Contents

[Abstract 2](#_Toc366513020)

[1. Introduction 2](#_Toc366513021)

[2. Biological Background 2](#_Toc366513022)

[2.1 Nucleotide 2](#_Toc366513023)

[2.2 Deoxyribonucleic acid (DNA) 3](#_Toc366513024)

[2.3 Amino Acids 3](#_Toc366513025)

[2.4 Proteins 5](#_Toc366513026)

[2. Information Theory Backgrounds 7](#_Toc366513027)

[2.1 Shannon Entropy 7](#_Toc366513028)

[2.2 Joint Entropy 7](#_Toc366513029)

[2.3 Conditional Entropy 7](#_Toc366513030)

[2.4 Mutual Information 8](#_Toc366513031)

[2.5 Normalized Mutual Information 8](#_Toc366513032)

[2.6 Kullback Leibler Divergence 8](#_Toc366513033)

[2.7 Jensen Shannon Divergence 8](#_Toc366513034)

[3. Methods 9](#_Toc366513035)

[3.1 Random Forests 9](#_Toc366513036)

[3.2 Information Theory based Signal Amplifying Framework 9](#_Toc366513037)

[3.3 DNA-binding Proteins Predictor 9](#_Toc366513038)

[4. Results and Evaluations 9](#_Toc366513039)

[4.1 Data Preparation 9](#_Toc366513040)

[4.2 Results & Evaluations 9](#_Toc366513041)

[5. Conclusions 9](#_Toc366513042)

[6. Discussions 9](#_Toc366513043)

[Bibliography 9](#_Toc366513044)

# Abstract

# Introduction

# Biological Background

In this section, we briefly presents the introductions of some biological term and phrases related to this thesis. After that, we mention a short overview of biological mechanisms which motivate us conduct this project. And last but not least, some useful tools, databases as well as web services relating to bioinformatics will be explained in detail.

## 2.1 Nucleotide

Nucleotides are biological molecules which are the building blocks of DNA or RNA. A typical nucleotide consists of a sugar (deoxyribose), a phosphate and a base (Figure 2.1). There are four types of nucleotides in DNA which are different due to their bases including Adenine(A), Guanine(G), Cytosine(C), and Thymine(T) (for RNA, Uracil(U) replaces Thymine(T)).

Table 2.: Full name and abbreviation of nucleotides

|  |  |
| --- | --- |
| Full name of base/nucleotide | Abbreviation |
| Adenine | A |
| Guanine | G |
| Cytosine | C |
| Thymine | T |
| Uracil | U |

Sugar

Base

Figure 2.: Structure of nucleotides

Further, Cytosine and Thymine (or Uracil in RNA) belongs to pyrimidine bases while Adenine and Guanine are in purine bases. Adenine always pairs with Thymine by 2 hydrogen bonds, while Guanine pairs with Cytosine through 3 hydrogen bonds, each due to their unique structures.

## 2.2 Deoxyribonucleic acid (DNA)

Deoxyribonucleic acid (DNA for abbreviation) is the macromolecule encoding genetic information. It is used in all biological mechanisms of all known living forms and virus. DNA is firstly isolated and identified by Friedrich Miescher in late 1890s and its double helix structure is discovered by James Watson and Francis Crick in 1950s. Roughly speaking, DNA comprises of two complement sequences of nucleotides in which pairs A, T and pairs G, C binding together (Figure 2.2[[1]](#footnote-1)).

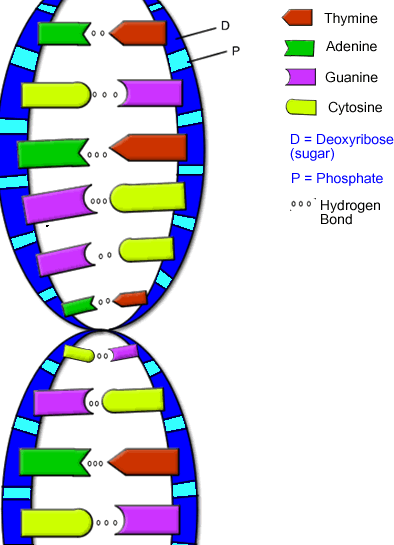


Figure .2: Structure of DNA

## 2.3 Amino Acids

Amino acids are the production of RNA translation which is a part of gene expression. In RNA-translation procedure, a triple nucleotides will be translated into one amino acid. An amino acid consists of central carbon atom (), a hydrogen atom (), free carboxyl group (), free amino group (), and side chain group (). The R group creates the distinction among amino acids due to its chemical structure.

H

COOH

R

Figure 2.: General structure of amino acid

Overall, there are 20 primary amino acids, and they are classified into two main groups due to their physical and chemical properties.

* Hydrophobic groups:
  + Nonpolar-aliphatic R groups: Glycine, Alanine, Proline, Valine, Leucine, Isoleucine, Methionine.
  + Aromatic R groups: Phenylalanine, Tyrosine, Tryptophan.
* Hydrophilic groups:
  + Polar-charged R groups: Serine, Threonine, Cysteine, Asparagine, Glutamine.
  + Positively charged R groups: Lysine, Arginine, Histidine.
  + Negatively charged R groups: Aspartate, Glutamate.

Table .2: Names and abbreviations of 20 types of amino acids

|  |  |  |  |
| --- | --- | --- | --- |
| Number | Full name | Three letters abbreviation | One letter abbreviation |
| 1 | Alanine | Ala | A |
| 2 | Arginine | Arg | R |
| 3 | Asparagine | Asn | N |
| 4 | Aspartic | Asp | D |
| 5 | Cysteine | Cys | C |
| 6 | Glutamic | Glu | E |
| 7 | Glutamine | Gln | Q |
| 8 | Glycine | Gly | G |
| 9 | Histidine | His | H |
| 10 | Isoleucine | Ile | I |
| 11 | Leucine | Leu | L |
| 12 | Lysine | Lys | K |
| 13 | Methionine | Met | M |
| 14 | Phenylalanine | Phe | F |
| 15 | Proline | Pro | P |
| 16 | Serine | Ser | S |
| 17 | Threonine | Thr | T |
| 18 | Tryptophan | Trp | W |
| 19 | Tyrosine | Tyr | Y |
| 20 | Valine | Val | V |

Moreover, all over 20 amino acids are necessary for protein synthesis to maintain the body growth and function. Among 20 amino acids, there are eleven ones (Alanine, Arginine, Asparagine, Aspartic, Cysteine, Glutamic, Glycine, Proline, Serine, Tyrosine) called non-essential amino acids because they can be synthesized within body. And the rest of nine amino acids are essential ones and only could be taken from additional nutrition [[1](#BHa04)].

## 2.4 Proteins

Protein is a large biological molecule consisting of several chains of amino acids. An individual chain of these amino acids has a specific spatial arrangement driven by non-covalent interactions such as hydrogen bonding, van der Waals forces, and hydrophobic packing. These spatial arrangements play the key role to define structure and functions of protein.

In general, a certain protein is described via four structural levels including primary, secondary, tertiary, and quaternary structure (figure 2.4[[2]](#footnote-2)).

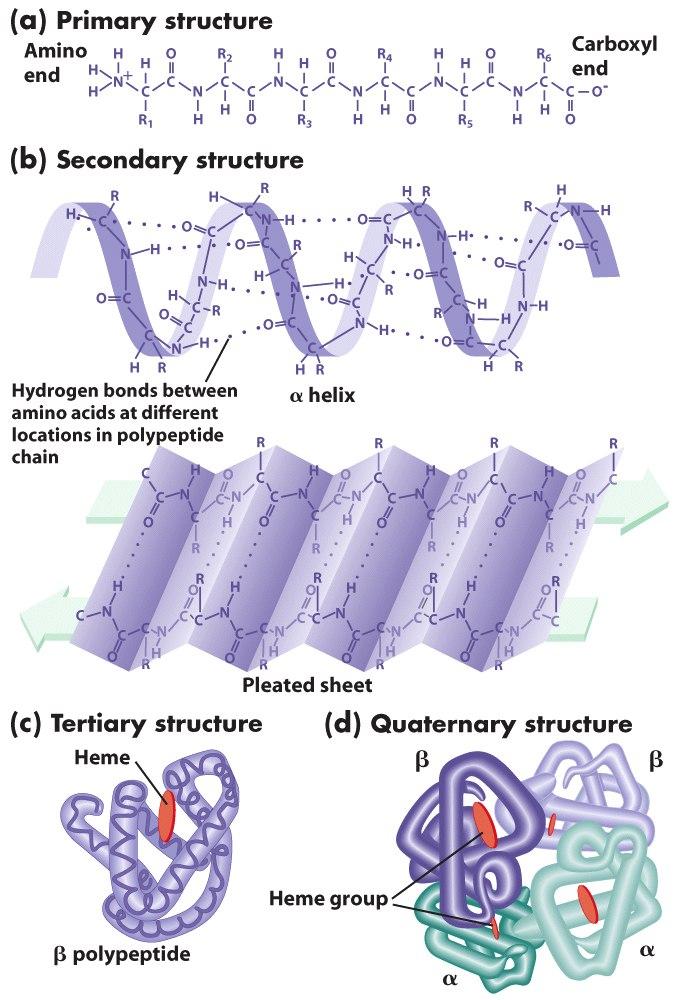
* Primary structure: the linear sequences of amino acids in polypeptide chain.
* Secondary structure: is highly regular local structures consisting of α-helices and β-strand (pleated sheet) driven by hydrogen bonds.

Figure 2.: a) primary structure: sequence of amino acids in chain. b) secondary structure: α-helix and β-strand (pleated sheet). c) tertiary structure: 3D structure of single chain of amino acids. d) quaternary structure: a complex of several protein chains.

* Tertiary structure: a three dimensional structure of protein chain, including secondary structure.
* Quaternary structure: a complex consisting of tertiary structure of several chains to describe a functional protein.

## 2.5 Bioinformatics tools and databases

This section would briefly present some common and useful databases in which all properties of proteins such as three dimensional structure and binding indexes can be found there. In addition, some web services used to detect protein DNA binding residual will be introduced as well.

### 2.5.1 Protein Data Bank[[3]](#footnote-3)

Protein data bank (PDB for abbreviation) is a worldwide collection of three dimensional structures of large biological molecules such as proteins and nucleic acids. This repository includes more than 86000 proteins, 2500 nucleic acids, and 4000 complexes. In particular, all identified proteins are indicated via four characters (A to Z or 0 to 9) such as “1V4S” for human glucokinase protein. In addition, in order to present the structural information of proteins, Proteins Data Bank has a conventional format called “pdb” format which allows researchers retrieve information of atomic coordinates, literature references and experimental details as well. Moreover, primary and secondary structure information such as disulfide bonds, helices, strand, binding sites, active sites are also presented in this format.

In sort, every protein available in PDB is formatted under extension “pdb” in which primary, secondary and tertiary structure information is recorded. Moreover, there are quite a bunch of records in PDB files which are explained in detail in [[2](#PDB)]. In the scope of this dissertation, we will briefly mention some important records which are used in feature extraction process.

#### HEADER and TITLE records

|  |
| --- |
| HEADER TRANSCRIPTION/DNA 27-NOV-97 1A0A  TITLE PHOSPHATE SYSTEM POSITIVE REGULATORY PROTEIN PHO4/DNA |

The HEADER tag consists of classification for the entry, the date when the coordinates were composited to PDB archive, and a unique identified PDB entry. And the TITLE record describes the name of experiment or analysis represented in the entry.

#### HELIX and SHEET records

|  |
| --- |
| HELIX 36 36 SER H 88 LEU H 98 1 11  HELIX 37 37 GLY H 101 THR H 119 1 19  SHEET 1 A 2 ARG A 83 PHE A 84 0  SHEET 2 A 2 THR B 80 VAL B 81 1 N VAL B 81 O ARG A 83 |

HELIX and SHEET record the secondary structure of the protein. Firstly, in HELIX records, the beginning and ending positions are located as well as their chain identity and types. The total length of helix is also mentioned at the end of record. Secondly, SHEET records describe another type of secondary structural information of protein. It records the begin and end position of β-strand as well as the length of this strand.

#### ATOM records

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Pos. | 1-6 | 7-11 | 13-16 | 17 | 18-20 | 22 | 23-26 | 27 | 31-38 | 39-46 | 47-54 | 55-60 | 61-66 | 77-78 | 79-80 |
| Eg. | ATOM | 1 | N | A | ARG | A | 14 |  | 11.281 | 86.699 | 94.383 | 0.50 | 35.88 | N |  |
| Desc. |  | | | | | | | | | | | | | | charge |
|  | | | | | | | | | | | | | Element symbol | |
|  | | | | | | | | | | | | Temperature factor | | |
|  | | | | | | | | | | | Occupancy | | | |
|  | | | | | | | | | | Z coordinate | | | | |
|  | | | | | | | | | Y coordinate | | | | | |
|  | | | | | | | | X coordinate | | | | | | |
|  | | | | | | | Code for insertion of residues | | | | | | | |
|  | | | | | | Residue sequence number | | | | | | | | |
|  | | | | | Chain identifier | | | | | | | | | |
|  | | | | Residue name | | | | | | | | | | |
|  | | | Alternate location indicator | | | | | | | | | | | |
|  | | Atom name | | | | | | | | | | | | |
|  | Atom serial number | | | | | | | | | | | | | |
| Record name | | | | | | | | | | | | | | |

The ATOM records consist of three dimension coordinate of atom in nucleic acid or amino acid. From this coordinate information, we can compute distance between any pair of atom, and easily detect binding residuals.

### 2.5.2 The HSSP database

# 3. Information Theory Backgrounds

In this section, we recapitulate some fundamental concepts in classic information theory. Firstly, we will provide some basic definitions such as Shannon entropy, joint entropy and mutual information. Then the next subsections discuss some divergent measurements which allow us to measure how much difference between two distributions.

## 3.1 Shannon Entropy

Given a random variable , Shannon Entropy is defined by:

Where is all possible outcomes, and p(x) = P(X=x) with . The entropy is maximal if is under uniform randomly distribution.

Moreover, there are several interpretations of Shannon entropy which belong to two main categories. The first one interprets Shannon entropy as a capacity in information channel as well as limitation of compression ratio. Another point of view is to consider Shannon entropy is the minimal number of bits (or yes/no questions) required to eliminate the uncertainty.

## 3.2 Joint Entropy

Given a pair discrete random variables () with a joint distribution , then joint entropy is defined:

Where χ, γ are outcome space of random variable and .

## 3.3 Conditional Entropy

Given a pair discrete random variables () with a joint distribution and conditional distribution , then the conditional entropy is defined as:

So, the conditional entropy is:

Where χ, γ are outcome space of random variable and .

When and are completely independent, the .

## 3.4 Mutual Information

Suppose we have two random variable and , and we would like to have a mean to measure the amount of mutual dependency between and . Thus, mutual information is a facility to measure how much and is different from each other.

So, the definition of mutual information is following:

There are two notices from above formulas. Firstly, mutual information will approach zero if and totally independent. Secondly, if and are completely correlated, we have .

## 3.5 Normalized Mutual Information

## 3.6 Kullback Leibler Divergence

Kullback Leibler divergence is a non-symmetric measure of the difference between two probability distributions  and . Its formula is given by (assume and are concrete probability distribution):

The value of Kullback Leibler divergence could diverse from zero (when P ≈ Q) to infinitive. Moreover, this divergence could be interpreted as the information lost when is used to approximate .

## 3.7 Jensen Shannon Divergence

Jensen-Shannon divergence is yet another popular method to measure the different between two different probability distributions. In contrast to Kullback-Leibler divergence whose value could go to infinitive, Jensen-Shannon ensures the range of score lie between zero and one; thus it enables us apply some threshold detection method. Overall, the formula of this divergence is defined by:

Where is Kullback-Leibler divergence and .

This divergence is symmetric and approach zero when P and Q are equivalent while get value of one when P and Q completely independent.

# 4. Methods

## 4.1 Random Forests

## 4.2 Information Theory based Signal Amplifying Framework

## 4.3 DNA-binding Proteins Predictor

# 5. Results and Evaluations

## 5.1 Data Preparation

## 5.2 Results & Evaluations

# 6. Conclusions

# 7. Discussions

# Bibliography

x

|  |  |
| --- | --- |
| [1] | B. Hansen, L. B. Jorde, "Lecture Notes Biochemistry," *KAPLAN medical*, 2004. |
| [2] | *Protein Data Bank Contents Guide: Atomic Coordinate Entry Format Description*, 320th ed.: wwPDB, 2012. |

x

1. <http://jameschung3b.edublogs.org/2011/11/28/dna-structure-notes/> [↑](#footnote-ref-1)
2. <http://ameliapsingh.wordpress.com/2013/04/14/duudddeee-its-proteins/> [↑](#footnote-ref-2)
3. <http://www.rcsb.org/pdb/home/home.do> [↑](#footnote-ref-3)