Contents

[Abstract 1](#_Toc362998256)

[1. Introduction 1](#_Toc362998257)

[2. Background 1](#_Toc362998258)

[2.1 Biological Backgrounds 1](#_Toc362998259)

[2.1.1 Nucleotide 2](#_Toc362998260)

[2.1.2 Deoxyribonucleic acid (DNA) 2](#_Toc362998261)

[2.1.3 Amino Acids 3](#_Toc362998262)

[2.1.4 Proteins 3](#_Toc362998263)

[2.2 Information Theory Backgrounds 3](#_Toc362998264)

[2.3 Random Forest 4](#_Toc362998265)

[3. Methods 4](#_Toc362998266)

[3.1 Information Theory based Signal Amplifying Framework 4](#_Toc362998267)

[3.2 DNA-binding Proteins Predictor 4](#_Toc362998268)

[4. Results and Evaluations 4](#_Toc362998269)

[4.1 Data Preparation 4](#_Toc362998270)

[4.2 Results & Evaluations 4](#_Toc362998271)

[5. Conclusions 4](#_Toc362998272)

[6. Discussions 4](#_Toc362998273)

[7. References 4](#_Toc362998274)

# Abstract

# Introduction

# Background

## Biological Backgrounds

In this sub-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.

### 2.1.1 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.

### 2.1.2 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)).

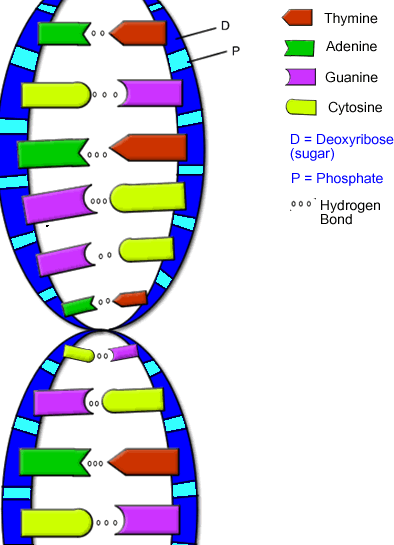


Figure .2: Structure of DNA

### 2.1.3 Amino Acids

### 2.1.4 Proteins

## Information Theory Backgrounds

Bb

## 2.3 Random Forest

# 3. Methods

## 3.1 Information Theory based Signal Amplifying Framework

## 3.2 DNA-binding Proteins Predictor

# 4. Results and Evaluations

## 4.1 Data Preparation

## 4.2 Results & Evaluations

# 5. Conclusions

# 6. Discussions

# 7. References

1. <http://jameschung3b.edublogs.org/2011/11/28/dna-structure-notes/> [↑](#footnote-ref-1)