondary structures from amino acid sequences. The proposed method is conducted for the three- and eight-class secondary structures of globular proteins due to the abundance of globular protein data freely available on public databases. This work compares the performance of the proposed method against publicly available methods. Several sets of protein data were used for evaluation, including datasets from the recent Critical Assessment of Structure Prediction (CASP), CASP12 (Moult et al., 2018) and CASP13 (Kryshtafovych et al., 2019), the latest proteins from the Continuous Automated Model EvaluatiOn (CAMEO) platform (Haas et al., 2018), as well as the latest protein dataset produced using the PISCES webserver (Wang & Dunbrack, 2003) , as described in Chapter 4.

1.8 Thesis Outline and Organization

The thesis is organized in seven chapters. This chapter introduced an overview of the research field. Discussions on research motivation, research problem, research objectives, research contributions, and scope of the study have been highlighted accordingly. The remaining of this thesis is organized as follows:

Chapter 2 presents the research background. It gives a brief introduction to the reader to the commonly used terminologies in biology of proteins. Then, the computational methods for protein secondary structure prediction, including a discussion on the theoretical limit of three-class secondary structure prediction accuracy, are presented. Finally, this chapter provides the terminologies and concepts of neural networks, including neural networks with sliding windows and bidirectional LSTM neural networks, to give a preliminary understanding about the application of neural networks in the area of computational biology.

Chapter 3 presents the literature review. It highlights the two important aspects of protein secondary structure prediction. First, it presents a comprehensive survey of different approaches and techniques used in protein secondary structure prediction. Then, it presents an overview of data features used in protein secondary structure predictions and describes how these data features are combined with different prediction models.

Chapter 4 depicts the research methodology used in this research. This chapter first provides the general research framework in this thesis with a brief description of the proposed methods related to the research objectives. Then, it presents the process of acquiring the datasets, the preparation of the datasets, the performance evaluation, and the benchmark algorithms. This chapter ends with the details of the software and hardware specifications used in this study.

Chapter 5 is dedicated to the first contribution in this thesis, which is the implementation of incorporating the FFNN with sliding window and bidirectional LSTM

networks for protein secondary structure prediction. The experiments on parameter tuning are also presented in this chapter. The proposed method is compared against a set of benchmark algorithms stated in Chapter 4. The experimental results on four protein datasets are presented and discussed at the end of the chapter.

Chapter 6 presents the second contribution: the use of amino-acid physiochemical properties to enhance prediction of the secondary structures of proteins with no sufficiently known evolutionarily related protein sequences. The relevant experimental results are compared against a set of benchmark algorithms and discussed at the end of the chapter.

Chapter 7 presents the conclusion and highlights some future directions of the research.

CHAPTER 2

BACKGROUND

2.1 Introduction

This chapter consists of three main parts. The first part presents a brief introduction to the basic concepts and the commonly used terminologies in the biology of proteins and highlights the role of local and long-range amino-acid interactions in the secondary structure formation. The second part covers the computational methods for protein secondary structure prediction, including a discussion on the theoretical limit of three-class secondary structure prediction accuracy. The last part of this chapter provides the commonly used terminologies and the basic concepts of neural networks, including neural networks with sliding windows and bidirectional LSTM neural networks.

2.2 Brief Biology of Proteins

In molecular biology, the central dogma describes the information flow from DNA to mRNA and finally to proteins (Zvelebil & Baum, 2007). Genes in DNA are transcribed into messenger RNAs and then translated into proteins. This section provides a brief overview of proteins, starting by clarifying some fundamental definitions of proteins. Subsequently, protein structure levels and the challenges of determining protein structures are briefly summarized. Finally, further elaborations on the protein secondary structure are highlighted.

2.2.1 Proteins

Proteins are biological macromolecules that are vital to support cell activities in living organisms. Proteins can act as transporting molecules (Klingenberg, 1981), responding to stimuli (Yoshida et al., 2005), catalysing metabolic reactions necessary for life possible (Margolis, 2008), storages to store nutrients and energy-rich molecules for later use (Zvelebil & Baum, 2007). Protein structure consists of amino acids that are connected to each other by one or more polypeptides chains. There are 20 different amino acids that take different combinations and different lengths to form a protein. These 20 amino acids which can be denoted using single alphabet or three alphabets are shown in Table 2.1. Amino acid denoted by single alphabet is known as a residue. A protein sequence can thus be seen as a string of characters drawn from these 20-membered residues.

Table 2.1: Names of amino acids and its respective three- and one-letter code.

Amino acid	Three-letter code	One-letter code
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic	Gln	Q
Glutamine acid	Glu	E
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

The chemical structure of an amino acid comprises a central α-carbon (C_α) atom, a carboxyl group (-COOH) and an amino group (-NH_2) on both ends, and a variable side-chain (-R) as illustrated in Figure 2.1. The side-chains are unique between different amino acids and can affect the physiochemical properties of the amino acid, such as mass, polarity, acidity, hydrophobicity, and electron charge. The functional properties of proteins are almost entirely due to the side-chain interactions of the amino-acid residues (Zvelebil & Baum, 2007). Although a protein can be simply thought of as a string of amino-acid residues, it folds into a unique, compact and stable 3D structure under physiological conditions to perform its biological functions.

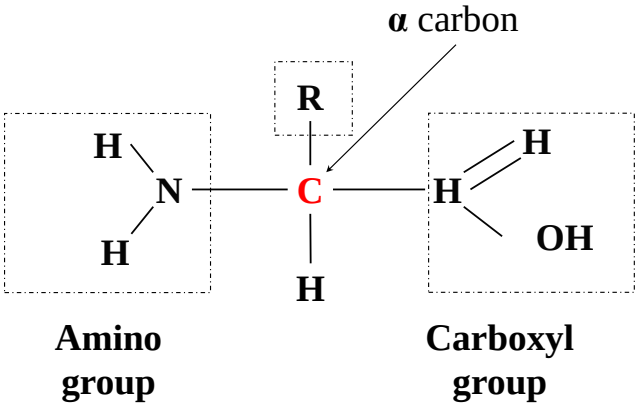


Figure 2.1: Chemical structure of a single amino acid

2.2.2 Protein Structure Levels

A protein molecule is formed from a chain of amino-acid residues sequence that folds into a complex 3D structure. The 3D structure of proteins comprises four levels (*primary structure, secondary structure, tertiary structure, and quaternary conformation*), as shown in Figure 2.2. The **primary structure** refers to the specific amino-acid residue sequence of the protein, plus the order of these amino acids along a polypeptide chain. The **secondary structure** is the first level of protein folding, in which parts

of the polypeptide chain fold to form generic structures that are found in all proteins. The **tertiary structure** is produced by the additional combination, folding, and packing together of these elements to ultimately provide 3D conformation that is exclusive to the protein. Many functional proteins are formed of more than one chain, in which case the individual chains are called protein subunits. The subunit composition and arrangement in such multi-subunit proteins is called the **quaternary conformation**.

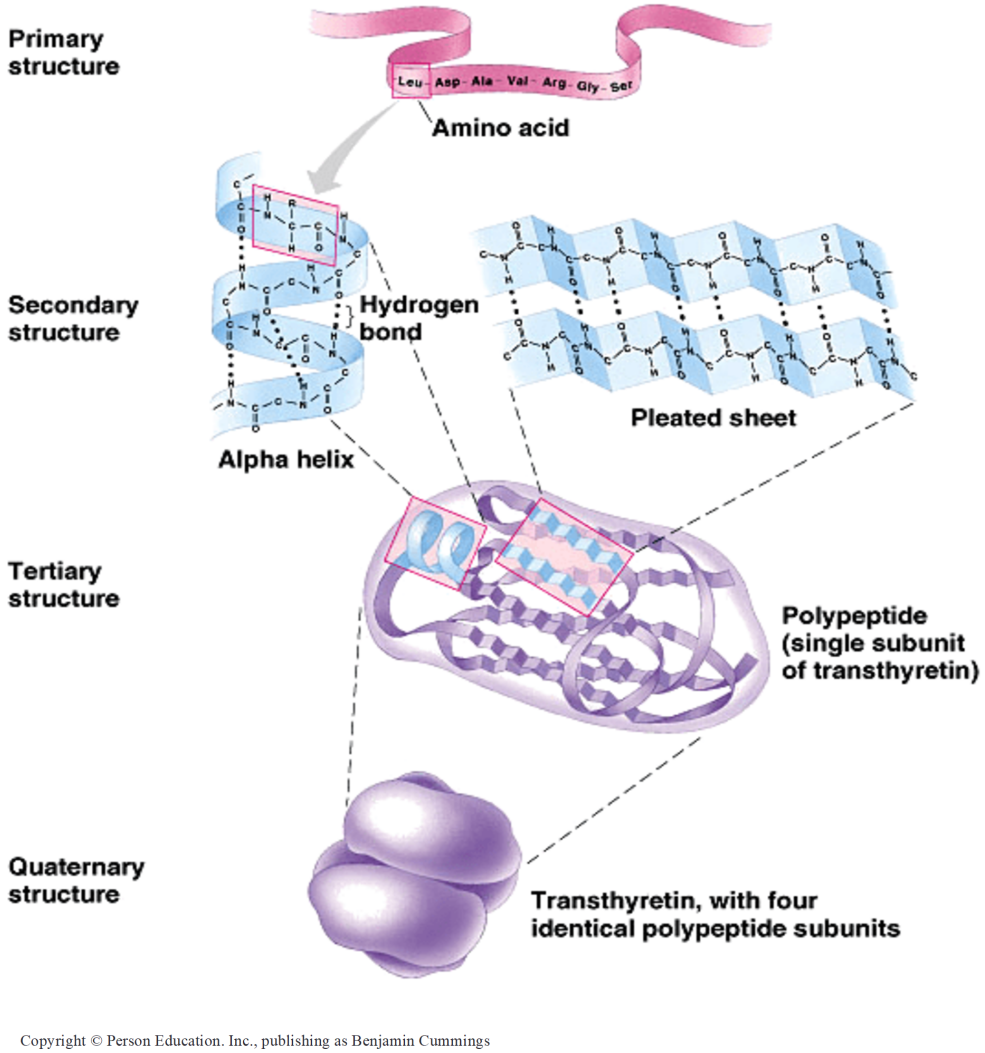


Figure 2.2: Protein structure levels, adopted from Campbell et al. (2007).

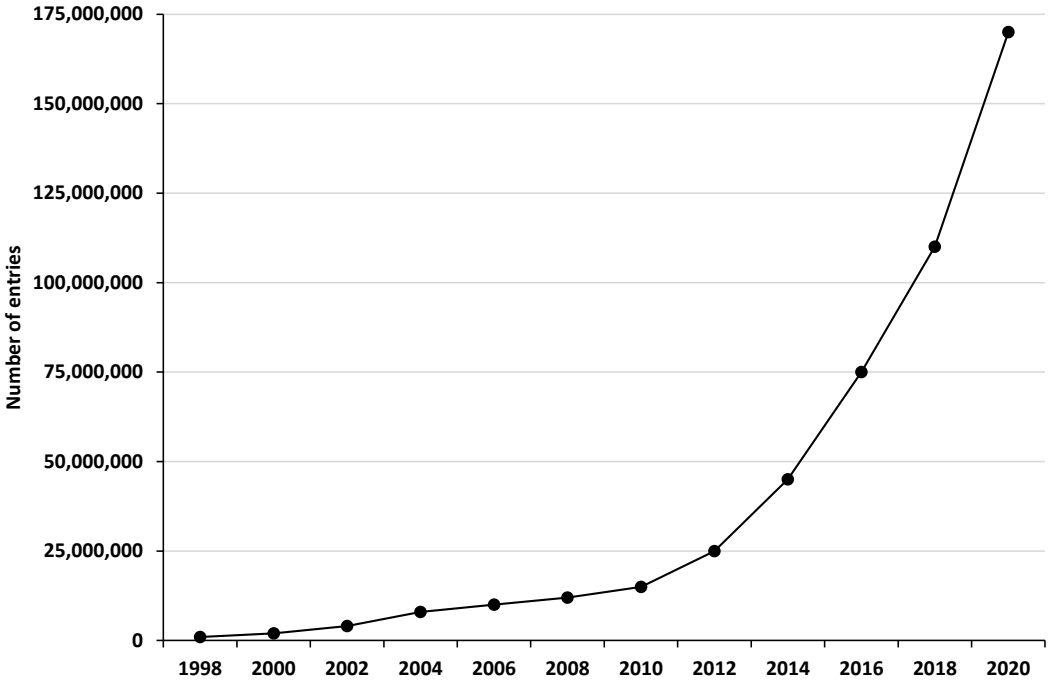
Numerous interactions between protein atoms stabilize the 3D structure of proteins (Ronner, 2017). Interactions between protein atoms can be either covalent or non-

covalent. A covalent bond forms when two atoms come very close together and share one or more of their electrons. Non-covalent interactions are short-range attractive forces and can take place by different means, such as **hydrogen bonding** or **van der Waals** or **electrostatic interactions** (Zvelebil & Baum, 2007).

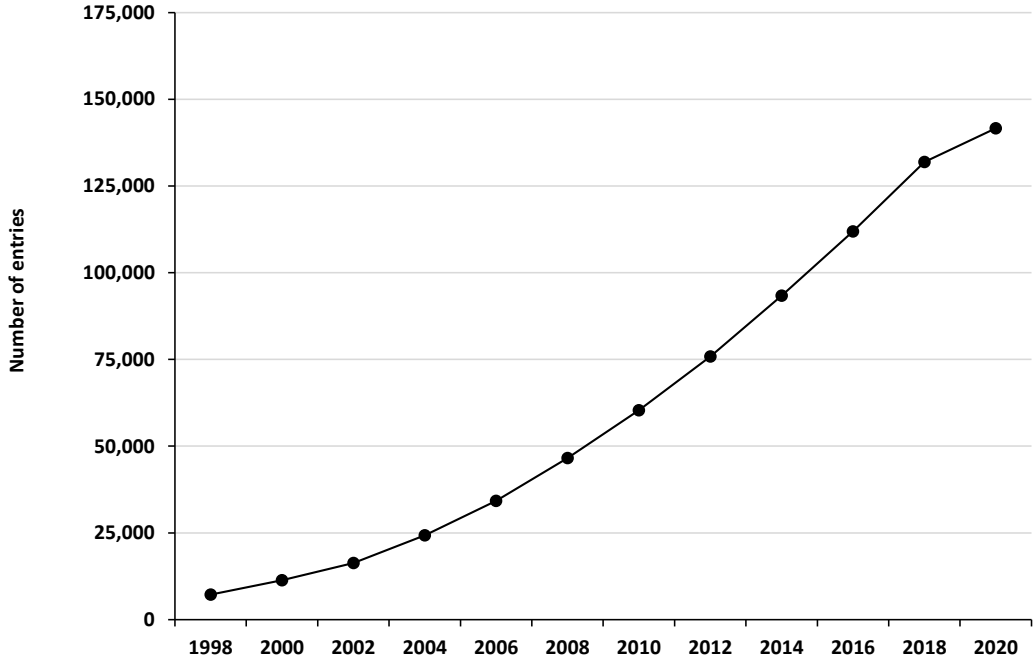
Hydrogen bonds form when there is a weak sharing of electrons between a hydrogen atom (donor) attached to a nitrogen or an oxygen and an acceptor atom, which in proteins is usually a nitrogen or an oxygen, although other atoms can act as acceptors as well (Baker & Hubbard, 1984). The peptide backbone contains hydrogen donors in the -N-H and hydrogen acceptors in the =O of each amide group. However, for steric reasons, the donor an acceptor must come from different amino acid residues (Bruce et al., 2015). The amino acid side chains contain the hydrogen donors -O-H (in Ser, Thr, Tyr), -N-H (in Trp, Asn, Gln, Lys, His, Arg), -S-H (in Cys), as well as the hydrogen acceptors =O (in Asn, Gln, Asp, Glu), -O (in Ser, Thr, Tyr), and -N (in His) (Ronner, 2017). *Electrostatic interactions* occur between groups that are of opposite charge. Atoms of the peptide backbone are essentially uncharged, except for the amino and carboxyl termini (Bruce et al., 2015). The amino acid side chains of Asp and Glu are negative charged, while those of His, Lys and Arg are positively charged (Zvelebil & Baum, 2007). Moreover, atoms can be thought of as a nucleus surrounded by a cloud of constantly moving electrons. Sometimes there are more electrons at one end of the atom than at the other; this forms a dipole. A *van der Waals interaction* occurs when atoms that have oppositely oriented dipoles are near each other (Zvelebil & Baum, 2007). The contact distance between any two non-covalently bonded atoms is the sum of their van der Waals radii (Bruce et al., 2015).

The function of a protein is determined entirely by its amino acid sequence, but the rules that govern how a protein chain of a given amino acid sequence folds up into the 3D structure are poorly comprehended (Zvelebil & Baum, 2007). To determine the 3D structure of proteins, a lot of efforts have been devoted to developing experimental methods, including electron microscopy, Nuclear Magnetic Resonance (NMR) spectroscopy, and X-ray crystallography. Nonetheless, due to costly experimental methods for determining protein structures, the number of known protein structures is much less than the number of known protein sequences, as shown in Figure 2.3. As a result, various studies on protein structure prediction have been conducted to bridge the gap between sequence and structure data via predicting protein structures using available protein sequences and proteins with known structures (Jumper et al., 2021).

However, direct prediction of the 3D structure of proteins from their sequence data remains a key challenge (Jiang et al., 2017). A preferred approach to resolving this prediction difficulty involves breaking down the problem into smaller structural problems. A number of these structural problems can be symbolized as 1D vectors along the protein sequence. Commonly used 1D structural features of proteins are secondary structures, backbone torsion angles, solvent accessibility, and residue contact numbers (Voet & Voet, 2011). Since this thesis focuses on secondary structure prediction, the discussion will focus more on the secondary structure.



(a) Growth of data points in Uniprot. The number of datapoints (in millions) was used; every subsequent point represents a novel Uniprot release.

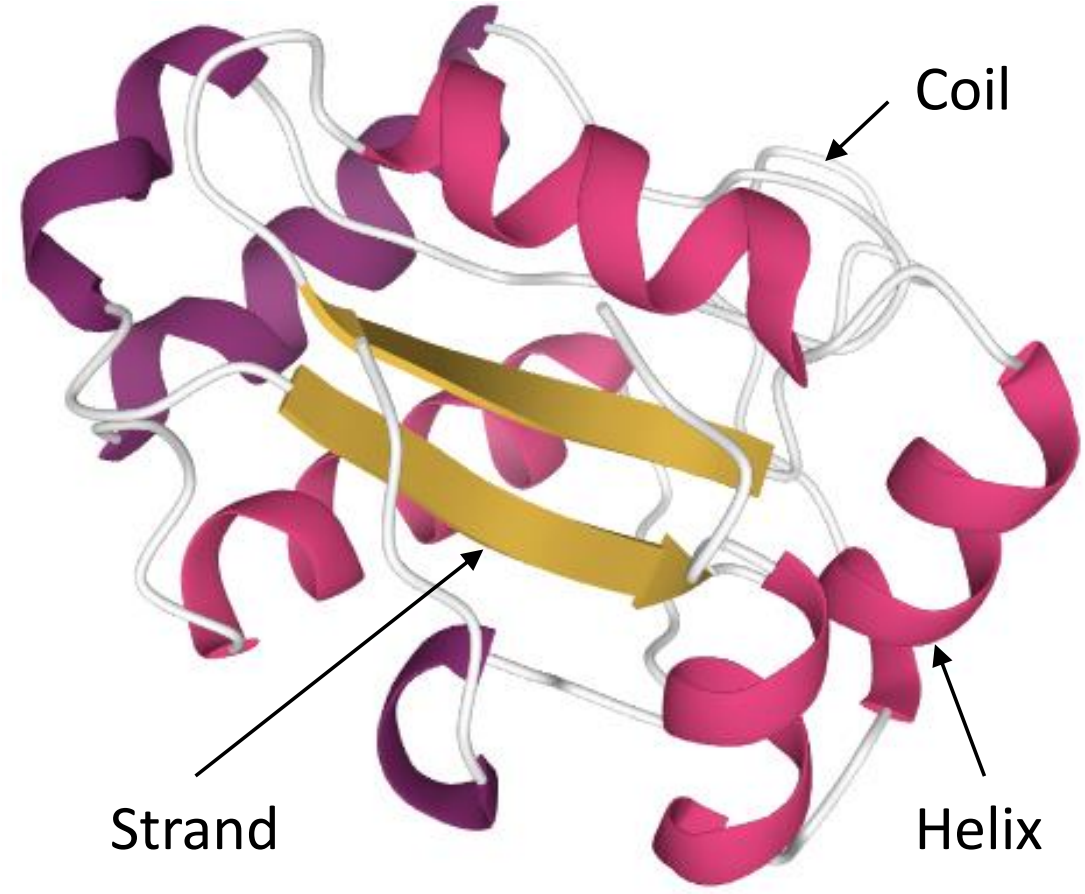


(b) Growth of data points in PDB. The number of datapoints (in thousands) was used; every subsequent point represents a novel PDB release.

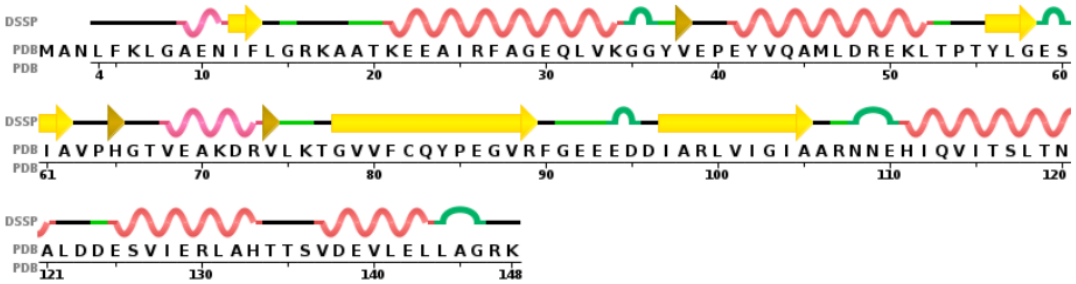
Figure 2.3: Protein databases growth. (a) growth of protein sequences. (b) growth of protein structures.

2.2.3 Protein Secondary Structure

Protein secondary structure plays a critical role in the study of proteins (Wang et al., 2017). It comprises the primary fold of the polypeptide chain. Thus, the spatial structure of the protein is based on its secondary component. The secondary structure of the protein denotes the local conformations of a protein segment formed by hydrogen bonds (Voet & Voet, 2011). As shown in Figure 2.4, there are two regular secondary structure patterns: α-helix (H) and β-strand (E) and one irregular secondary structure pattern: coil region (C). Helices arise as a result of energetically favourable hydrogen bonding between atoms of the backbone of the protein chain (Zvelebil & Baum, 2007). Within the helix, all the backbone groups that can be involved in hydrogen bonding. This gives a very regular and stable arrangement (see Figure 2.4). Strands are another common regular repeating structure found in globular proteins that consisted of extended strands aligned with each other to permit favourable hydrogen bonding (Zvelebil & Baum, 2007). A single extended chain of this type is called a β-strand, and a set of β-strands hydrogen bonded together side by side forms a β-sheet (Figure 2.4). Other structures which cannot be classified as one of the standard two classes is usually grouped into a category called "Coils". It should be noted that the structures found in proteins are not perfectly regular, so it is frequently difficult to define the precise ends of the structure, and in some cases, the hydrogen-bonding patterns are intermediate between these idealized forms (Zvelebil & Baum, 2007). Therefore, prediction of these structures using bioinformatics programs is made more difficult.



(a) The three-dimensional structure, highlighting helices, strands and coils.



(b) The amino acid sequence and the respective DSSP-assigned secondary structures.

AA:	K	L	T	P	T	Y	L	G	E	S	I	A	V	P	H	G	T	V	E	A	K	D	R	V	L	K	T	G	V	V
SS8:	H	H	S	L	L	E	E	E	T	T	E	E	L	L	B	L	L	G	G	G	G	G	G	B	L	L	L	E	E	E
SS3:	H	H	C	C	C	E	E	E	C	C	E	E	C	C	E	C	C	H	H	H	H	H	H	E	C	C	C	E	E	E

(c) The amino acid sequence from 51-81 residues and the respective eight- to a three-class mapping of the DSSP-assigned secondary structures.

Figure 2.4: Graphical representation of the Escherichia coli phosphotransferase IIA-mannitol (PDB ID: 1a3a chain A, 148 amino-acid residues). The figure created using Mol* (Sehnal et al., 2021)

Furthermore, the Define Secondary Structure of Proteins (DSSP) program defines the secondary structure into eight fine-grained elements: three elements for helix (α-helix (H), 3_{10}-helix (G), and π-helix (I)), two elements for strand (β-sheet (E) and β-bridge (B)), and three elements for coil (Bend (S), Turn (T), and others or loops (L)) (Kabsch & Sander, 1983). Most of the developed methods for secondary structure prediction are trained and evaluated for only three classes, so the eight classes are mapped to three. Several different rules have been used to conduct the mappings, as shown in Table 2.2. This work follows Rule #3, the hardest three-class rule to predict, while the easiest three-class rule follows Rule #1 (Shapovalov et al., 2020). Rule #3 treats α-helix (H), 3_{10}-helix (G), and π-helix (I) as 'helix', β-sheet (E) and β-bridge (B) as 'strand', and all other classes treated as 'coil'.

Table 2.2: Eight-class to three-class reduction Rules

Rule	DSSP-assigned class to three-class reduction
Rule #1	H to H, E to E. Rest to coil.
Rule #2	H and G to H, E and B to E. Rest to coil.
Rule #3	H, G and I to H, E and B to E. Rest to coil.
Rule #4	H and G to H, E to E. Rest to coil.
Rule #5	H, G and I to H, E to E. Rest to coil.

Application of alternative eight- to three-class mappings shows a deviation in the prediction accuracy of more than three percent (Shapovalov et al., 2020; Cuff & Barton, 2000). It is therefore necessary to assess the accuracy of a set of predictions to determine the most accurate method available, as well as to identify successful improvements in developing new methods. Alternative measures have been proposed, some looking at the level of individual amino-acid residues, such as Q_3 accuracy for three-class secondary structures and Q_8 accuracy for eight-class secondary structures, others focusing on the complete secondary structure elements, such as SOV_3 and SOV_8

(Zemla et al., 1999) for three- and eight-class secondary structures, respectively. The aforementioned accuracy measures will be described later in Section 4.5 in Chapter 4.

Table 2.3 shows a comparison of two prediction accuracy measures, Q_3 and SOV_3. Five possible predictions of a single observed helix are shown. Predictions 1 and 2 have the same number of helical amino-acid residues, and yet 1 is completely unrealistic. Although SOV_3 registers this, Q_3 does not. Predictions 2, 3, 4, and 5 demonstrate that SOV_3 gives higher scores when only one helix is predicted even if fewer individual amino-acid residues are correctly predicted to be helical, contrary to the behaviour of Q_3 (Zemla et al., 1999). In general, both measures are only meaningful if they refer to the prediction of a large number of nonhomologous structures because the methods are intended to be widely applicable.

Table 2.3: A comparison of two prediction accuracy measures, Q_3 and SOV_3

	Secondary structure	Q_3%	SOV_3%
Observed	C**HHHHHHHHHH**C		
Prediction 1	C**H**C**H**C**H**C**H**CC	58.3	12.5
Prediction 2	CCC**HHHHH**CCCC	58.3	63.2
Prediction 3	C**HHH**C**HHH**C**HH**C	83.3	40.6
Prediction 4	C**HH**CC**HHHHH**CC	75.0	52.3
Prediction 5	CCC**HHHHHH**CCC	66.7	80.6

The formation of all secondary structure classes is guided by local and long-range interactions of amino-acid residues in the proteins (Bruce et al., 2015). The local amino-acid interactions refer to the interactions between the nearest neighbouring amino-acid residues that are in contact along the protein sequence, while the long-range amino-acid interactions denote the spatially mediated interactions between the amino-acid residues which are distant along the protein sequence (Ronner, 2017). For example, a β-sheet (also called a pleated sheet) is composed of two or more strands

that are held together by hydrogen bonds, but that can be situated very far apart in the protein sequence. As shown in Figure 2.5, the amino-acid residues of the pleated sheet in this protein are spatially close in 3D but have distant positions in sequence. A similar effect is observed for cysteines that are linked by disulfide bonds (Bruce et al., 2015). The problem, then, is how to transfer proteins to computers and how to make computers understand them in order to predict the secondary structures, taking into account the local and long-range interactions among amino-acid residues in the proteins. It has been discussed that one of the main reasons for secondary structure prediction limitations conceivably comes from long-range amino-acid interactions, which may overwrite local sequence propensity of secondary structures (Heffernan et al., 2017).

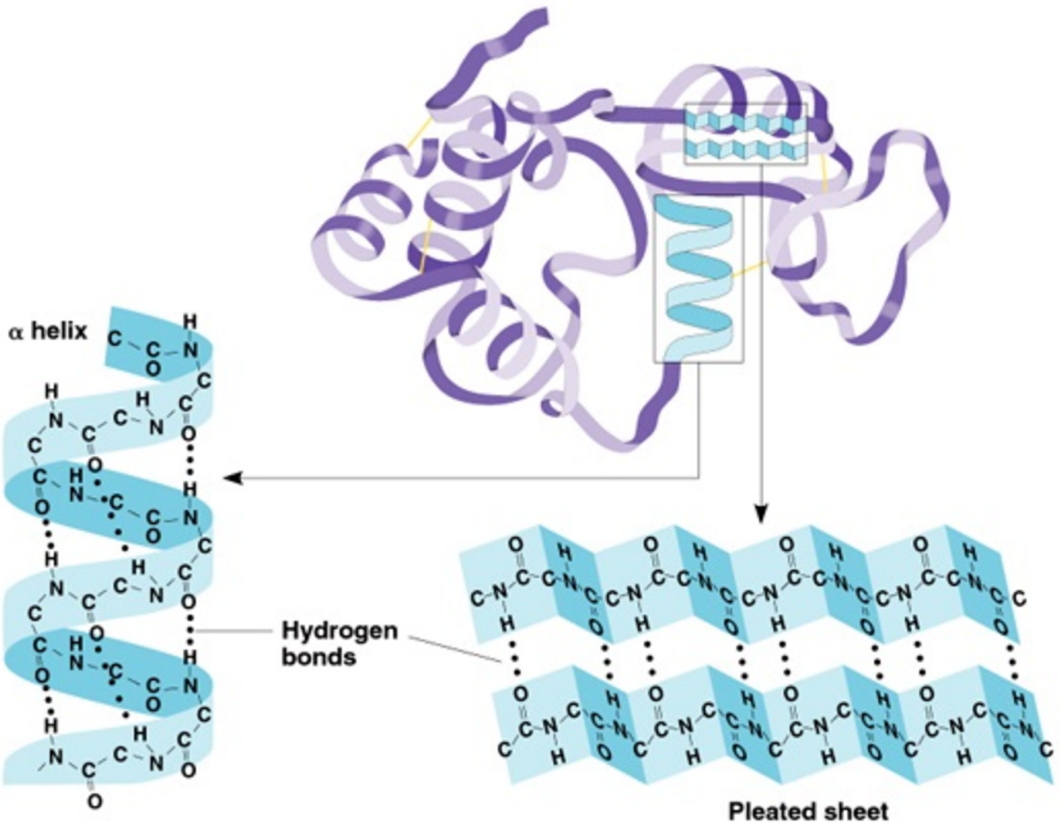


Figure 2.5: Secondary structure elements in proteins , adopted from Campbell et al. (2007).