DOCTOR OF PHILOSOPHY

Discovering properties of new DNA-binding activity of proteins

Xueting Wang

2014

Aston University

Discovering Properties of New DNA-binding Activity of Proteins

XUETING WANG

Doctor of Philosophy

– ASTON UNIVERSITY –

November 2013

ASTON UNIVERSITY

Discovering Properties of New DNA-binding Activity of Proteins

XUETING WANG

Doctor of Philosophy, 2013

Thesis Summary

Protein-DNA interactions are an essential feature in the genetic activities of life, and the ability to predict and manipulate such interactions has applications in a wide range of fields. This Thesis presents the methods of modelling the properties of protein-DNA interactions. In particular, it investigates the methods of visualising and predicting the specificity of DNA-binding Cys_2His_2 zinc finger interaction. The Cys_2His_2 zinc finger proteins interact via their individual fingers to base pair subsites on the target DNA. Four key residue positions on the α- helix of the zinc fingers make non-covalent interactions with the DNA with sequence specificity. Mutating these key residues generates combinatorial possibilities that could potentially bind to any DNA segment of interest. Many attempts have been made to predict the binding interaction using structural and chemical information, but with only limited success.

The most important contribution of the thesis is that the developed model allows for the binding properties of a given protein-DNA binding to be visualised in relation to other protein-DNA combinations without having to explicitly physically model the specific protein molecule and specific DNA sequence. To prove this, various databases were generated, including a synthetic database which includes all possible combinations of the DNA-binding Cys_2His_2 zinc finger interactions. NeuroScale, a topographic visualisation technique, is exploited to represent the geometric structures of the protein-DNA interactions by measuring dissimilarity between the data points. In order to verify the effect of visualisation on understanding the binding properties of the DNA-binding Cys_2His_2 zinc finger interaction, various prediction models are constructed by using both the high dimensional original data and the represented data in low dimensional feature space. Finally, novel data sets are studied through the selected visualisation models based on the experimental DNA-zinc finger protein database.

The result of the NeuroScale projection shows that different dissimilarity representations give distinctive structural groupings, but clustering in biologically-interesting ways. This method can be used to forecast the physiochemical properties of the novel proteins which may be beneficial for therapeutic purposes involving genome targeting in general.

Keywords: DNA-binding Cys_2His_2 zinc finger interaction, dissimilarity measures, high-dimensional data visualisation, new data prediction.

To
my parents.

Acknowledgement

First and foremost I would like to express my gratitude to my supervisors Prof. David Lowe and Dr. Anna Hine, for the patience, understanding and much needed guidance offered to me during my studies at Aston. Without their generosity of knowledge sharing, endless support and encouragement, it would not be possible for me to carry out this research programme and complete the thesis. Moreover, I am grateful for all members of the NCRG group, especially Prof. David Saad, Prof. Ian Nabney for useful discussions and to Vicky Bond, Kanchan Patel and Susan Doughty for dealing with all administrative jobs.

My sincerest thanks to Dr. Diar Nasiev who helped me a lot on mathematics and programming. Also to Dr. Rajeswari Matam and Dr. Michel Randrianandrasana who gave me lots of suggestion during my research.

During my years at Aston I have been receiving the strongest support from my family and friends in China. I feel so much grateful to my parents and my husband for help and encouragement.

Contents

List of Figures

List of Tables

1 Introduction

CONTENTS

This thesis addresses a key unsolved problem in the analysis of very high-dimensional yet topologically-ordered data, and in particular focusing on predicting properties of protein-DNA interactions. Protein-DNA interactions as the process of proteins recognizing nucleic acids, play a central role in transcriptional regulation and other biological processes (Wolfe et al., 2000). In general, the interactions are of mainly two types: specific interaction and non-specific interaction (CA et al., 1998). In this work, the specific interaction is selected as the main research target. Existing protein engineering techniques (Isalan et al., 2001; Wu et al., 1995; Rebar and Pabo, 1994) make it possible to manufacture proteins which can bind with specific DNA sequences. However, unless the properties of the protein are determined beforehand, there is an uncertainty as to whether the manufactured protein would bind with the desired DNA sequence. Usually, such binding activity can be verified through experiments, which is time-consuming and repetitive, especially when there are millions of potential combinations. To make it more efficient, various mathematical prediction models have been applied in this field, which have shown some promising results in predicting the binding status (Morozov et al., 2005; Siggers and Honig, 2007; Persikov et al., 2008; Nakata, 1995). In a broader sense, additional to binding status it would be useful for biological discovery to be able to estimate other properties of such molecular interactions, such as hydrophobicity and hydrophilicity with a target molecule.

The work presented in this thesis concentrates on studying the interaction between Cys_2His_2 zinc fingers and DNA sequences. One of the challenges is to construct effective models that can search for and discover implied relationships between the interactions that can be verified by experiments. Moreover, this thesis will also investigate whether these models can be applied to analyse the biochemical characteristics of novel DNA-binding Cys_2His_2 zinc finger interactions.

The following sections in this chapter will discuss the state of the art in understanding the Cys_2His_2 zinc finger binding DNA interaction, present the objectives of this work and give a summary of results. Finally, an outline of the thesis is provided.

1.1 The study of Cys_2His_2 zinc finger binding DNA interaction

The protein-DNA combination interacts when a protein binds a molecule of DNA to regulate the biological function of DNA, which usually is the expression of a gene. In this thesis, we focus on a small but highly significant subset of all possible protein-DNA interactions by analysing Cys_2His_2 zinc fingers. It is significant because it is the common type of DNA-binding domain found in the majority of eukaryotic genomes (JP and M., 1998; Shastry, 1996). Moreover, the original discovery of the Cys_2His_2 zinc finger and the elucidation of its structure led it to be a major focus of research over the past decades, especially the ability to recognise specific DNA sequences (Vallee and Auld, 1993; Pavletich and Pabo, 1991; Elrod-Erickson et al., 1996; Wolfe et al., 2000). Although the Cys_2His_2 zinc finger is a relatively simple motif, there still are significant challenges in understanding this protein and in developing methods which are designed to find widespread application in biochemical research and gene therapy (Wolfe et al., 2000). In addition, it is expected that the developed method can be extended to study more complex protein-DNA interactions.

Since there are twenty naturally occurring amino acids and four positions within a Cys_2His_2 zinc finger to make non-covalent interactions or bonds with the four bases denoted by the standard positions of (-1, 2, 3 and 6) within a DNA sequence [1], the total number of possible protein-DNA binding sites reaches almost 41 million, which makes it very difficult to discover potential protein-DNA interactions through random laboratory experiments. Although some physical techniques, such as X-ray crystallography and Nuclear magnetic resonance (NMR) spectroscopy (Pavletich and Pabo, 1991; Elrod-Erickson et al., 1996), have successfully shown the structure of binding interactions, various prediction methods have only managed to show limited results in identifying and predicting the protein-DNA binding specificity and affinity. For example, combinatorial randomized protein libraries (Hughes et al., 2005) have been generated for the purpose of identifying

[1] The detailed structure of the interaction and the mechanism of the interaction will be explained in Section 2.1.2.

novel zinc finger proteins without display, purification or sequencing.

Recently, a range of statistical models have been developed with the aim of trying to predict protein-DNA interactions (Morozov et al., 2005; Siggers and Honig, 2007; Kaplan et al., 2005; Wingender et al., 2001). According to the type of experimental data, the methods can be divided into two classes: structure-based and sequence-based. The structure-based methods mainly depend on crystallographic information of protein-DNA interaction. Through studying the structural characteristics of the typical DNA-binding Cys_2His_2 zinc finger interactions, such as binding energy [2] and amino-acid-nucleotide distance (Morozov et al., 2005; Siggers and Honig, 2007), knowledge-based parameters of the prediction models can be determined. Accordingly, for a novel prospective interaction, the models have the ability to evaluate relative affinities. However, as the structure-based models strongly rely on the selected template, the range of application is restricted. This problem can, however, be effectively tackled by using the sequence-based methods, where only the information of protein and target DNA sequence is required for constructing prediction models (Kaplan et al., 2005; Wingender et al., 2001).

The binding predictions made by either structure-based or sequence-based models are purely based on current knowledge. Therefore, it is rather difficult to predict the binding status for previously unseen protein-DNA interactions. This deficiency can, however, be rectified through investigating relationships between all possible DNA-binding Cys_2His_2 zinc finger interactions, which is a key contribution of this thesis. This has the disadvantage of generating huge amounts of high-dimensional data which therefore presents an enormous analysis problem. To overcome the analysis problem, various visual informatics approaches to the representation and characterisation of complex data will be introduced into the study of the protein-DNA interaction focussed on the Cys_2His_2 zinc finger, which can provide more insight into the data before implementing prediction models or experimental protocols.

[2] Definition of binding energy and relevant impact factors will be explained in Section 2.1.1.

1.2 Objectives

Visual informatics is the area of statistical pattern processing concerned with the discovery, analysis and interpretation of structure in complex data primarily through low-dimensional visualisation techniques. *Topographic* visualisation is one of the visual informatics methods used to map the data from a high dimensional space into a low dimensional space by preserving the structure of the data. The geometric structures of the data are usually defined through measures of relative dissimilarities between the data samples in the high dimensional space, where a suitable dissimilarity metric is used to reflect the prior knowledge of the domain (Sivaraksa and Lowe, 2008). In this thesis, the study has been divided into four phases which are shown in Figure 1.1: raw data collection and representation, data relationship visualisation, visualisation results verification and novel data investigation. The purpose of this section is to elaborate objectives of each phase, respectively.

To better understand the relationships between the possible DNA-binding Cys_2His_2 zinc finger interactions, it is essential to select an appropriate representation model to describe the features of available interactions as comprehensively as possible. Meanwhile, as visualisation intends to reflect the relationships between both real and theoretical interactions, it requires a representation model capable of demonstrating the core characteristics of any zinc finger binding DNA interactions, even when the experimental data is sparse. All these problems have to be solved in phase 1 in Figure 1.1.

Although various approaches have been developed to implement the function of finding data structures, a topographic visualisation method discussed in phase 2 will be selected in this work, which has the capability to visualise novel data directly by appropriately optimised models. Moreover, the format of the reconstructed data is also considered a key factor. Since a good visualisation method is expected to uncover biologically-useful structural and functional relationships between the data samples, which in other words, is to relate data samples with similar structures into similar groups, the study of similarity measures becomes another major objective.

Besides analysing the structural relationships of interactions discovered through the

data visualisation techniques, it is necessary to exploit the visualisation representations in predicting DNA-binding or other properties, of Cys_2His_2 zinc finger interactions. This can be achieved by using conventional (nonlinear) prediction models, which will be discussed in depth later on (in phase 3). Finally, novel DNA-binding protein interactions will be studied based on the developed visual informatics framework and verified through experiment in the laboratory in phase 4.

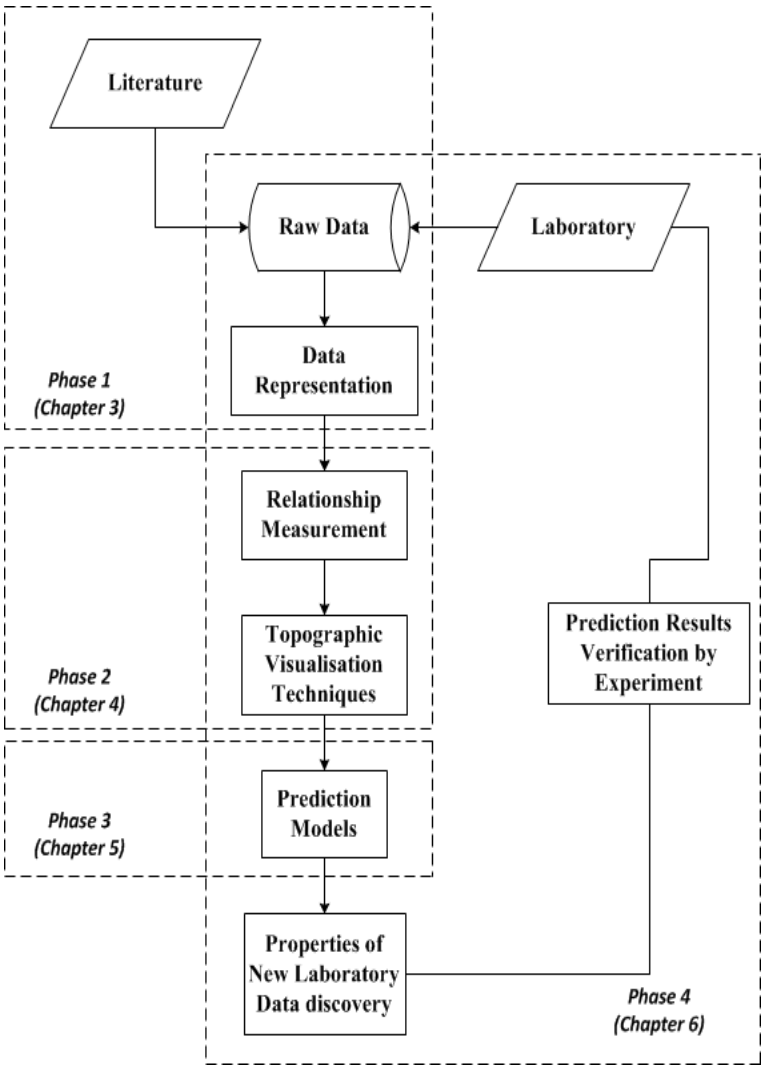


Figure 1.1: Process of studying the DNA-binding Cys_2His_2 zinc finger interactions. The flowchart shows the four phases in studying the structural relationships of the interactions: data representation, data relationship visualisation, visualisation results verification and novel data investigation. In phase 4, the in-silico predicted properties of novel potential zinc finger-DNA interactions can be verified through in vivo experiment in laboratory.

1.3 Contributions

In the process of achieving the objectives set out in Section 1.2, a series of interesting discoveries have been made. Among them, the most important is that the developed model allows the binding properties of a given protein-DNA interaction to be visualised in relation to other protein-DNA combinations without having an explicitly physical model, of the specific protein molecule and specific DNA sequence. In addition, through studying the visualisation results, the binding properties may be determined using a relatively comprehensive input training data set. In other words, since the visualisation representation is mainly implemented by a topographic feature extraction method, the *functional* properties of a given target interaction may be determined from the structural properties; i.e. neighbours which have similar topological properties in the data space, which can be verified by the relevant visualisation results based on experimental databases. In the process of creating the topographic representation model in Chapter 4, it is stressed that the binding status of known pairings is not applied. Apart from the experimental databases, a synthetic database which includes all possible combinations of the DNA-Cys_2His_2 zinc finger interactions will be generated for this thesis. Based on partial data randomly selected from the synthetic database as a training set to create a topographic projection, it will be shown that the experimental interactions occurring naturally are probably evolutionary favoured combinations.

In order to represent the relationships in the data and the binding properties of the protein-DNA interactions, various visualisation methods based on machine learning were investigated. A topographic transformation method was selected to preserve the geometric structure of the data in transforming from the original configuration space to the feature space. In such modelling, the geometric structure is described by relative dissimilarities between the data. Another contribution of the thesis was to realise that biological knowledge may not be best represented by a Euclidean dissimilarity, and we explored for the first time the use of an indefinite metric to reveal different structure in the Cys_2His_2 zinc finger -DNA binding data.

Since visualisation is beneficial for providing insight into the properties of the DNA-

binding Cys_2His_2 zinc finger interaction, relevant results will be exploited to construct non-linear prediction models in order to verify the patterns seen in the visualisation. So a final contribution of the thesis was to verify that there exists predictive information on structural and physical properties of DNA-Cys_2His_2 zinc finger combinations based solely on the data distribution in the encoding space as determined by the dissimilarity metric.

1.4 Outline of thesis

According to the flowchart in Figure 1.1, the thesis is designed to consist of seven chapters, each of which, except this chapter, is summarised as follows.

Chapter 2 reviews applied methods for studying the interactions between the zinc fingers and the target DNA sequences. X-ray crystallography and NMR spectroscopy approaches are briefly discussed to illustrate the background and evidence to support the structural mechanism of the processes in which the Cys_2His_2 zinc fingers recognise specific DNA sequences. Two classes of prediction approaches, the structure-based models and the sequence-based models, are discussed. A literature comparison between the approaches is also included in this chapter.

The DNA-zinc finger protein database used to support this thesis is described in Chapter 3. There are two completely different sources of data used in this thesis: one is a meta dataset taken from multiple publications from the literature, and the other is data extrapolated from laboratory experiments conducted at Aston. The characteristics of the data sources led to the decision that the canonical structure model has more advantages for data representation than the method known as biomedical fingerprinting. Also found in this chapter is the description of generating the database, and the conversion process in which the original data is transformed into binary feature vectors.

The methods of visualising the converted 320-dimensional database are the main topics of Chapter 4. Various dimension reduction methods are reviewed, most of which are deemed inapplicable for this work after analysing the characteristics of the reconstructed database. The method known as NeuroScale was the chosen topographic feature

extraction method applied to implement the lower dimensional topographic mapping for the 320-dimensional data visualisation. Euclidean and Minkowski metrics will be used to represent the dissimilarities in the high-dimensional space. The related visualisation results are also analysed from different aspects.

Whereas previous chapters focussed on unsupervised methods of data analysis, the focus of Chapter 5 is on investigating various supervised prediction models from machine learning, and comparing the accuracy of each model based on Receiver Operating Characteristic (ROC) curves. The visualisation results from Chapter 4 suggest the possibility of predicting the interaction by using the low-dimensional projected data. The prediction results are evaluated by selected quality criteria. The performance comparison between the prediction models based on different types of data is achieved by cross validating the quality criteria. A conclusion is provided at the end of the chapter as a reference for new data analysis.

The selected visualisation models are validated in this Chapter 6 through visualising groups of novel experimental data. The characteristics of novel data are described in the chapter. Apart from the visualisation results, some prediction methods are selected to quantify the visualisation results. During this process, a method is discussed to overcome some issues related to extreme outliers in the data samples.

The conclusions and contributions of the work and the recommendations for future research are given in Chapter 7.

This thesis is the work of the author but parts of it have appeared in the public domain including:

Conference Paper:

- Xueting Wang, Anna V. Hine and David Lowe. Signal processing issues of high-dimensional visual informatics: A study in protein-DNA binding patterns. In 9th IMA International Conference on Mathematics in Signal Processing, Birmingham, UK, 2012.

Patent:

- Xueting Wang, Anna V. Hine and David Lowe. Predicting properties of molecules. Patent: UK Patent Office. Patent Application Number: 1222627.0, Filing Date: 14 Dec. 2012.

2 DNA-binding Protein Interactions and Information Processing

CONTENTS

Cys_2His_2 zinc fingers, as a class of transcription factors, play a pivotal role in the genetic transcription from DNA to mRNA. In Chapter 1, the challenge of predicting the protein binding specificity and affinity has been presented. The main purpose of this chapter is to review some applied methods of studying the interaction between the zinc fingers and the target DNA sequences. Initially the background knowledge of the DNA-binding protein interaction, related biochemical concepts and interaction mechanisms are discussed. Then, the commonly employed prediction methods, which are based on both biochemical structure and sequence information, are discussed in the first part of Section 2.2. To conclude, some specific methodologies which have been applied to analyse and predict interactions between DNA and proteins are briefly introduced.

2.1 Protein-DNA interaction

Protein macromolecules and nucleic acids, are responsible for providing the behaviour of cells and performing various functions associated with life. The interaction between protein and DNA is a process whereby the protein recognises the nucleic acids by the basic rule of macromolecular recognition (Berg et al., 2006). Usually the interaction regulates the biological function of DNA and provides structural and catalytic roles in other cellular processes. The proteins involved in this process are transcription factors (TF) that can activate or repress gene expression in the vicinity of the binding site. In general, the proteins contact with the bases of DNA in the major groove, although there are also some known minor groove DNA-binding ligands (CA et al., 1998), such as Netropsin, Distamycin, Pentamidine amongst others. For the purpose of this thesis, those within the major groove are considered.

In order to elucidate the mechanisms of DNA-binding protein interactions, X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy have been employed to provide three-dimensional structural models of the interaction. In this section, the binding energy of protein-DNA interactions is introduced firstly. As the foundation of this work, the characteristics of Cys_2His_2 zinc finger protein, the function of the protein, and relevant X-ray structures are reviewed as well as the research methods that were used to

interpret the protein-DNA binding process interaction.

2.1.1 Protein-DNA binding energy

In the DNA-binding protein interaction, the chemical bonds are considered as primary factors in structure-based approaches. The chemical bonds can be categorised into two groups: covalent bonds and non-covalent bonds (Berg et al., 2006). A covalent bond is formed by the sharing of a pair of electrons between adjacent atoms. It defines the structure of molecules, which is the strongest among all chemical bonds (Berg et al., 2006). In contrast to the covalent bonds, the non-covalent bonds are weaker, but crucial for biochemical processes such as the formation of a double helix. There are four fundamental non-covalent bond types: (1) electrostatic interactions; (2) hydrogen bonds; (3) van der Waals interactions and (4) hydrophobic interactions (Berg et al., 2006). According to the three-dimensional structures of protein-DNA complexes, the contact between the protein and DNA backbone mainly involves non-covalent bonds (Carl, 1984) where the hydrogen bonds play a crucial role during recognition.

Binding energy is the physical index of assessing binding specificity of the proteins when interacting with DNA sequences. It is usually divided into two parts: specific and non-specific (Berg and von Hippel, 1987; Gerland et al., 2002). The specific binding energy exhibits a very strong dependence on the actual nucleotide sequence such as the hydrogen bonding, electrostatic and hydrophobic interactions. The non-specific part arises from interactions that do not depend on the DNA sequence which the TF is bound to, such as interactions with the phosphate backbone. To estimate the binding affinity, the binding free energy, defined as a sum of an intermolecular energy, a solvation free energy term and an entropic term is applied. This free energy can be modelled by a ‘position weight matrix’ (PWM)(Stormo et al., 1982) which will be discussed later as a prediction parameter. In general, when the bound pair has lower free energy, the protein-DNA binding is thought to occur with higher affinity.

2.1.2 Cys_2His_2 zinc finger

By studying the structures of various regulatory proteins that bind to specific DNA sequence, it has been revealed that roughly 80% of such proteins can be assigned to one of three classes based on their possession of one of three small, distinctive structural motifs: the helix-turn-helix (HTH), the zinc finger, and the leucine zipper (bZIP) (Berg et al., 2006). The Cys_2His_2 zinc finger known as "classic zinc finger", was first identified in the X*enopus laevis* transcription factor TFIIIA (H et al., 1995), and its three-dimensional structure was elucidated thereafter. Since the Cys_2His_2 zinc finger class has common interaction mechanisms when binding to the DNA sequence, it was selected as main focus of analysis in this thesis.

Characteristics of Cys_2His_2 zinc fingers

About one-third of the proteins in the Protein Data Bank (PDB) (Bernstein et al., 1978) contain metals, yet metal atoms are reported to be critical to the function, structure and stability of proteins (Shu et al., 2008). Approximately another one-third are metalloproteins (Holm and Sander, 1996) that are capable of binding to one or more metal ions (Passerini et al., 2007).

Zinc is the second most important metal [1] playing crucial roles in many biological functions. In zinc proteins, zinc ions can be observed in catalytic, co-catalytic or structural roles. For example, a catalytic zinc ion directly participates in the bond-making or bond-breaking step at the active site of an enzyme (McCall et al., 2000); in a co-catalytic zinc site, there are several metal ions bound in proximity to one another, where one plays a catalytic role and other metal ions enhance the catalytic activity of the site (Vallee and Auld, 1993); in structural zinc sites, the zinc ion mainly stabilizes the structure of the enzyme. The zinc finger is a nucleic acid binding motif[2]. The zinc fingers coordinate one or more zinc ions with a combination of cysteine and histidine residues to stabilize the protein architecture which is named as a fold. They can be classified by the type and order of the following zinc coordinate residues: Cys_2His_2, Cys_4 and Cys_6. This thesis will

[1] Iron is the most important metal in the biological functions.

[2] motif: is a sub-sequence which is thought to be an independent component or region of amino acids in one protein.

be focused on Cys_2His_2 zinc fingers, which are the most common type of DNA-binding domain found in the majority of eukaryotic genomes (Figure 2.1).

Each Cys_2His_2 zinc finger contains approximately 30 amino acids and comprises an antiparallel β-sheet followed by an α-helix around a tetrahedrally coordinated single zinc ion. Its structure is described as $CX_{2-6}CX_{12}HX_{2-6}H$, shown in Figure 2.1. In the structure, C denotes Cysteine, an α-amino acid with the chemical formula $HO_2CCH(NH_2)CH_2SH$ where SH is thiol. Cysteine is a non-essential amino acid, which means that it is biosynthesized in the human body. The side chain on cysteine is thiol, which is non-polar. Therefore, cysteine is usually classified as a hydrophobic amino acid. H in the structure represents Histidine, which is an essential amino acid that can not be synthesized within a human body and must be supplied through diet. It also has a positively charged functional group. X in the structure below can be any amino acid. Each Cys_2His_2 zinc finger domain has a conserved ββα structure, and amino acids on the surface of the α helix can recognize bases in a contiguous DNA sequence.

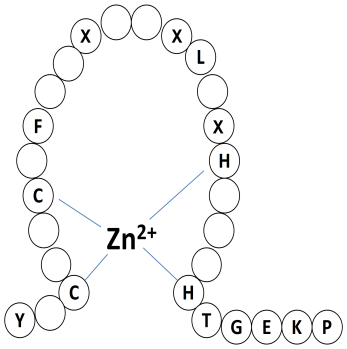


Figure 2.1: Cys_2His_2 zinc finger structure. Based on the figure in (Berg et al., 2006). C: Cysteine, H: Histidine. Other amino acids located in the rest part of the structure, such as: L:Leucine, F:Phenylalanine, P:Proline, G:Glycine, E:Glutamate, K:Lysine, Y:Tyrosine, X is any possible amino acid.

In the 1980s, the first high-resolution three-dimensional structure of a Cys_2His_2 zinc finger was determined by Nuclear magnetic resonance (NMR) spectroscopy, and its structure was analysed by using distance geometry and molecular dynamics (MD) calculations (MS et al., 1989). In general, the typical Cys_2His_2 zinc finger domain has a conserved ββα structure as shown in Figure 2.2. The anti-parallel β sheets encompass the two cysteine ligands which coordinate the zinc ion. The α helix contains the two histidine residues

that complete the zinc ion coordinate sphere (Laity, 2006). The zinc ion is buried in the core of the protein, and the structure of the protein is stabilised by the coordinate bonds between the cysteine, histidine residues and the zinc ion. According to the crystal structure of the interaction complex which was determined in the 1990s, the four amino acid residues are localised in specific positions (-1, 2, 3 and 6) on the surface of the α helix, which participate in DNA recognition by interacting with hydrogen donors and acceptors exposed in the DNA major groove.

Figure 2.2: Three-dimensional structure of Cys_2His_2 zinc finger. Taken from http://emergentcomputation.com/endo.html

Based on the location and number of zinc fingers, the Cys_2His_2 zinc finger proteins can be divided into three major groups (Iuchi, 2001). The first group, which is the main research target in the work, consolidates the proteins containing one cluster of three close zinc fingers. The proteins in the second group contain one pair or more of zinc fingers but with increased distance from each other. The third group of zinc finger proteins is characterised by the composition of four or more zinc fingers.

The major functional role of the Cys_2His_2 zinc finger, which is also characteristic of the protein, is to influence transcription of individual genes or gene groups. As transcription factors, the Cys_2His_2 zinc finger proteins can control the flow of genetic information from DNA to messenger RNA through binding to specific DNA sequences (Latchman,

1997; Karin, 1990). Amidst Cys_2His_2 zinc finger, transcription activation proteins promote the recruitment of RNA polymerase and vice versa suppressors. In the next subsection, the interaction between DNA sequences and the Cys_2His_2 zinc finger protein will be explained.

2.1.3 DNA-binding Cys_2His_2 zinc finger interactions

Zinc finger proteins are responsible for DNA or RNA binding, protein-protein interactions and membrane association. Specific to Cys_2His_2 zinc fingers, many of these proteins are transcription factors[3] that can be used to recognize specific DNA sequences. The α-helical portion of each finger fits in the major groove of the DNA sequence, and the binding of successive fingers causes the protein to wrap around the DNA. The majority of base contacts occur in three pair segments along the primary strand of the DNA. The sequence recognition is mediated mainly by amino acid in positions -1, 3 and 6 of the α helix (Figure 2.3) (Hughes et al., 2005), whereas the amino acid at position 2 can contact to the complementary strand of the DNA to stabilize the interaction.

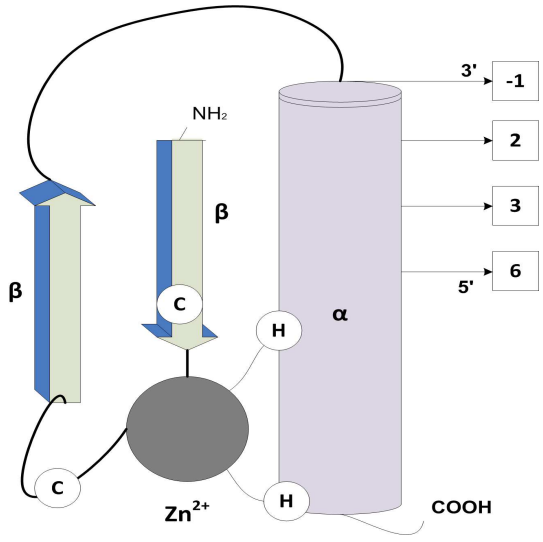


Figure 2.3: Cys_2His_2 zinc finger 3D structure with 3 base subsites. Based on the figure in (Tachikawa and Briggs., 2006).

[3] transcription factors which control when, where, and how efficiently RNA polymerases function, are vital for the normal development of an organism.

Research methods in DNA-binding Cys_2His_2 zinc finger interaction

Since the Cys_2His_2 zinc finger has the ability to recognize a variety of different sequences and may be able to "mix and match" fingers for new sites (Wolfe et al., 2000), many methods have been developed to understand the zinc finger-DNA interaction, and analyse base contacts that are made by zinc finger-like proteins.

X-ray crystallography is commonly used to determine the atomic and molecular structure of a crystal. The principle of this method is mainly based on measuring the angles and intensities of X-ray beams which are diffracted by the crystalline atoms. Through crystallography, a three-dimensional picture can be produced, which describes the density of electrons within the crystal. Thereby the mean positions of the atoms, the chemical bonds, the disorder and various information can be determined. Nuclear magnetic resonance (NMR) spectroscopy is another widely used method to study molecular structure. The energy of electromagnetic radiation which is absorbed and re-emitted by nuclei in a magnetic field depends on the strength of the magnetic field and the magnetic properties of the atom isotopes. NMR spectroscopy is frequently used to investigate the properties of organic molecules. These methods provide the chance to study the zinc-finger-DNA interaction in depth.

Since the Cys_2His_2 zinc finger protein can recognize the DNA target site with high affinity and specificity, there must exist structural properties which permit the linking between specific residues in the helix with identified bases in subsite locations (Wolfe et al., 2000). These properties ought to be evident in features extracted from experimental evidence. The experimental results can provide new information about the best protein finger for recognizing a given DNA subsite. Meanwhile, additional studies (Tsuchiya et al., 2004; Kaplan et al., 2005; Wingender et al., 2001; Morozov et al., 2005; Siggers and Honig, 2007; Nakata, 1995; Persikov et al., 2008) have been developed to predict the interaction between a functional protein and entirely novel sites. Although these methods have constructed libraries which contain a broader range of sequences as predicting references, mathematical methods are widely applied in the research.

2.2 Methodologies for biochemistry information analysis

As there are twenty naturally occurring amino acids in position -1, 2, 3 and 6 of a zinc finger to interact with the four bases of a DNA sequence, this makes the total number of possible protein-DNA binding sites, for zinc fingers alone, in the order of 41 million. It is infeasible to construct libraries and investigate binding possibilities through experiments for all these combination. Therefore, numerous attempt (Tsuchiya et al., 2004; Kaplan et al., 2005; Wingender et al., 2001; Morozov et al., 2005; Siggers and Honig, 2007; Nakata, 1995; Persikov et al., 2008) have been made in order to predict the interaction using models derived from structural and chemical information based on smaller experimental data sets.

In this section, the methods which are commonly applied to predict the interactions between proteins and DNA sequences will be reviewed first. Then, both structure-based and sequence-based prediction methods for DNA-binding Cys_2His_2 zinc finger interaction will be discussed, followed by a few specific methodologies used in this work.

2.2.1 General discussion of methods in biochemistry binding prediction

Structural and physical properties of DNA provide important constraints on the binding sites with which only the specific protein has the ability to recognise and interact. The three-dimensional crystallographic information of the protein-DNA structures provide the opportunities to understand the mechanism and the characteristics of the interaction. In this section, some typical prediction methods such as quantitative structure-activity relationship (QSAR), docking methods and molecular dynamics (MD) simulations will be discussed.

Quantitative Structure-Activity Relationship (QSAR)

Structure-Activity Relationship (SAR) (Sims and Sommers, 1985) is the relationship between the chemical or three-dimensional structure of a molecule and its biological ac-

tivity. It can be represented by molecular descriptors[4]. The QSAR models as a kind of prediction tool was developed by analysing the computational data based on molecular descriptors (the SAR data) to measure the binding likelihood (Zheng et al., 2006). This idea comes from the physicochemical properties of the compound that the variations of the chemical structure would affect biological activities (either reduce or increase activity) (Hames, 2000). The biological activities of a group of compounds are studied mathematically based on the physicochemical properties or theoretical molecular descriptors of chemicals. Moreover, quantitative values are measured or calculated for the physical features. In general, the QSAR models can be divided into five main types according to the different training algorithms: *k*-nearest neighbours (*k*-NN) (Altman, 1992), support vector machine (SVM) (Cortes and Vapnik, 1995), multiple linear regression (MLR), artificial neural network (ANN) (McCulloch and Pitts, 1943) and partial least square (PLS) (Wold et al., 2001). To implement a prediction, the QSAR models first extract the relationship between chemical structures and biological activity in a training dataset of chemicals. Secondly, the optimised QSAR models provide an estimate of the likely biological activities of new chemicals.

Docking methods

Docking algorithms which were first suggested in 1978 (Wodak and Janin, 1978; Janin and Wodak, 1985) are the methods for predicting preferred orientation of one molecule to another when they bound together to form a stable complex. Based on the knowledge of the preferred orientation, the binding affinity between the two molecules can be estimated through using various scoring functions[5]. The basic idea of molecular docking is to computationally simulate the molecular recognition process which can be thought as

[4] As Roberto Todeschini (Puzyn et al., 2010) defined: "Molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardised experiment."

[5] Scoring functions are fast approximate mathematical methods which are used to predict the binding affinity between two molecules after they have been docked (Jain, 2006). The scoring functions can be divided into three classes: force field which by estimating the sum of strength of intermolecular van der Waals and electrostatic; Empirical which is based on counting the number of various types of interactions between the two binding molecules (Bohm, 1998); knowledge-based which is based on statistical observation of intermolecular close contacts in large three-dimensional databases.

a " lock-and -key" relation[6]. An accurate predictive docking method can provide substantial structural knowledge about complexes. In general, there are two popular docking approaches: shape complementarity (Meng et al., 2004; Morris et al., 1998; Goldman and Wipke, 2000) and simulation (Feig et al., 2004). The shape complementarity methods describe the protein and ligand as a set of features that make them dockable. The simulation approach focuses on mimicking the actual docking process. For this method, the related energy cost due to any "moves" in the process of the ligand finding and binding to the active site of the protein is calculated. In general, the prediction accuracy of the docking methods is limited by the type of molecules and the biochemical information. However, current docking methods are accurate enough to guide drug design or for rational mutagenesis studies (Mendez et al., 2003).

Molecular dynamics (MD) simulation

Molecular dynamics (MD) was originally conceived within theoretical physics in the late 1950s (Alder and Wainwright, 1959; Rahman, 1964). In 1977, the first molecular dynamics simulation of a macromolecule of biological interest was published (McCammon et al., 1977). Today, MD simulation has been developed as an important tools for understanding the physical basis of the structure and function of biological macromolecules. The principal concept of the method is using computer to describe the interactions between the atoms and moleculars which govern microscopic and macroscopic behaviours of physical systems (Rahman, 1964). Specific to biophysical problem, MD simulation can provide detailed information on the fluctuations and conformational changes of proteins and nucleic acids (McCammon et al., 1977). Moreover, it has been shown that MD simulations together with free-energy calculations can provide quantitative predictions of protein-DNA binding energies (Yamasaki et al., 2012).

Applying MD simulations to study the interactions between biomolecules, the model uses a total potential energy function (**??**) to describe the molecule as a collection of atoms which are connected by harmonic bonds (two-body interactions), angles (three-

[6]Lock or receptor is the "receiving" molecule, most commonly a protein or other biopolymer; Key or ligand is the molecule which binds to the receptor.

body interactions) and dihedrals (four-body interactions) and the interaction forces such as the Coulomb and van der Waals potentials. Based on the MD simulation models, the data structures obtained from experiment can be determined or refined. It also can provide the description of the physical system which includes structural and motional properties, and examine the actual dynamics (Karplus and McCammon, 2002). The energy equation is usually expressed as follow:

$$\begin{aligned} E = & \sum_{bonds} K_b(r-r_0)^2 + \sum_{angles} K_\theta(\theta-\theta_0)^2 + \sum_{dihedras} K_\phi[1+\cos(n\phi-\delta)] \\ & + \sum_{impropers} K_\psi(\psi-\psi_0)^2 + \sum_{i>j} \varepsilon\left[\left(\frac{r_m}{r}\right)^{12} - 2\left(\frac{r_m}{r}\right)^{6}\right] + \sum_{i>j} \frac{q_i q_j}{4\pi\varepsilon_0\varepsilon r} \end{aligned} \tag{2.1}$$

where $\varepsilon\left[\left(\frac{r_m}{r}\right)^{12} - 2\left(\frac{r_m}{r}\right)^{6}\right]$ reperesents the potential of van der Waals interactions, $\frac{q_i q_j}{4\pi\varepsilon_0\varepsilon r}$ is used to calculate the Coulomb potential comes from the electrostatic interactions.

Methods in DNA-binding Cys_2His_2 zinc finger interaction prediction

Which method should be used for predicting the DNA-binding Cys_2His_2 zinc finger interaction usually depends on data representation models. Generally, there are two types of commonly used methods, namely structure-based (Tsuchiya et al., 2004) and sequence-based (Kaplan et al., 2005; Wingender et al., 2001) prediction methods for representing the data of the zinc finger-DNA interaction. The biochemical fingerprinting model is the foundation for the structure-based method and is focused on describing the kinetics of several biochemical reactions, while the sequence-based prediction method relies on a canonical structure model which only describes the structure of DNA-binding Cys_2His_2 zinc finger. The two methods will be elaborated in Section 3.2.

Structure-based prediction of DNA-binding sites on proteins

The transcription factors, as mentioned in Subsection 2.1.2, are groups of proteins that bind to specific DNA sequences, known as the transcription factor binding sites that control the transcription of genetic information from DNA to mRNA. The structure-based prediction method uses the crystallographic information of the protein-DNA structures

obtained by either X-ray crystallography or NMR spectroscopy to predict binding specificity and affinity.

Given a structure of a protein-DNA complex, a model is needed to quantify the similarity and evaluate relative affinities, such as the one developed by Alexandre et al. (Morozov et al., 2005) that is based on all-atom[7] free protein-DNA binding energy. As proteins can recognize specific DNA sequences mainly by way of direct readout through base-amino acid contact and indirect readout through DNA conformation, free energy is defined to consist of protein-DNA energy[8] and DNA conformation energy[9] . Another well-known method uses knowledge-based structure potentials (Siggers and Honig, 2007). These potentials are based on the selected structural parameters, such as amino-acid-nucleotide distance; twist, roll and tilt parameters of the base-pair (Siggers and Honig, 2007). Since the prediction is based on the structure of the protein-DNA complex, for a novel pair of proteins and target DNA sequences, selecting a suitable structurally homologous protein that has a sufficiently similar structure as a template becomes a necessary prerequisite. Due to the limited number of protein-DNA complexes obtained by experiments, and since these models strongly rely on the existing data examples in protein databases (PDB), the prediction accuracy is restricted by the similarity between the structure template and the target protein-DNA complex. To overcome this obstacle, the sequence-based prediction method is introduced as an alternative to predict the protein-DNA binding interaction.

Sequence-based prediction of DNA-binding sites on proteins

Sequence-based prediction can be understood as using statistical estimation procedures to estimate the context-specific DNA-recognition preferences based on a set of pairs of transcription factors and the target DNA sequences (Kaplan et al., 2005; Wingender et al., 2001). The set of pairs of transcription factors and target DNA sequences can be represented by a canonical structure model [10] which describes the residues and nucleotides

[7] All-atom means every atom in the protein.

[8] The protein-DNA interaction energy is used to describe direct readout of the DNA sequence by the protein, such as polar interactions (electrostatics and hydrogen bonds), van der Waals forces and solvation energies.

[9] The DNA conformation energy considers distortion of B-form DNA caused by protein binding.

[10] Details of the canonical structure model will be discussed in Section 3.2.2.

participating in the protein-DNA interaction. With pre-processed data, various prediction methods can be employed, such as support vector machines (SVM), probabilistic models and multilayer perceptrons.

As one example of a prediction model, the SVM as a supervised learning algorithm is widely employed in bioinformatics (Persikov et al., 2008). In DNA-binding Cys_2His_2 zinc finger protein prediction research, the SVM classifies the feature vectors of proteins as positive (DNA-binding) and negative (non-DNA-binding) based on the posterior probability of zinc-binding for a residue in the chain. In Persikov's research (Persikov et al., 2008), both linear and polynomial kernels have been employed in SVM models to predict zinc finger protein-DNA binding on the basis of the canonical binding model. Unlike the structure-based prediction method, known examples of non-binding zinc finger-DNA pairs are also incorporated in the database. The model has been evaluated by cross-validation tests and the results show that the prediction accuracy of the SVM is better than previously published methods when comparing receiver operating characteristic (ROC) curves and area under the curves (AUC) (Persikov et al., 2008).

For the probabilistic models, first of all, four matrices of conditional probabilities of the four nucleotides given all 20 amino acid are calculated as DNA-recognition preferences for the model. Since the database only reports the DNA sequences that contain the binding sites, without the exact binding locations provided, iterative expectation maximization (AP et al., 1977) is used to learn both the probabilities associated with the contacts in the canonical model as well as the binding locations. Then, using the appropriate set of DNA-recognition preferences, given a novel pair of zinc finger protein and a target DNA sequence, the potential binding probability can be calculated.

Besides the two methods mentioned above, multilayer perceptrons (MLP) can also be applied to anonymous protein sequence analysis of zinc-binding sites (Nakata, 1995). The MLP provides an optimised non-linear mapping function that maps the input feature vector x to an output that represents the binding affinity.

Between the structure-based and sequence-based prediction methods, the most obvious advantage of the latter is that the model is independent of the integrity of the crystal-

lographic information. Moreover, since the dependence of the sequence-based model on the similarity between the structure template and the target protein-DNA complex is less significant than that of the structure-based model, the limitation which is caused by the number of existing data samples is reduced.

2.2.2 Specific methodologies used in this thesis

To study the important role of Cys_2His_2 zinc finger proteins in sequence-specific DNA-binding interactions, besides the methods discussed above, some specific methodologies are adopted. It mainly includes a data coding model for pre-processing; high dimensional data visualisation; and non-linear prediction models of the DNA-binding Cys_2His_2zinc finger interaction. Attempts to characterise similarity between a new data set and a known data set are also addressed herein.

Sequence-based data coding

As discussed in Subsection 2.2.1, the method of selection for studying the DNA-binding Cys_2His_2 zinc finger interaction strongly depends on the properties of the original data samples. In the current study, the canonical binding model as a sequence based method is employed to convert the original data to a sparse 320 dimensional binary vector, which will be the main topic of Chapter 3.

Topographic Visualisation

Before constructing a prediction model, this thesis develops various data visualisation models to gain insights into the relative distributions of the protein-DNA combinations that exist in nature. Although several techniques exist to represent high dimensional data as low dimensional objects, NeuroScale as a topographic feature extraction method will be selected to implement a lower-dimensional topographic mapping representation for high-dimensional data visualisation (Tipping, 1996). It employs a nonlinear transformation to preserve geometric structure while mapping the data from the original configura-

tion space into the feature space. The geometric structure can be described by relative 'dissimilarities' which are the distances between feature vectors in the original and transformed spaces. More discussion and related results will be presented later in Chapter 4.

Prediction models

Different prediction models will be applied to predict the zinc finger-DNA binding affinity. Besides the commonly used neural networks and SVM model (Nakata, 1995; Persikov et al., 2008), the *k*-nearest neighbours (*k*-NN) algorithm, the relevance vector machine (RVM) and linear regression are also investigated in this thesis. The *k*-NN algorithm will focus on studying the projected visualisation results that reflect the distribution of data samples in the high-dimensional feature space. Meanwhile, the RVM (Tipping, 2001) which is a probabilistic model and has similar structural form to the SVM, will be utilised to derive a prediction model based on both visualisation results and the high-dimensional original space. In Chapter 5, the prediction results obtained by the various models will be evaluated and compared using ROC/AUC and prediction error criteria.

Similarity measures

In this thesis, we explore several measures of dissimilarity, in both the data and visualisation spaces. In this work, NeuroScale which is used to study high-dimensional data, requires a measure of the dissimilarity between two pattern vectors. Moreover, the similarity is applied to evaluate the possibility of predicting binding status of novel data. In general, the suitability of a measure depends on the data characteristics and the problem domain and should ideally be driven by expert knowledge. There are many dissimilarity measures for numeric variables, such as Euclidean distance, City-block distance and Minkowski distance. For binary variables, other measures which are more specific to discrete data are also available, including Hamming and Jaccard distances (Webb, 1999). In Chapter 4, different measures of dissimilarity and the results of various experiments based

on different dissimilarity measures will be explained and analysed. It is a topic for future research to determine what measure of similarity is optimally suggested by the biological prior knowledge.

2.3 Summary

In this chapter, basic biochemical concepts and the principles of the DNA-binding Cys_2His_2 zinc finger interaction were introduced. The Cys_2His_2 zinc fingers, as one of the most common transcription factors, have the ability to recognise specific DNA sequences principally by the amino acids in positions -1, 3 and 6 of the α helix. In order to understand the zinc finger-DNA interaction, X-ray crystallography and NMR spectroscopy have been applied by others to analyse the protein-DNA complexes. Different data representation models will be used in models to predict the likelihood of binding interaction between functional zinc finger proteins and an entirely novel DNA-binding site. Correspondingly, two classes of prediction approaches were mentioned: the structure-based models and the sequence-based models. The structure-based methods are based on physical and chemical structures of observed experimental protein-DNA complexes. The prediction accuracy of these models strongly rely on the structure of the observed experimental template. Contrasting with the structure-based models, sequence-based models depend only on pairs of transcription factors and the target DNA sequences. The SVM and multilayer perceptron to be discussed in Chapter 5 are two typical sequence-based models applied to estimate the context-specific DNA-recognition preferences.

This thesis will focus on the sequence-based prediction methods. The experimental data represented by the canonical binding model will be introduced in Chapter 3. Visualisation models, utilised to project the high-dimensional data into the low-dimensional feature space based on the dissimilarities between data samples will be discussed in Chapter 4. By analysing the visualisation representations, groups of prediction models will be developed and discussed in Chapter 5. In the next chapter, alongside the data representation models, the characteristics of the selected experimental data and the process of creating the various databases used by the predictive models will be outlined.

3 Coding and creation of Cys_2His_2-DNA binding databases

CONTENTS

The purpose of this chapter is to describe the special coding scheme and the creation of the DNA-zinc finger protein database used for subsequent analysis and model-building. The DNA-binding zinc finger (ZF) protein interaction is one of the essential features in the genetic activities of life. More and more researchers are setting their sights on the ability to predict and manipulate such interactions. With the improving experimental techniques and methods in the past two decades, databases that contain realistically possible interactions have been constantly enriched, leading to more open research opportunities.

In this thesis, all experiments and discussions are based on real experimental data obtained from different laboratories. Although the total possibilities of particular protein-DNA binding sites are almost 41 million [1], relatively few interactions occur in nature. Therefore, how to process the existing data sources, and extract and represent the features by an appropriate model become the first of the crucial problems.

The focus of the chapter is on the analysis of characteristics of the experimental data for the database generation based on the canonical structure model. The properties of the samples in the original data sets are analysed in Section 3.1. A range of biochemical data representation models are discussed in Section 3.2, where the canonical binding model is selected to implement the data features representation. The process of generating the database is described in Section 3.2.2. Finally, a demonstration of the converting process from original data to analytical vector is provided in Section 3.3.

3.1 Data source

Data samples, collected from numerous experiments, provide the opportunity to study and understand the principle of the DNA binding zinc finger interaction using mathematical methods. In this thesis, some publicly available experimental data from different sources have been studied. Meanwhile, a randomized protein library will be selected as a novel data set (out of sample) for validation. The characteristics of each data source are also discussed in this section. Moreover, the basic principle of data selection and database

[1] There are $20\times 4 = 80$ possibilities in each binding position. Considering each binding pair includes four positions, the total possibilities is $80\times80\times80\times80 = 40960000$.

generation are introduced.

3.1.1 Characteristics of published experimental data and selection

The original set of data was kindly provided by Anton V. Persikov [2](Persikov et al., 2008). The original data consists of 26 separate literature data sources representing experiments performed across separate laboratories around the world, looking at the binding status of DNA and Cys_2His_2 zinc finger proteins. An example of the main structure of the data is shown in Table 3.1, and more detailed examples can be found from Table A.2 in Appendix A. All original data samples include only the information of the primary chain of DNA sequence in the 5'-3' order, the amino acids in every single zinc finger from left to right which are numbered as -1 to 6[3] and the quantitative information about binding affinity (e.g. K_d[4])

DNA	f1	f2	f3	K_d (nM)
ctcgcgGGGgcggcc	KSADLKRHIRI	RSDHLTTHIRT	RSDERKRHTKI	0.5

Table 3.1: Example of original data. In the first column, the primary sequence of the DNA is provided in the 5' to 3' direction. The capital letters are the bases which would be contacted by amino acids at -1, 3 and 6 positions. The second to fourth columns contain the detailed information of 3 zinc fingers, but in some cases, only the second zinc finger is studied during the designed experiment. The last column includes quantitative information about binding affinity, such as '0.5 nM' in this example.

By re-organizing the original data, the 26 data sources are divided into three groups based on different characteristics. The first group which is listed in Appendix A Table A.3 only features the interaction between the DNA sequence and the second zinc finger. The data sources which discuss the interactions between the DNA sequence and three zinc

[2]With thanks to Dr. Persikov for direct correspondence and access to data.

[3]The variable part of each zinc finger (such as the sequence in f1: 'HIRI' in Table 3.1) normally holds exactly same information in the same experiment, and makes no contribution toward the interaction, it is omitted from further processing.

[4]A dissociation constant $K_d^{-1} = \frac{[RL]}{[R][L]}$ is a specific type of equilibrium constant that measures the propensity of a larger object to separate reversibly into smaller components. Units of K_d is nM, when K_d<200 nM, the data sample can be considered as a positive binding example.

fingers were assigned to the second group (Appendix A Table A.4). The third group as listed in Appendix A Table A.5, consists of the publications which contain comparative examples without any quantitative information about binding affinity. These comparative examples are generated by comparing the value of K_d [5]. In the process of the database creation, all duplicated data samples have been filtered out from the original dataset. Only a small number of data samples (a total of 31 data samples) which share the same protein-DNA pairs but reported with contradictory binding status in different experiments are kept and used as part of a validation data set later in Chapter 6.

3.1.2 Characteristic of laboratory data and selection

Besides the published data described in Subsection 3.1.1, a combinatorial randomized protein library is selected to create a test dataset. This original data set was provided by Dr. Anna V. Hine [6] (Hughes et al., 2005). In the data set, the primary DNA sequence 5'-T_{10}GGGXXXGCTT_{10}-3' where 'XXX' refers to any codon at positions -1, 3 and 6 [7] is designed to interact with various zinc finger proteins.

Figure 3.1 shows an example of the original data sample from the randomized protein library [8]. All theoretical interactions between each specific DNA sequence and 8,000 proteins at position -1, 3 and 6 can be described by three graphs. The resulting data of each position are normalized [9] and sorted by the highest signal (Hughes et al., 2005). When identifying possible candidate proteins for interaction with a target DNA sequence, data from both three and four washes need to be considered. Although the data from three washes are considered as the staple factor which can basically reflect the interaction trend, the data from four washes show similar trends to that from three washes, there are some exceptions that may be used to eliminate possible interactions (Hughes et al., 2005).

[5] In this case, a protein-DNA pair with K_{d1} is considered to have a stronger binder than a protein-DNA pair with K_{d2} ($K_{d1} < K_{d2}$)(Persikov et al., 2008). By reorganizing the comparative examples, finally gives a total of 673 data samples without binding affinity.

[6] Thanks for data source providing and research guidance.

[7] All 64 codons are included in the data set (4 bases$\times$4 bases$\times$4 bases).

[8] There are 60 randomized protein libraries reconstructed by using 'MAX' randomization (Hughes et al., 2003). Each library contains compounds with one specific residue 'fixed' as a single building block and the remaining residues fully randomized. Then, the data from the library screening is scaled according to the total amount of GFP fluorescence present in each library (Hughes et al., 2005).

[9] Where 100% = the highest signal after three washes (Hughes et al., 2005)

Figure 3.1: Processed screening data for zinc finger libraries. The triplet in DNA sequence is 'ATA'. '-1', '3' and '6' indicate the positions of contacting residues. Blue bars represent data measured after three washes and red bars represent data obtained after four washes (Hughes et al., 2005). Values on the ordinate are normalized DNA binding signals, and letters on the abscissa are the amino acids at three positions.

3.2 Data processing

As discussed in Chapter 2, the study of protein-DNA interactions has proved zinc fingers can be used in developing novel transcription factors which would regulate the transcription of genetic information from DNA to mRNA(Hughes et al., 2005). More and more biologists are trying to develop various methods to construct a wide range artificial zinc finger libraries. In order to utilize and study the collected data sources introduced in Section 3.1 efficiently, the canonical binding model is employed in this thesis. Compared with other structural models (e.g. biochemical fingerprinting[10]), such a canonical model only relies on the interaction information between a DNA sequence and proteins. It reduces the requirement of the amount of information, but should still be able to describe the features of each data sample properly. In this section, the canonical structure model is first introduced followed by the description of database creation.

3.2.1 Data representation in binary format - Canonical structural model

Through studying the structures of the DNA-binding protein interactions, the canonical binding model as a structural model was recommended (Elrod-Erickson et al., 1996; Pavletich and Pabo, 1991) and widely applied in predicting DNA-binding protein interaction (Persikov et al., 2008; Kaplan et al., 2005).

The basic idea of the canonical structural model focuses on describing the structure of DNA-binding Cys_2His_2 zinc finger. The α-helix in each finger fits into the major groove of the DNA, and each consecutive finger contacts the nucleotides within four base subsites. For each zinc finger, there are three amino acid positions: -1, 3 and 6 which contact the primary DNA strand, while the amino acid at the 2^{nd} position makes contact with the complementary DNA strand. A simple model which explains the principal of representing the experimental data based on the canonical binding model is shown in Figure 3.2. The amino acids numbered as a_{-1}, a_3 and a_6 contact the bases b_3, b_2 and b_1

[10]Biochemical fingerprinting is defined as a phenotyping method in which the kinetics of several biochemical reactions are recorded based on specialized analytic techniques. Specific to the DNA-binding zinc finger protein interaction, the methods, such as X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy, are selected to describe the structure of the interaction.

respectively. Only a_2 contacts b'_4 in the complementary strand. As the experimental data only includes the information of the primary DNA chain, b'_4 should be the paired base of b_4 in the complementary chain. According to Figure 3.2 and Table A.9 in Appendix B, each experimental data sample can be denoted in a binary 1×320 vector and studied by different data analysis models. The process of creating a training data set using the canonical binding model is introduced in Subsection 3.2.2, and an example is provided in Section 3.3.

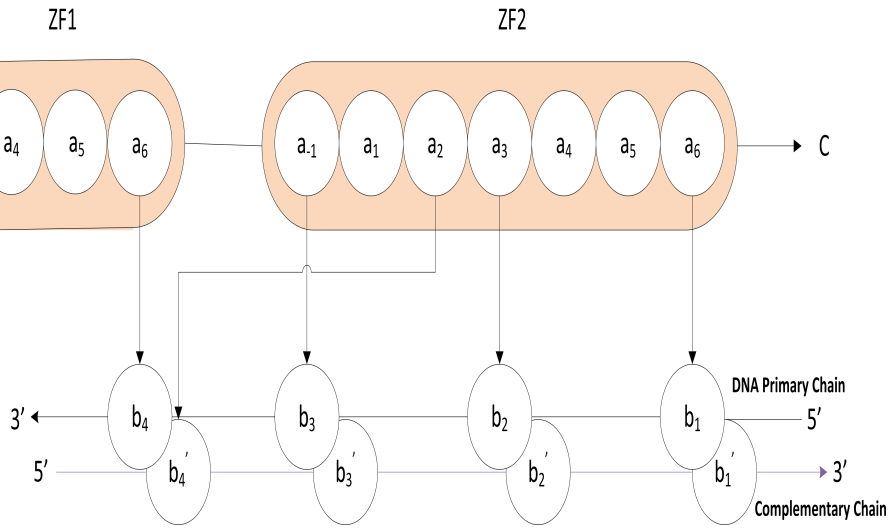


Figure 3.2: The canonical DNA binding Cys_2His_2 zinc finger model based on the figure in (Persikov et al., 2008). Residues at position 6, 3, 2 and -1 in the α-helix interact with nucleotides which are numbered sequentially from 5 to 3 of the primary DNA chain, and are primed in the complementary DNA chain.

3.2.2 Database creation description

Out of the original 26 data sources, 25 data sources were selected to generate the training dataset used later in this thesis. The 26th data source is excluded because the related data samples provide implicit binding statuses[11]. As mentioned in Subsection 3.1.1, the selected data sources are divided into 2 groups depending on the number of zinc fingers. In this thesis, the analysis is focused on the 13 citations in Table A.3 where one zinc finger

[11]The 26th data source contains comparative examples without any quantitative information about binding affinity

is studied and the 12 citations in Table A.4 where three zinc fingers are investigated. The unused data source and part of the data samples from the selected data sources, where the binding status is also implicit, will be exploited as a validation data set. Therefore, there are 1860 data samples available in total for the study.

Based on the concept of the canonical structural model, each data sample can be expressed as a 320-dimensional vector[12]. As shown in Figure 3.3, each vector is divided evenly into four sections. Each section denotes one binding position of the interaction. For example, the first section, i.e., from index 1 to 80, indicates the binding of one amino acid with one nucleotide on complementary strand at position 2. The second section (81 to 160) for position -1, the third (161 to 240) for position 3 and the fourth (241 to 320) for position 6 . Once the nucleotide and amino acid in the four positions are determined, it is easy to convert the binding pairs into a set of model numbers by using Table A.9 in Appendix A [13]. The model numbers of each data sample indicate the indices of four elements in each vector that are set with 1. The rest of the vector is filled with 0 as shown in the fourth layer in Figure 3.3.

[12]Every amino acid $a \in \{Ala, Cys, \ldots, Trp\}$interacting with base $b \in \{A, C, G, T\}$ at specific contact position can be defined using the binding model. All possible combinations are numbered and marked in a feature space containing 320 dimensions representing (20 amino acids $\times$4 bases $\times$ 4 contacts).

[13]Since the binding interaction at position 2 occurs on the complementary DNA chain, it is necessary to convert the base to paired base before checking the index.

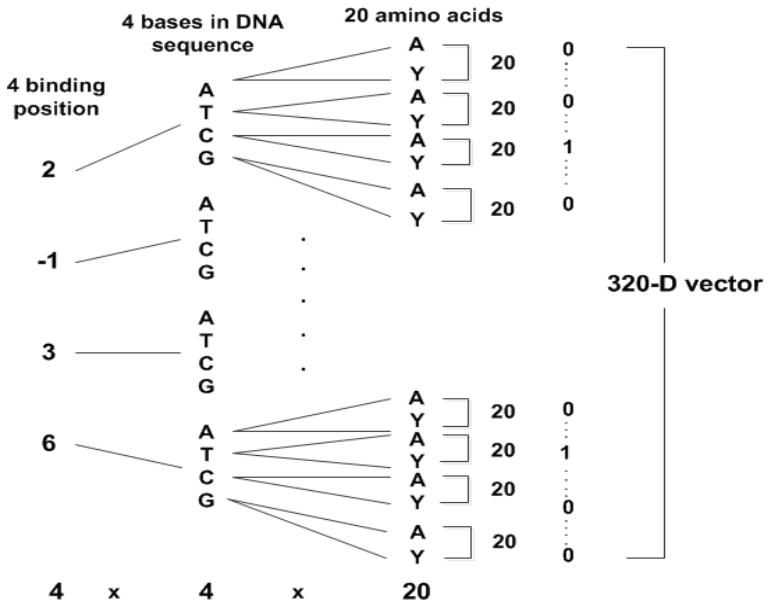


Figure 3.3: The architecture of the 320-dimensional vector. In the figure, four layers are used to describe the principle of the vector generation. The first layer indicates the four binding positions of each data sample. At each binding position, there are four possible bases in the second layer which can be recognised by an amino acid shown in the third layer. By checking the look-up dictionary in Table A.9, the elements with connections are set to '1'; other elements are set to '0'.

Table 3.2 shows an example of a 320-dimensional vector which represents the binding pair: GCGg (DNA) and THRD (zinc finger). The binding status is defined as [0 1] for binding or [1 0] for non-binding based on the K_d value or the binding status provided in the data sources. A threshold of $K_d = 200$nM is used to specify binding or not. The detailed process of the database generation is explained in Appendix A.4.

$1 \cdots 22$	**23**	$24 \cdots 134$	**135**	$136 \cdots 266$	**267**	$268 \cdots 296$	**297**	$298 \cdots 320$
$0 \cdots 0$	**1**	$0 \cdots 0$	**1**	$0 \cdots 0$	**1**	$0 \cdots 0$	**1**	$0 \cdots 0$

Table 3.2: Example of an 320-dimensional vector. This 320-dimensional vector example is created based on the binding pair: GCGg (DNA) and THRD (zinc finger). Since the model number of the binding pair at 2 position is 23, the twenty-third element in the 320-dimensional vector is set with 1. Other elements between 1 and 80 are set with 0, and so on for position -1, position 3 and position 6.

As discussed in Subsection 3.1.1, the removed duplicate samples with unknown binding status are re-organized to create a test dataset, while the remaining data with vague

binding information are merged with the unused data source as a validation data set. Moreover, as there are twenty naturally occurring amino acids, and four positions within a zinc finger to interact with the four bases within a DNA sequence, this makes the total number of possible protein-DNA binding sites, for zinc fingers alone, almost 41 million possible configurations. A synthetic database (DB5 in Table 3.3) has been created which contains all the 41 million possibilities. It will be used to study high dimension visualisation in the next chapter. Table 3.3 lists all generated databases which will be utilised in this work. DB1 as the training data set will be applied to study the relative distributions of the protein-DNA combinations that exist in nature. The detailed information of the data sampels are listed in Appendix A Table A.10. Moreover, it will be used to train the prediction models before involving the test data set (DB2) and the validation data sets (DB3 and DB4).

Database (DB)	Type of Database	Total number of samples	Sources
DB1	Training data set	1860	Published papers
DB2	Test data set	673	Comparing data from papers
DB3	Validation data set	7615	Laboratory data
DB4	Validation data set	31	Duplicated data from papers
DB5	All combination	41 million	Sythetic data

Table 3.3: Summary of databases. DB1 is the combined, filtered data samples listed in Table B.1.2 and A.3. DB2 is the database which only includes the filtered data which is listed in Table A.4. DB3 is generated based on the laboratory data. Details of creating the data will be described in Chapter 6. DB4 only has 31 data samples. These samples are the filtered duplicated data without binding status from the published papers listed in Table B.1.2 and A.3.

Interaction status	Number of data samples	Propotion
Binding	882	47.42%
Non-binding	978	52.58%
Total	1860	100%

Table 3.4: Detailed information of database DB1. In database DB1, the 1860 data samples consist of 882 binding and 978 non-binding examples where the number of the non-binding samples is slightly more than the binding samples.

3.3 A data reconstruction example

In this section, an example is provided to demonstrate how an experimental data sample is converted to a 320-dimensional vector, shown in Figure 3.4.

Step 1: Order an original data sample as a DNA sequence (**5'-3'**): ctcgatTGGgcggcc, three fingers: KSADLKRHIRI, RSDHLTTHIRT, TSGNLVRHTKI and a K_d value;

Step 2: By determining the primary interaction finger, the interaction bases 'TGGg' which bind to specified amino acids 'RSDHLTTHIRT' are selected from the primary DNA chain ;

Step 3: The sequence of target bases is reordered from TGGg (**5'-3'**) to gGGT (**3'-5'**) for convenience in the future;

Step 4: The binding pairs at each position are stored. In this example, according to the rule of interaction, bases and amino acids are stored in binding pair format: Base+Amino Acid (CD, GR, GH, TT);

Step 5: According to the definition in Table A.3 in Appendix, the binding pairs are numbered based on amino acid positions as: 01cD(2 position), 02gR(-1 position), 03gH(3 position), 04tT(6 position);

Step 6: The numbered pairs are represented by a serial number with respect to Table A.9: 01cD(23), 02gR(135), 03gH(207), 04tT(317);

Step 7: A 1×320 zero vector is created for the binding pairs. Four elements with specified indices in the vector are marked 1, otherwise, marked 0. While, the threshold of K_d value is set as 200nM. When K_d is smaller than 200nM, it is stored as binding [0 1], otherwise not binding [1 0]. Since 400>200, K_d is stored as [1 0] in this example. This label will be used in classification experiments in Chapter 5.

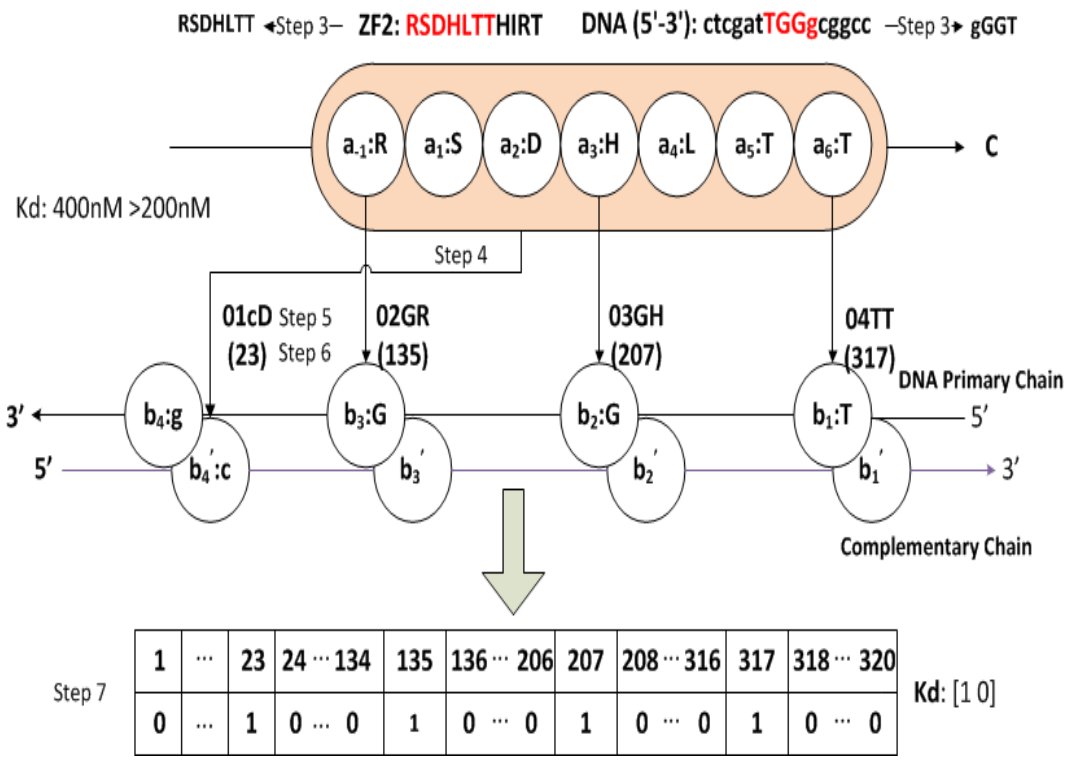


Figure 3.4: Example of 320-dimensional vector generation. In the figure, the first two steps are omitted as the binding pair information is highlighted in red. The rest of the steps are expressed using the canonical binding model.

3.4 Summary

In this chapter the coding, creation and characteristics of the collected data sources were described. In total, 25 data sources were selected to form the training data set. Another combinatorial randomized protein library based on experiments will be used as the test data set. Two representation models, biochemical fingerprinting and canonical binding model are available to describe the characteristics of the original data. Although biochemical fingerprinting can retain observed biochemical structure information, the limited number of available data samples compared to that of the 41 million samples in the theoretical database limits its capability. Therefore, the canonical binding model is selected as the primary data representation model to convert the original data to the sparse binary vector. It has the advantage of preventing the database from including the biased binding data samples, as it only focuses on describing the structure of the DNA-binding

zinc finger protein. Even though the number of the experimental data is limited, a theoretical database is generated by considering all possible combinations between 4 bases in a DNA sequence and 20 amino acids in a zinc finger protein based on the canonical binding model. The theoretical database will be used to validate the effect of the lack of experimental data in the next chapter. In order to explain the process of database generation, an example which demonstrated how to use a 1×320 vector to represent an original datum was provided in the last section and the details of each step are included in Appendix A.

In this thesis, the converted database is the basis for predicting the DNA-binding Cys_2His_2 zinc finger interaction. Since it is difficult to study the structural relationship in high dimensional space, various visualisation methods will be applied and discussed in the next chapter.

4 Analysis methods (I): Data visualisation

CONTENTS

In Chapter 3, through analysing the characteristics of the selected data sources applied in this work, the canonical binding model was employed as the primary data representation model to convert the original data to the high dimensional sparse binary vector. This chapter discusses how the constructed 320-dimensional database was visualised. In particular, NeuroScale will be discussed as the main visualisation method in this work, which is used to implement lower-dimensional topographic mapping representation for the 320-dimensional data visualisation. This chapter begins with reviewing various visualisation methods for the high-dimensional database. Then, the characteristics of the represented 320-dimensional database are analysed in Section 4.2, followed by the visualisation results of various standard visualisation techniques. In Section 4.3, the dissimilarity and relevant preconditioning and quality criteria of the NeuroScale are introduced at the beginning. Then, before analysing the visualisation results of the high-dimensional binary data using NeuroScale in Subsection 4.3.5, the visualisation of both high-dimensional and low-dimensional numerical data is discussed. Finally, as a supplementary analysis, the visualisation results of generated synthetic data are provided in Subsection 4.3.6.

4.1 High–dimensional data visualisation methods

Data visualisation is an important means of extracting useful information from large quantities of raw data. When such data is in a high dimensional space, data visualisation makes the data more understandable to researchers and helps to unveil some properties within the data that are difficult to observe in the high dimensional space. Various techniques for dimensionality reduction have been developed, which are increasingly essential in analysing biology related data.

According to the structural properties of data, the data transformation can be divided into two classes: linear and non-linear. Principal Component Analysis (PCA)(Pearson, 1901) is a classical linear projection method. As the most commonly used feature extraction and visualisation technique, it is widely applied in practice due to its speed and easy to implement advantages in computing. However, this method is only suitable for the linear datasets. The introduction of the non-linear dimensionality reduction techniques

effectively remedy this drawback. The non-linear methods can be broadly classified into two groups based on their functions. One group focuses on mapping the data either from the high dimensional space to the low dimensional embedding or vice versa. Another group just provides a visualisation. A visualisation method required in this work ought to retain the structure of the high dimensional dataset in the low dimensional projection space.

Topographic models based on the conception of topographic mapping is considered in this work. Generally, the topographic models can be subdivided into deterministic projection methods and probabilistic and generative models (Sivaraksa, 2008). Generative models such as Generative Topographic Mapping (GTM) (Bishop et al., 1998) and Stochastic Neighbour Embedding (SNE) (G.E.Hinton and Roweis, 2002), use probabilistic intuition by assuming a Gaussian distribution centred around each data point. On the contrary, the deterministic methods provide more direct projections without use of distributions over generator space. In addition, the approaches can also be categorised into global and local techniques, each of which has pros and cons (Silva and Tenenbaum, 2003). Local algorithms such as Locally Linear Embedding (LLE) (Roweis and Saul., 2000), Laplacian Eigenmaps (Belkin and Niyogi, 2001), attempt to preserve the local geometry of the data by seeking to map neighbouring points on the manifold to nearby points in the low-dimensional representation. Global methods such as NeuroScale (Lowe and Tipping, 1997) and Isomap (Tenenbaum et al., 2000) preserve geometry at all scales. This means overall properties and structure are retained while local models sometimes may not reflect global metric properties. However, the representational capacity of local methods make them attractive, when the intrinsic distance is different from the global metric properties.

Specific to this work, the selected approach should be capable of projecting new unseen binary data. Although both probabilistic based GTM and SNE can preserve the topology of the data, they are weak on projecting sparse binary data. On the contrary, a latent-variable density model (Tipping et al., 1999) which is also based on the distribution of the two-dimensional latent variable vector, is proposed to visualise high dimensional binary data. However, since this distribution must be a priori specified, it is not suitable for visualising new unseen data directly. NeuroScale which has the advantage of pre-

serving the structure of the data by measuring dissimilarities between the data samples becomes the most appropriate approach to project such special datasets. In Section 4.2, characteristics of the converted DNA-binding protein dataset will be reviewed, followed by a discussion of selected dimensionality reduction methods based on their visualisation results of the dataset.

4.2 DNA-binding protein interaction information visualisation

Although there are various dimensionality reduction methods that can represent the high-dimensional data into the low dimensional space, it is difficult to achieve satisfying visualisation results for the DNA-binding protein interaction information due to its specific structure. In the previous section, various visualisation methods have been discussed. The purpose of this section is to analyse the characteristics of the created database and discuss the visualisation results of selected standard visualisation models based on the database.

4.2.1 Characteristic of data

As introduced in Subsection 2.1.1, the Cys_2His_2 zinc fingers recognise specific DNA sequences via the amino acids on the surface of the α helix to contact the nucleotides within four base subsites in the target DNA sequence. Given that any one of 20 amino acids may preferentially bind to one of four bases A, C, G, T, there are 80 possible combinations for each binding site on the helix which can be represented as a 1-from-80 binary coding scheme. Therefore, using 4 sites in the canonical model gives rise to 320 binary dimensions where only 4 '1's are present for the specific interaction positions and 316 '0's. Moreover, as mentioned in Section 3.2.2, since there are 20 naturally occurring amino acids and 4 positions within a zinc finger that interact with 4 bases in a DNA sequence, a synthetic database with almost 41 million data examples are generated and will be used to verify and explain the visualisation results based on the experimental data samples.

In this chapter, the created database DB1 defined in Table 3.3 is selected to study

the high dimensional structural relationship. There are only 1860 available data examples which is 0.0045% of all potential binding sites [1] in this database, due to the limited number of real data sources, and the restriction from the binding status requirement. Therefore, the database DB1 can be represented as a 1860×320 matrix. The relevant visualisation results will be discussed in the following sections.

4.2.2 Visualisation results from standard visualisation techniques

At the beginning of this chapter, the techniques for visualising the high dimensional data have been discussed. By studying the characteristics of the converted database in the last subsection, some standard visualisation techniques are attempted to implement the visualisation of this database.

Principal Components Analysis (PCA)

Principal Components Analysis (PCA) is a technique to reduce a large number of correlated variables into a new set of uncorrelated variables (Jollife, 2002). These uncorrelated variables are called Principal Components (PCs) and the first few PCs are considered to retain the maximum variance of the original data. In this work, it is employed to extract the structural features from the 320 dimensional sparse binary database and project the data samples into a 2-dimensional feature space using the first two PCs. As shown in Figure 4.11, the proportion of variances which is explained by different number of PCs increases gradually from 9.72% to 100%. According to Table B.1 in Appendix B, the first two PCs only represent 16.10% feature information of the original data. Therefore, the PCA model is not suitable to visualise the database DB1 in the low-dimensional space, where the projection result is plotted in Figure 4.2. In this figure, the 1860 data samples are generally projected into four clusters. From each cluster, the data samples labelled with binding or non-binding are difficult to separate. Therefore, although the first two PCs retain the maximum variance of the dataset, they still fall short in describing the structural features of the dataset.

[1] 1860 data samples compare with 41 million possible binding sites.

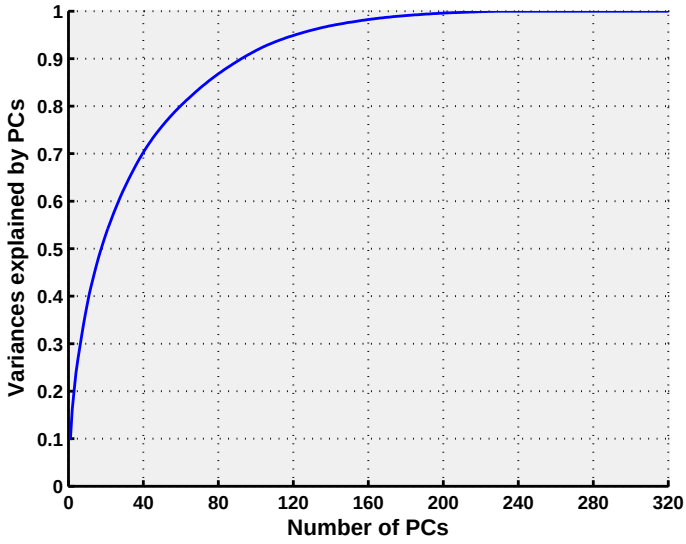


Figure 4.1: Variances explained by different principal components (PCs). In the figure, only 16.1% variances can be explained by the first two PCs. This proportion increases gradually and reach 100% when 232 eigenvectors are used. Relative information can be found in Appendix B Table B.1

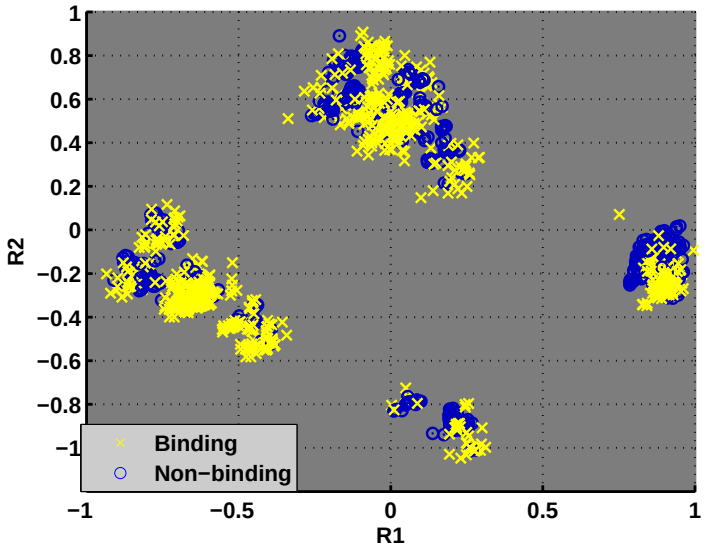


Figure 4.2: The visualisation of PCA. All data samples are generally projected into four clusters, but as in each cluster, the data samples with binding/non-binding status are overlapped, it is hard to study the structural properties.

Generative Topographic Mapping (GTM)

The Generative Topographic Mapping (GTM) is a probabilistic model which can project the data from a high dimensional data space into a low dimensional visualisation space by using a generative model, transforming from the latent variable to the data space by

using a Radial Basis Function (RBF) network (Bishop et al., 1998). This algorithm is based on the constrained mixture of Gaussians, and uses the Expectation Maximisation (EM) algorithm to optimise the parameters of the Gaussians. Because of the probabilistic approach in the GTM model, it is more tolerant to noise in the data. However, since the number of the RBF basis functions and distribution of the latent space sample points are chosen by hand, the visualisation result strongly depends on the choices of these parameters. Figure 4.3 shows the visualisation result based on the GTM approach. This figure is plotted with magnification factors which is used to ensure that the projected data samples can be well represented. Referencing the colour bar on the right hand side, the areas with white colour indicate high probability or vice versa. Since there are many data samples projected on the same point, it is difficult to distinguish them and study the relevant structure properties.

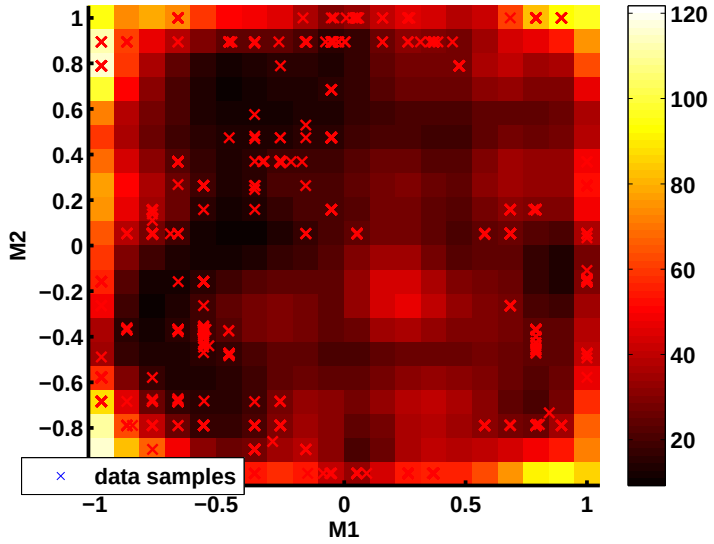


Figure 4.3: The visualisation result of GTM. The cross denotes the 1860 data samples. In this figure, the white area indicates high probability or vice versa. There are many data samples are projected into the same location which is hard to study the structure properties of them.

Locally Linear Embedding (LLE)

Locally Linear Embedding (LLE) (Roweis and Saul., 2000) is a local method which focuses on preserving the topographic distance in small neighbourhoods by using an eigenvector method (Saul and Roweis, 2003). It begins by finding a set of the K nearest neighbours for each point. Then, it computes a set of weights for each point that can

best describe the point based on these K nearest neighbours. Finally, the eigenvector-based optimisation technique is applied to find the low dimensional embedding of points. In this algorithm, the number of neighbours per data point, K, is a key parameter to be defined. Higher values of K cause the algorithm to be more similar to the PCA model. Otherwise, it will be hard to preserve the topographic structure of the data point in the low dimensional space. In this experiment, as shown in Figure 4.2, a few eigenvectors are unable to represent the majority of information of the original data, the visualisation result of the LLE model which depends on the selection of the eigenvectors is also affected [2]. Through comparing the visualisation results with respect to the number of neighbours, Figure 4.4 plots the best projection result with 12 neighbours in the LLE algorithm for the DNA-binding protein database. In this figure, due to most data samples concentrated around the origin point, the structural relationships between the data samples can not be confirmed.

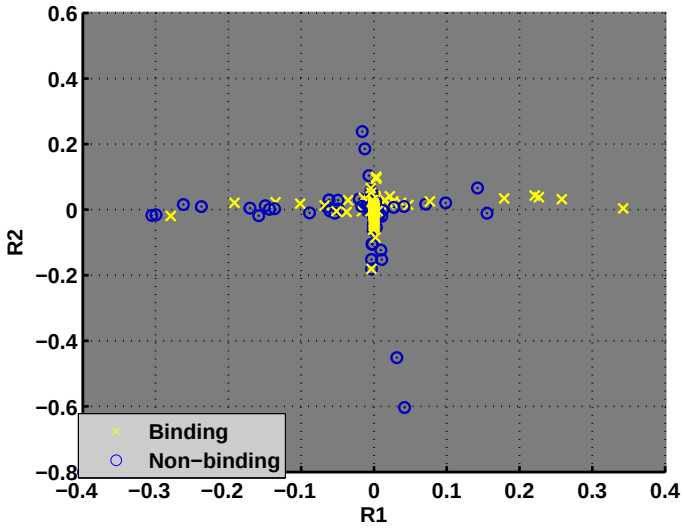


Figure 4.4: The visualisation result of LLE using 12 nearest neighbours. All data samples are projected into a cross and most points concentrate in the origin point which is hard to investigate the structure distribution.

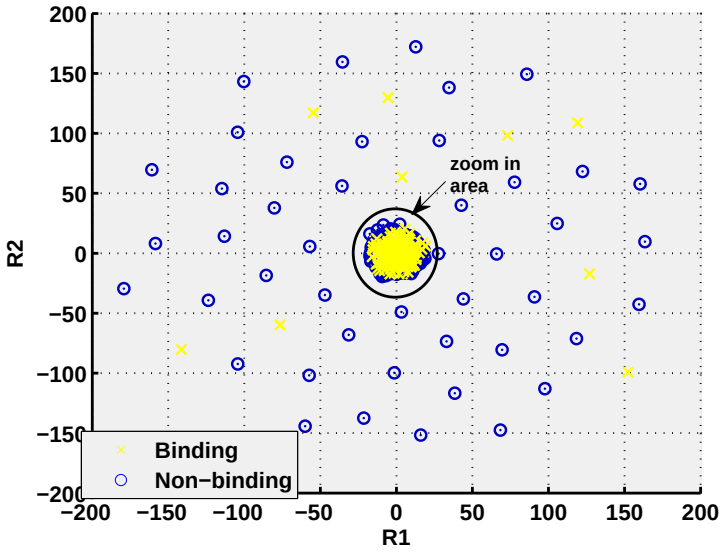
Stochastic Neighbour Embedding (SNE)

Stochastic Neighbour Embedding (SNE) (G.E.Hinton and Roweis, 2002) is a non-linear dimensionality reduction method which measures dissimilarities between points using a probabilistic distance approach to preserve the neighbourhood identities. A Gaussian

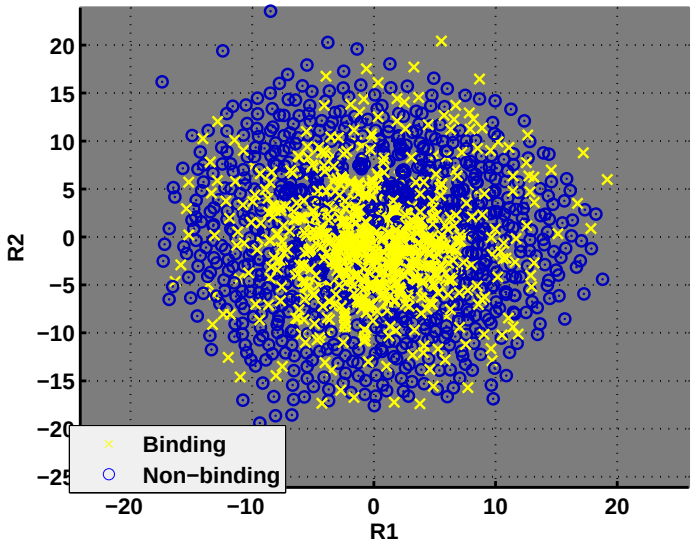
[2] Details of the LLE model can be found in Appendix B.1.2.

distribution is centred on each data sample in the high dimensional data space and a probability distribution is defined over all the potential neighbours of the point. This approach permits a 1-to-many mapping of the high dimensional data samples to the projection space (Sivaraksa, 2008). In accordance with the probability distribution [3], the high dimensional space is determined by the dissimilarity which can be scaled by a smoothing factor σ_i. If the value of σ is too large, the projection data is likely to collapse to a single point. However, there lacks a well defined approach of determining the smoothing factor. In this work, Figure 4.5(a) presents the visualisation result based on the SNE algorithm. Although the σ is adjusted to 5, most of data samples are still projected into a small area. By zooming in this area, the detailed structure can be checked in Figure 4.5(b). Compared to the data samples labelled as non-binding, the points with a binding symbol are closer to the origin. However, it is impossible to obtain any structural properties from the figure.

[3]The probability distribution is defined as $p_{ij} = \frac{\exp(-d_{ij}^2)}{\sum_{k \neq i} \exp(-d_{ik}^2)}$ where i is data point in the high dimensional space, j denotes each potential neighbour. $d_{ij} = \frac{\|x_i - x_j\|}{\sigma_i^2}$ is the dissimilarity between each point and its neighbours.



(a) In this figure, most of data samples are projected into a small area which is hard to obtain the detailed distribution of them.



(b) This figure plots the zoomed in area. Comparing with the data samples with non-binding status, most of binding samples concentrates in the centre of the visualisation area, but no cluster information can be found from this result.

Figure 4.5: The visualisation result of SNE using $\sigma = 5$. (a) is the visualisation result and (b) plots the zoomed in area.

Sammon mapping

The Sammon mapping (Sammon, 1969) is an algorithm that maps data samples from high dimensional space to a space of lower dimensionality by minimising the differences

between the corresponding inter-point distances in the two spaces. Unlike traditional linear dimensionality reduction techniques (such as PCA), the Sammon mapping does not explicitly represent the transformation function. Instead, it provides an error function that is defined as

$$E = \frac{1}{\sum_{i<j} d^*_{ij}} \sum_{i<j}^{N} \frac{\left(d^*_{ij} - d_{ij}\right)^2}{d^*_{ij}}. \tag{4.1}$$

where the distance between ith and jth data points in the original space is denoted by d^*_{ij}, and d_{ij} is their distance in the projection space. To minimise the error, gradient descent can be applied. The Sammon mapping algorithm uses the first two Principal Component from PCA as an initial configuration [4], and gradient descent is used to minimise the error. This approach is not sensitive to the dimensionality, as it only depends on the measured dissimilarities between the data samples which is irrelevant to the dimensionality of the original data. The relevant projection result is plotted in Figure 4.6. In this figure, the 1860 data samples are projected into different clusters which may reflect distinctive structural properties. In the next section, NeuroScale, a Sammon mapping related algorithm will be introduced and the related visualisation results will be discussed.

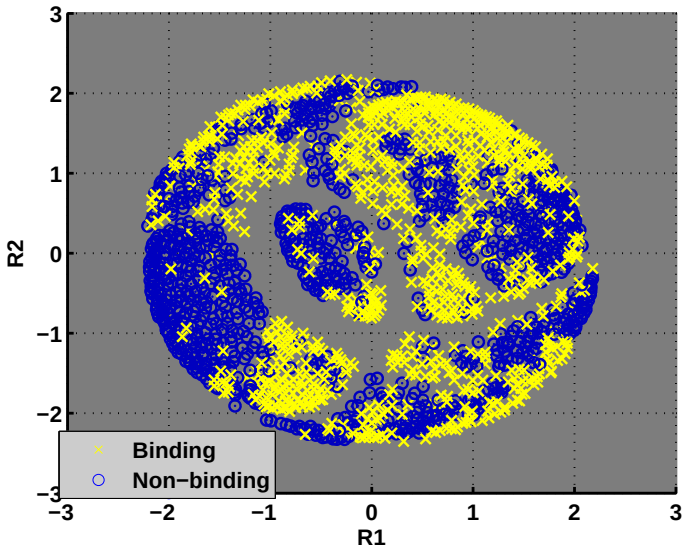


Figure 4.6: The visualisation result of Sammon mapping. In the figure, all data samples are projected into different clusters which may reflect distinctive structural properties. The data samples with binding/non-binding status can not be separated based on the result.

[4] Similar to the PCA model, only 9.72% of the variance can be explained by the first PC.

4.3 NeuroScale in DNA-binding protein interaction information visualisation

The characteristics of the database (Database DB1) for the visualisation study have been discussed in Subsection 4.2.1. Through analysing the visualisation results of such a dataset using the selected dimensionality reduction methods, it was discovered that the Sammon Mapping provides the best projection result for the high dimensional sparse binary database. According to the objective of this work to develop a system that can indicate the structural properties of the novel data samples, the relevant visualisation method is expected to have the capability of representing the new data without re-training the model. Therefore, the Sammon Mapping-related NeuroScale approach is exploited as a topographic feature extraction method to visualise the protein-DNA interaction data samples in this work. During visualisation, the class of non-linear parametrised transformations provided by Radial Basis Function (RBF) networks is chosen, and the model parameters are optimised through minimising the *Sammon stress metric* (Sammon, 1969) which will be introduced in Section 4.3.2. In general, the metric is developed based on the error function 4.1 mentioned in Section 4.2.2.

In this section, the visualisation mechanism of the NeuroScale model will be explained in Subsection 4.3.1. Then, the preconditioning of the RBF network and *Sammon stress metric* as the quality criteria will be introduced, which is followed by the discussion of different dissimilarity measurements employed to describe the geometric structure of the input data. Since the protein-DNA interaction data samples are represented in a high dimensional sparse binary matrix, it is essential to study the visualisation results of numerical data using the NeuroScale approach before discussing the projection results of the selected DNA-binding protein database.

4.3.1 Visualisation mechanism of NeuroScale

The NeuroScale approach, as discussed in Section 4.1, is a topographic feature extraction method which employs a nonlinear transformation $\{ f : \mathrm{boldsymbolR}^{\mathrm{d}} \rightarrow \mathrm{boldsymbolR}^{\mathrm{m}} : f(\mathbf{x}) =$

$\mathbf{y}\}$ from the original configuration space that maps into the feature space. The architecture of this model is shown in Figure 4.7.

In this figure, the input data $\mathbf{x}_i$ is projected into the transformed feature space as $\mathbf{y}_i$ by a class of non-linear parametrised transformations provided by Radial Basis Function (RBF) networks (Lowe and Tipping, 1997). The advantage of this approach is that a transformation can be obtained, while interpolations still allowed. Since the weights in the output layer of the RBF model are used to indirectly determine the location of the feature points, the method of initialising the weights has to be decided. There are two choices available: randomly generated or using Principal Component Analysis (PCA) to project the input data $\mathbf{x}$ and find the output layer using a least squares fit (Nabney, 2002). In this work, both of them have been attempted and evaluated by the *Sammon stress metric* (defined as: *STRESS* value). By comparing the *STRESS* value, the PCA algorithm is selected to initialise the weights at the beginning of the visualisation process. Then, the temporary points $\mathbf{y}$ are generated by the RBF network, given the data points as input. That is, $\mathbf{y}_q = f\left(\mathbf{x}_q;\theta\right)$, where f is the non-linear transformation effected by the RBF model with parameters, i.e. output layer weights and kernel smoothing factors, θ. The model parameters are adjusted to minimise the global *STRESS*: $E_{sam} = \sum_{i=1}^{N}\sum_{j>i}^{N}\left(d_{ij} - d_{ij}^{*}\right)^2$, where $d_{ij}^{*} = \| \mathbf{x}_i - \mathbf{x}_j \|$ are the distances between data points in the original space and $d_{ij} = \| \mathbf{y}_i - \mathbf{y}_j \|$ are the distances in the transformed space. Considering the non-linear transformation and the relevant parameters, the (squared) 'distance' in the feature space may thus be given by

$$d_{ij}^2 = \| f(y_i) - f(y_j) \|^2 = \sum_{l=1}^{n}\left(\sum_k w_{lk}\left[\phi_k\left(\| \mathbf{x}_i - \mu_k \|\right) - \phi_k\left(\| \mathbf{x}_j - \mu_k \|\right)\right]\right)^2 \qquad (4.2)$$

where ϕ_k are the basis functions of the RBF network, μ_k are the fixed centres of those functions [5], and w_{lk} are the weights from the basis functions to the output (output layer weights) (Lowe and Tipping, 1997).

Since the topographic nature of the transformation is imposed by the *STRESS* term

[5] In this work, data samples in the original space are randomly selected to be the centres μ_k of those functions.

which attempts to match the inter-point distances in the feature space with the dissimilarities in the input space, there is no specific target for each y_i where a relative measure of target separation between each $(\mathbf{y_i}, \mathbf{y_j})$ pair is provided. Therefore, no 'target' information such as binding properties is required as only the distance measurement between points is used, meaning that the training set makes no assumptions as to the binding. Moreover, as the dissimilarity measurement is irrelative with the dimension of the input data, this model is not sensitive to the high dimensionality.

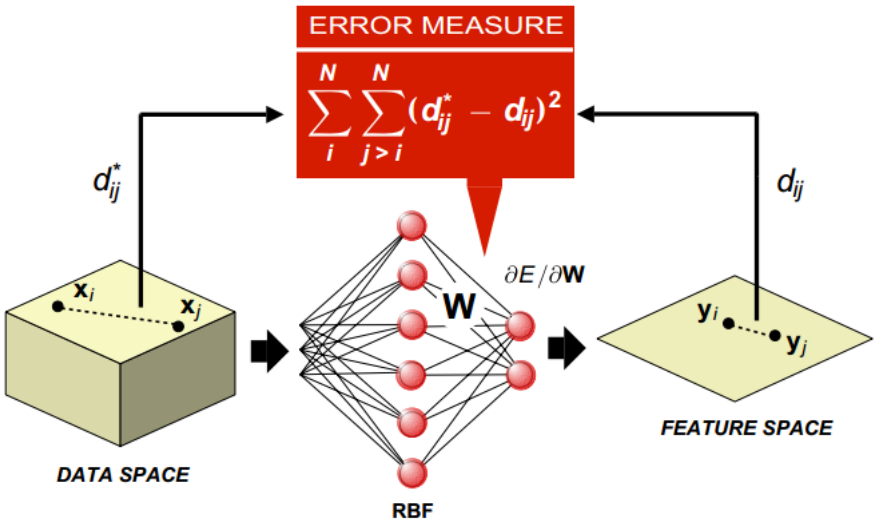


Figure 4.7: The NeuroScale architecture (Lowe and Tipping, 1997).In this figure, the RBF model is used to implement the projection function. The relevant parameters such as weights and any kernel smoothing factors, are optimised by minimising *Sammon stress metric* between the inter-point distance d_{ij} in the feature space and the distance d_{ij}^{*} in the input space.

In next subsection, the preconditioning of the RBF network and the *Sammon stress metric* will be introduced, where the latter will be applied to optimise the relevant parameters of the nonlinear transformation. In addition, the distance measurement as the key factor in the projection process will be discussed in Subsection 4.3.3.

4.3.2 Preconditioning and quality criteria

In the process of building the NeuroScale model, the RBF network is used to transform the represented data samples in the 320 binary dimensions to a corresponding set of feature vectors in a two-dimensional space. Moreover, the quality of the projected feature vectors

is measured by the *Sammon stress metric*. The aim of this subsection is to discuss the preconditioning of the RBF by using the Principal Component Analysis (PCA) and how the *Sammon stress metric* works to control the quality of the visualisation results.

Principal component analysis (PCA) in RBF

From the definition of the RBF network, it comprises a single hidden layer of *h* neurons which represents a set of basis functions, each of which has a centre selected from the input data examples in this work. The hidden units implement a radial activated function. The output is a weighted sum of the hidden unit outputs. In NeuroScale, the output layer weights of the RBF are optionally initialised using a principal component projection of the training data to set an initial projection of patterns. Otherwise an initial random choice of projections can be made. In this work, both of them are attempted, and the relevant visualisation results are evaluated by the *Sammon stress metric*. The advantage of applying PCA to initialising the weight is to shorten the number of optimisation steps because the initial value the *Sammon stress metric* is closer to the minimum value. On the contrary, the randomly selected weight usually causes a much worse maximum value at the beginning of the optimisation, but may have a better minimum *STRESS* value for the visualisation.

Sammon stress metric

In order to project the data, the model parameters controlling the behaviour of the RBF network which govern the position of the projected patterns, **y**, are adjusted to minimise the *Sammon stress metric* . The stress metric is expressed as :

$$E = \sum_{p=1}^{P} \sum_{q<p} \left[d_n(\mathbf{p}, \mathbf{q}) - d_2(\mathbf{p}, \mathbf{q}) \right]^2 \tag{4.3}$$

where $d_n(p,q) = \|\mathbf{x}_p - \mathbf{x}_q\|$ is the distance between data points in the original space and $d_2(p,q) = \|\mathbf{y}_p - \mathbf{y}_q\|$ are the distances in the transformed space. The topographic nature of the transformation is imposed by the *STRESS* term which attempts to match the inter-point distances in the feature space with the dissimilarities in the input space. Specific to the DNA-binding protein interaction, only a relative measure of target separa-

tion between each $(\mathbf{y}_q, \mathbf{y}_p)$ pair is provided, and no 'target' information such as binding properties.

4.3.3 Dissimilarities

As mentioned in Subsection 2.2.1, the dissimilarity measure as a metric[6] is a concrete way of describing the similarity between two data samples. In NeuroScale modelling, the dissimilarity measure is used to represent the structural relationship between the DNA-binding protein interactions. In general, many dissimilarity measures have been proposed for distance measurement (Webb, 1999), such as Euclidean distance, City-block distance, p norm distance and Minkowski distance. The Euclidean distance or Euclidean metric is the 'ordinary' distance between two points that can be measured by the Pythagorean formula: $d_e = \left[\sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{y}_i)^2\right]^{\frac{1}{2}}$. As the distance measure of the Euclidean metric complies with human visual experience, it is widely used in the dissimilarity measurement. The City-block distance, which is also known as the Manhattan or box-car or absolute value distance is defined as the sum of the differences of the corresponding components of two points: $d_{cb} = \sum_{i=1}^{n} \mid \mathbf{x_i} - \mathbf{y}_i \mid$(Webb, 1999). This metric is suitable for calculating the distance between points that follow a grid-like path. The p norm or L^pnorm distance is a more general form of the Euclidean and City-block distances. For a real number $p \geq 1$, the metric is $d_p = \left(\sum_{i=1}^{N} \mid \mathbf{x}_i - \mathbf{y}_i \mid^p\right)^{\frac{1}{p}}$ where the City-block metric is equivalent to $p = 1$, and the Euclidean metric is equivalent to $p = 2$. Finding an appropriate value for p depends on whether the large difference is preferred. Normally, larger value of p gives progressively more emphasis towards the larger differences.

Different from the metrics explained above, the Minkowski distance or Minkowski inner product is based on the concept of the Minkowski space or Minkowski spacetime[7] in mathematical physics. To unify space and time, the Minkowski metric is defined as $(x_1, x_2, x_3, x_4) = (x, y, z, ict)$[8] where c is the speed of light (Petkov, 2010). For two events, the separation between them is measured by the interval between the two events, which

[6] The dissimilarity can also be understood as distance.

[7] Different from the three ordinary dimensions of space, the Minkowski space also has one timelike dimension.

[8] x, y, z are the three variables in the space, t is the time.

take into account not only the spatial separation, but also the temporal separation. The interval, s^2, between two events is defined as:

$$s^2 = \triangle \mathbf{r}^2 - c^2 \triangle t^2 \tag{4.4}$$

where $\triangle \mathbf{r}$ and $\triangle t$ denote differences of the space and time coordinates, respectively(Petkov, 2010). The choice of signs for s^2 follows the space-like convention (- + + +)[9].

Specifically to the dissimilarity measure in the NeuroScale model, since the radial basis function used in the Neuroscale model can deploy non positive definite metrics and basis functions, the inner product can be negative. Considering the properties of the zinc finger-DNA interactions, the Minkowski inner product as well as the classical Euclidean metric as a benchmark to measure the dissimilarities are exploited to describe the structural relationships between the DNA-binding protein interactions. As introduced in Subsection 2.1.2, the sequence recognition is mediated mainly by an amino acid in positions -1, 3 and 6 of the α helix, and the amino acid at position 2 contacts to the complementary strand of the DNA to stabilise the interaction. To reflect this property in the dissimilarity metric, the Minkowski indefinite inner product is selected, where the dimensions of the input space corresponding to connections to the complementary DNA strand are weighted with -1 and the connections related to connections to the primary DNA helix are weighted with +1. In this work, by fixing the weights for the connections in the primary strand, the weight for the connection in the complementary strand are adjusted from 0 to -3. This range is defined by considering the function of each connection position in the interaction. If the weight is set to 0, the visualisation model ignores the contribution from position 2 for the interaction. If the weight is -3, the binding pair at this position is considered to have equal effect on the interaction with other binding positions in the primary strand. Through verifying the global STRESS and comparing the projection results, the weight for position 2 is defined as -1. Moreover, cubic basis functions are used for the interpolation model inside NeuroScale.

Since the input dataset is a 320 dimensional sparse binary matrix, the dissimilarity

[9]The squared differences in the space coordinate are defined as positive, where the difference in the time coordinate is negative.

measures for binary variables are also considered, such as the Hamming distance. The Hamming distance $d(\mathbf{x},\mathbf{y})$ between two vectors $\mathbf{x}$, $\mathbf{y} \in R^d$ is the number of coefficients in which they differ. As each data sample in this work is converted into a 1×320 vector in which only four '1's are present for the specific interaction positions and the rest are all '0's. The dissimilarities between two data samples which are calculated by the Hamming distance can only be 0 or 2 or 4 or 6 or 8, which is same as the squared Euclidean distance. Using the NeuroScale model for further investigation, the projection result is same as the Euclidean metric based result. Due to the particularity of the DNA-binding zinc finger protein interactions, any dissimilarity metric discussed above can be exploited in the model. Moreover, some additional metrics (i.e. Bregman divergence) can also be explored for the visualisation study (Sun, 2011).

4.3.4 Computational Methodology

Specific to this work, the implementation process of the projection is presented in Figure 4.8. As mentioned in subsection 4.3.1, the created 320 dimensional sparse binary database DB1 is applied as the training data set for the visualisation model. Then the RBF network is chosen to predict the coordinates of the data point in the transformed feature space. To initialise the weights of the RBF model, the PCA algorithm is applied. Through minimising the *STRESS* value, relevant parameters such as output layer weights and kernel smoothing factors, are optimised. One novelty of this thesis is that different distance metrics are attempted to measure the dissimilarities between the data samples.

The results which will shown in following sections were all generated using bespoke Matlab code based on the Netlab library (Nabney, 2002). The code was modified to employ different metric functions in input (data) space and output (visualisation) space. The input data is taken from the generated datasets and the relevant database DB1, and calculated by the appropriate dissimilarity metrics. An additional dimension of colour is used in the output space to represent additional properties such as hydrophobicity and DNA labels in Subsection 4.3.6. However, these additional properties were not used as part of the metrics or as part of the learning process.

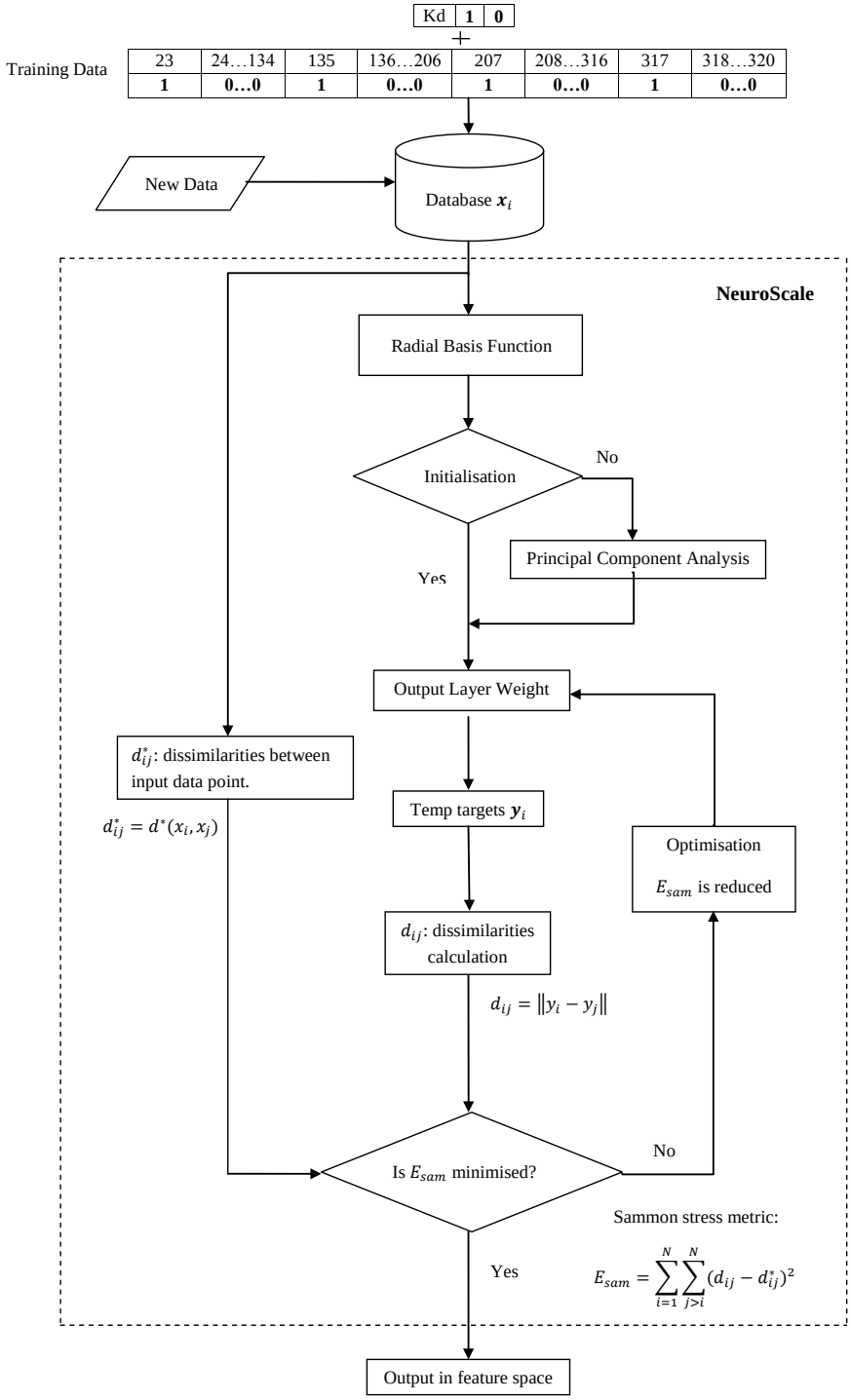


Figure 4.8: Visualisation mechanism of NeuroScale. The flowchart illustrates the process of visualising the given dataset by the NeuroScale approach. The RBF network is applied to implement the transformation, while the relevant parameters are optimised by minimising the *Sammon stress*. The PCA is used to initialise the output layer weights.

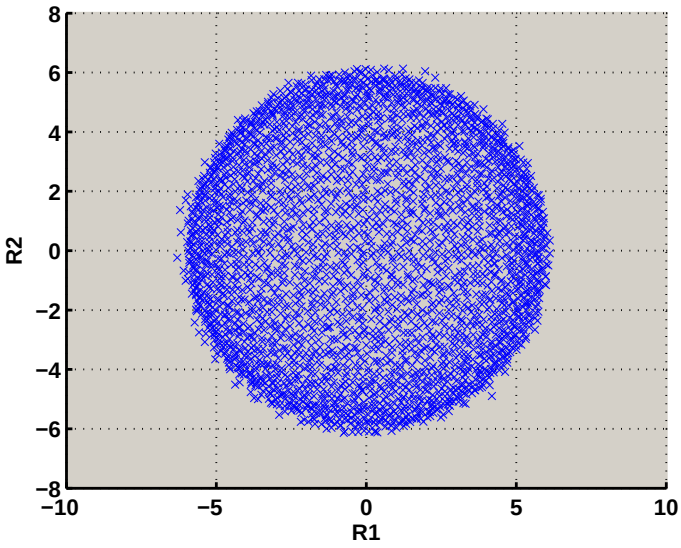
4.3.5 Discussion of numerical data visualisation

In the last subsection, various dissimilarity metrics have been discussed. Since the Euclidean distance and Minkowski metric are not usually defined for binary variables, it is worth examining the visualisation results of the numerical dataset based on the dissimilarity metrics before applying them in the NeuroScale model to implement the projection. In this subsection, two respective datasets with 320 dimensional and three dimensional numerical data samples are generated. Combined with the histograms of the distances between the data samples in the original space, the relevant visualisation results will be discussed.

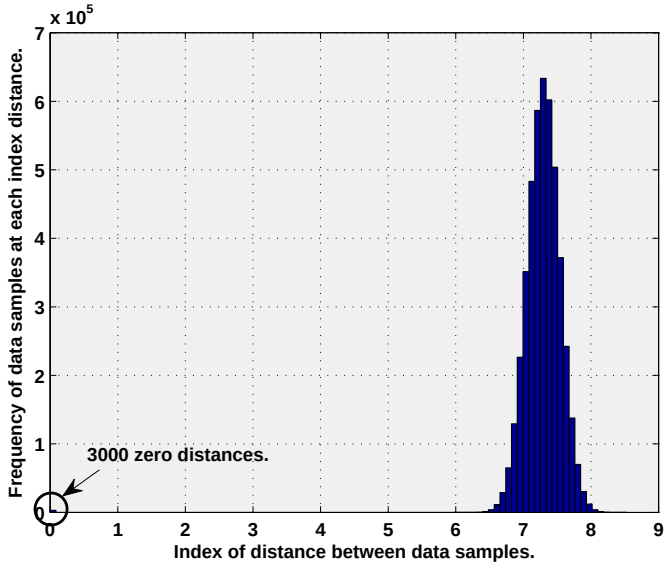
Visualisation of the 320 dimensional numerical dataset

For this experiment, a 320 dimensional numerical dataset is created, which includes 3000 data samples containing pseudo-random values drawn from the standard normal distribution. It is considered as a reference for the 320 dimensional sparse binary database. Figures 4.9 and 4.10 plot the relevant projection results with different dissimilarity metrics using the NeuroScale approach.

Figure 4.9(a) shows the representation result of the 320 dimensional numerical dataset where the dissimilarities in both original space and feature space are measured by the classical Euclidean metric. In the 2-D feature space, the projected data samples are distributed as a sphere, where the density of data samples is seen highest at the edge of the sphere and lower gradually towards the centre. Imagine that the 3000 data samples form a hypersphere in the data space, when the data samples have similar structure information, the dissimilarities between them can be very small, and vice versa. Applying the NeuroScale model to project these data by preserving the structure relationships, the data samples with similar structures in the original space should be projected into the same area. This expectation is supported by the histogram of the Euclidean distance between data samples in the data space shown in Figure 4.9(b). As highlighted in the corner of the figure, zero distances indicate the data samples themselves. The distances between them and other data samples vary between 6 and 9. At a distance of 7.4, there are more data samples than at any other distances.



(a) This is the visualisation of generated 320-D numerical dataset based on the Euclidean metric. In the figure, the projected data samples are distributed as a sphere, where the density of data samples is seen highest at the edge of the sphere and lower gradually towards the centre.



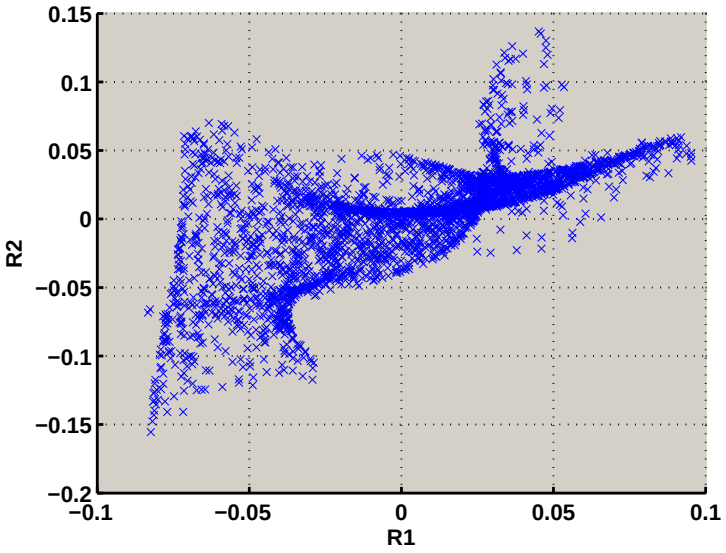
(b) This is the histogram of the Euclidean distance between the generated dataset in the 320-D data space. As highlighted in the figure, there are 3000 zero distances which indicate the data them sample themselves. The distances between the most of data samples are falling into the range from 6 to 9.

Figure 4.9: Analysis result of generated 320-D numerical dataset based on the Euclidean metric. (a) is the 2-D visualisation result and (b) is the relevant histogram of the Euclidean distance between data samples in the original space.

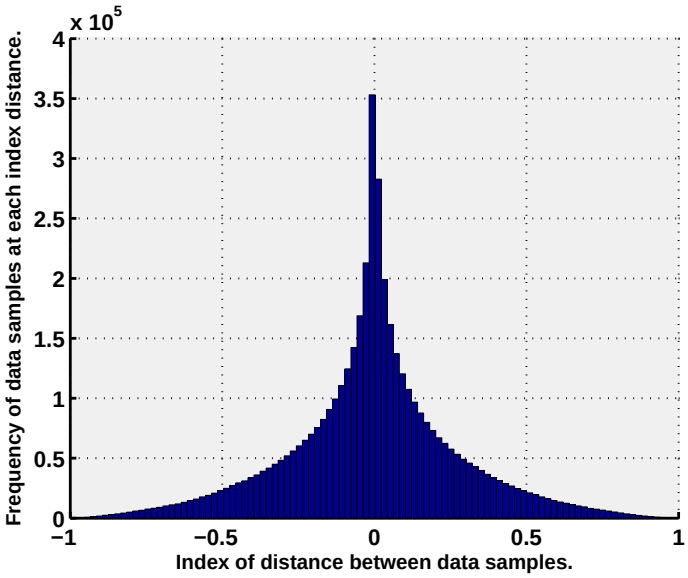
Compared with Figure 4.9(a), the Minkowski inner product based visualisation result

of the 320 dimensional numerical dataset is more complicated and elusive[10]. As shown in Figure 4.10(a), the plane in the projective space is twisted several times, and the densities of the represented data samples in the twisting areas are significantly higher than in other areas. Through verifying the relevant histogram of the Minkowski distance between data samples in the original space shown in Figure 4.10(b), most of the data samples are found to have similar structures with the range of the measured dissimilarities between -1 and 1. Since all visualisation results represented in this subsection are reliant on PCA to initialise the weights of the output layer of the RBF network, it needs some attention whether such initialisation affects the projection result. Therefore, the visualisation experiments without PCA are also conducted several times. By comparing the global *STRESS* between the models with and without the PCA initialisation, it is found that the *STRESS* values of the randomly initialised models are always higher than the PCA initialised the models. This indicates that the random weights initialisation has no beneficial effect on the projection result. However, there may still be some other factors that can cause this result such as the dissimilarity metric applied in the feature space and the selection of the smoothing function for the NeuroScale model. The investigation of these factors will be left as a direction of future research.

[10]The Minkowski metric applied here is defined as: $d_{ij} = -\sum_{k=1}^{80}\left(x_{ik}-x_{jk}\right)^2 + \sum_{k=81}^{320}\left(x_{ik}-x_{jk}\right)^2$ where k is the numbered coordinates. Specific to this work, the first 80 coordinates are weighted with -1, the remaining coordinates are weighted with +1.



(a) This is the visualisation of generated 320-D numerical dataset based on the Minkowski metric. In this figure, the plane in the projective space is twisted several times, and the densities of the represented data samples in the twisting areas are significantly higher than in other areas.



(b) This is the histogram of the Minkowski distance between the generated dataset in the 320-D data space. According to the figure, most of data samples have similar structure information as on the zero point the number of data samples reaches the highest value.

Figure 4.10: Analysis result of generated 320-D numerical dataset based on the Minkowski metric. (a) is the 2-D visualisation result and (b) is the relevant histogram of the Minkowski distance between data samples in the original space.

Visualisation of the three dimensional numerical dataset

Besides discussing the visualisation of high dimensional numerical dataset , the projection of the low dimensional data samples is also worth studying. Figures 4.12 and 4.13 plot the projection results of a three dimensional numerical dataset. Similar to the 320 dimensional numerical dataset, 5000 three dimensional data samples are also generated by containing pseudo random values drawn from the standard normal distribution. As in each coordinate, each data sample is between 0 and 1. All generated data samples are distributed in a cube as shown in Figure 4.11. When applying the Euclidean metric to measure the dissimilarities between these data samples in the original space, the histogram is presented in Figure 4.12(b). As before, zero distances indicate the data samples themselves. The distances between the most of data samples are falling into the range from 0.5 to 0.7. The relevant visualisation result is shown in Figure 4.12(a). In this figure, the distribution of the projected data samples is similar to a square which has a relatively clear contour. This result coincides with the structure relationship presented in Figure 4.11. For the Minkowski metric, since there are only three coordinates existing in the data space, the X coordinate is weighted with -1, when the Y and Z coordinates are both weighted with 1. Compared with Figure 4.10(b), the histogram of the distances between the three dimensional data samples is similar to that of the 320 dimensional dataset, but the visualisation result in Figure 4.13(a) is more reasonable. Since most of the data samples have zero distance from others, when projecting them into the feature space, the origin shows the highest density. With the expansion of the projection range, the density reduces gradually. By investigating the visualisation results of the numerical datasets, it becomes clearer about the effect of dissimilarity metrics on data visualisation. In the next subsection, the projection results of the 320 dimensional sparse binary database will be presented and discussed.

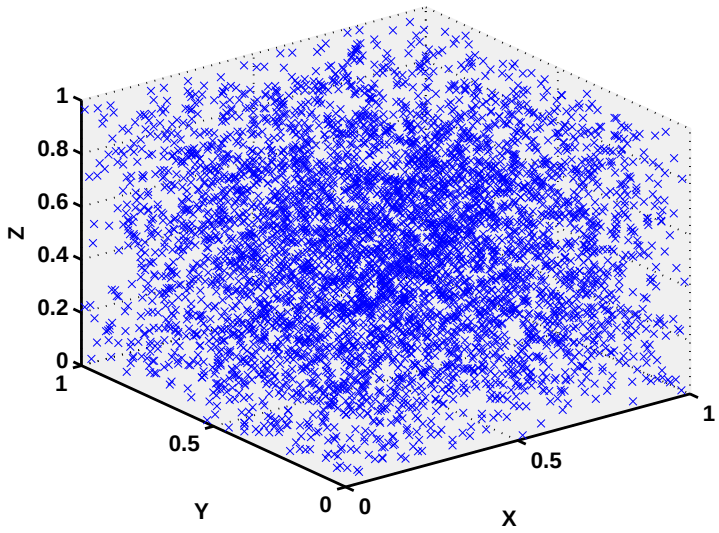
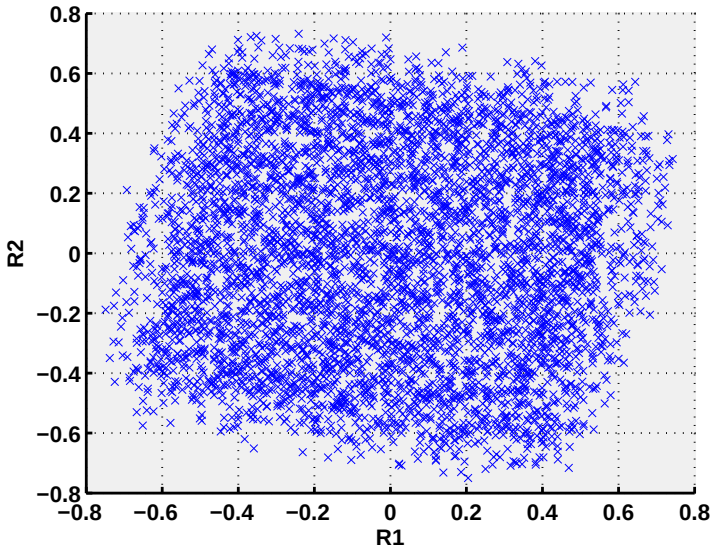
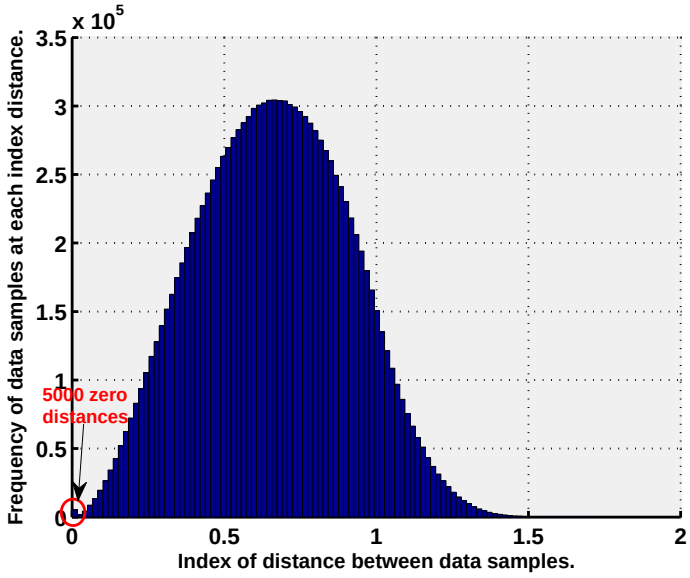


Figure 4.11: Distribution of the generated 3-D numerical dataset. All data samples distributed in a cube.

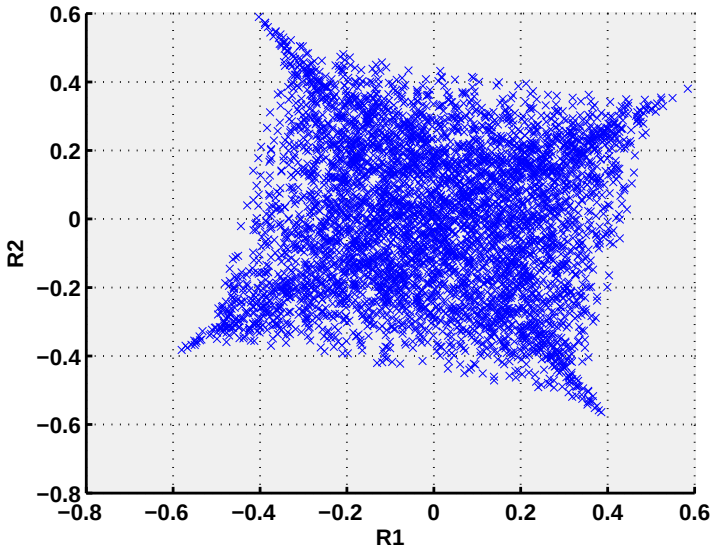


(a) This is the visualisation of generated 3-D numerical dataset based on the Euclidean metric. In the figure, the distribution of the projected data samples is similar to a square which has a relatively clear contour.

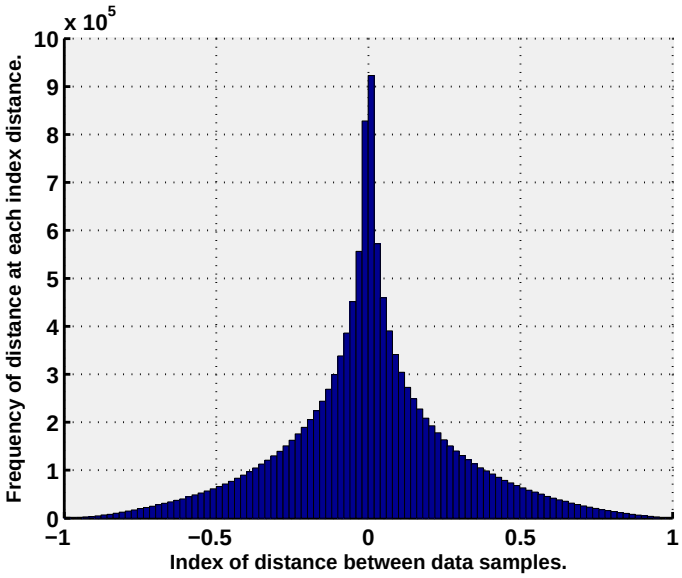


(b) This is the histogram of the Euclidean distance between the generated dataset in the 3-D data space. As highlighted in the figure, there are 5000 zero distances which indicate the data them sample themselves. The distances between the most of data samples are falling into the range from 0.5 to 0.7.

Figure 4.12: Analysis result of generated 3-D numerical dataset based on the Euclidean metric. (a) is the 2-D visualisation result and (b) is the relevant histogram of the Euclidean distance between data samples in the original space.



(a) This is the visualisation of generated 3-D numerical dataset based on the Minkowski metric. In the figure, lots of data samples are concentrated to the cross which is centred on the origin, and on the four edges, the densities of data samples are lower than the centre.



(b) This is the histogram of the Minkowski distance between the generated dataset in the 3-D data space. According to the figure, most of data samples have similar structure information as on the zero point the number of data samples reaches the highest value.

Figure 4.13: Analysis result of generated 3-D numerical dataset based on the Minkowski metric. (a) is the 2-D visualisation result and (b) is the relevant histogram of the Minkowski distance between data samples in the original space.

4.3.6 Visualisation results

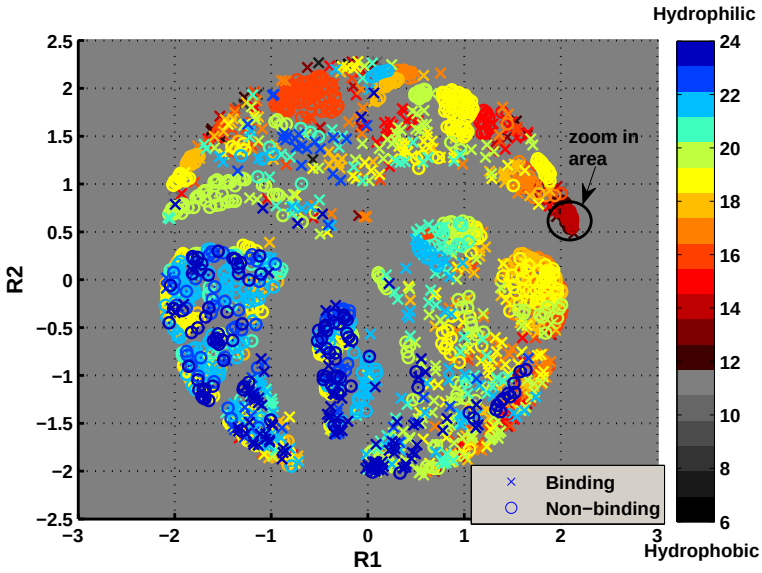
In Subsection 4.3.5 the visualisation results and relevant histograms of the generated numerical datasets in both the high-dimensional (320-D) and the low-dimensional (3-D) spaces have been presented and analysed as the references. The purpose of this subsection is to compare the resulting NeuroScale projections of the sparse binary DNA-binding Cys_2His_2 zinc finger interaction data samples in the 2-D Euclidean space by using either the Euclidean dissimilarity metric or Minkowski indefinite inner product in the input space. Moreover, through colouring the data samples based on different conditions, such as DNA sequence, amino acids combination and binding statues, some revelatory properties are shown in the results.

Visualisation results coloured based on amino acids

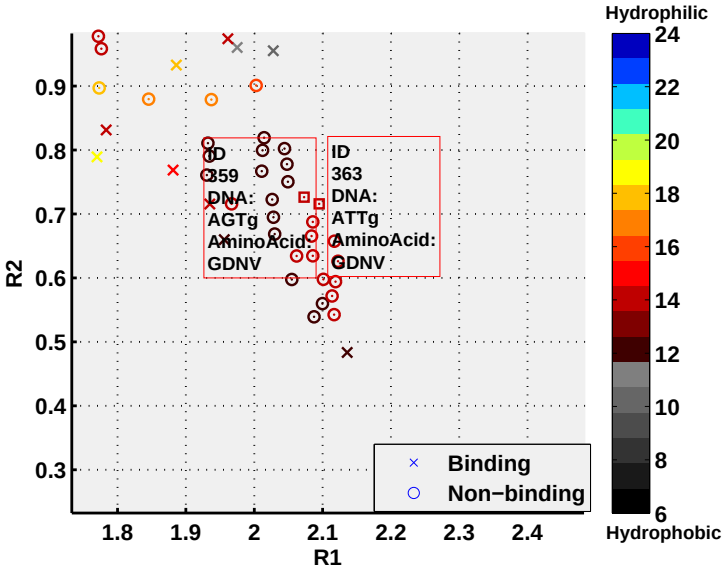
In chemistry, hydrophobicity is the physical property of a molecule that is repelled from a mass of water (Ben-Naim, 1980). In contrast, a hydrophilic molecule is the one that has a tendency to interact with or be dissolved by water and other polar substance (McNaught and Wilkinson, 1997). Specific to the Cys_2His_2 zinc finger, this pair of physicochemical properties can affect the protein interaction or forming a functional domain. Therefore, it is useful to observe the coloured result of the projected DNA-binding zinc finger interaction data samples based on the hydrophobic or hydrophilic properties of the amino acids combinations in the zinc finger proteins. Figures 4.14 and 4.15 show highly structured relationships for both different types of the metric space: the classical Euclidean metric and the Minkowski indefinite inner product. Moreover, in these figures, the relevant physicochemical properties of the amino acids combinations are defined in different colours for which the definition details can be found in Appendix B Table B.2.

Figure 4.14(a) plots the representation result of the 1860 data samples with the Euclidean dissimilarity metric applied in the input space. In the figure, the data samples with different structure properties are projected into different clusters. In addition, given the colour map provided on the right hand side of the figure, most of the zinc finger proteins with similar hydrophobic or hydrophilic properties can be generally represented in the same areas. For example, most of the data samples with high hydrophilicity are mainly

represented at the bottom left area. However, it is noted that some data samples which also have high hydrophilicity are projected external to the area. This phenomenon is considered to be caused by the similarities of the structures of the DNA sequences. To verify the inferences, the highlighted area in Figure 4.14(a) is zoomed in and plotted in Figure 4.14(b). From Figure 4.14(b), it is clear that the two selected data samples with the same physicochemical properties have the same amino acids combinations and similar DNA sequences. Therefore, for the data samples projected into the clusters which present different hydrophobic or hydrophilic properties, their structural features must be similar to the neighbours.



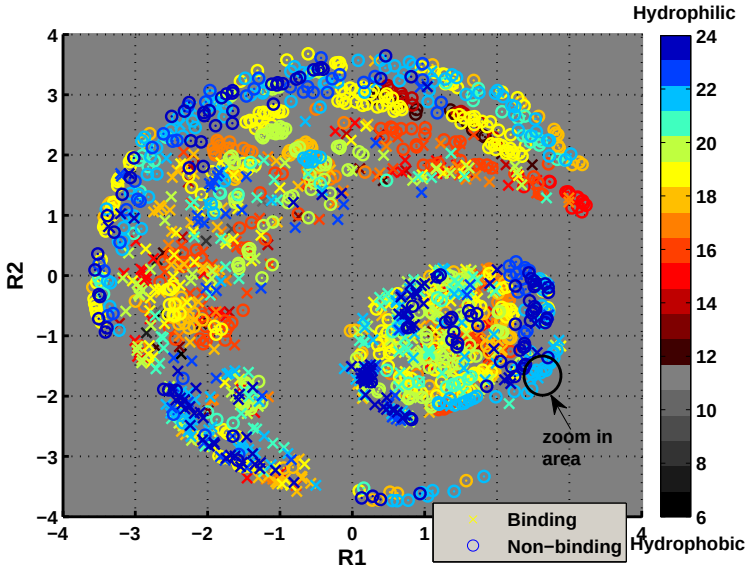
(a) This is the visualisation using the Euclidean metric in the data space to measure the dissimilarities. From this figure, it is found that proteins with high hydrophilicity properties are mainly projected on the bottom left of the main visualisation area, while the amino acid combinations with relatively lower hydrophilicity properties are represented at the top and right hand side.



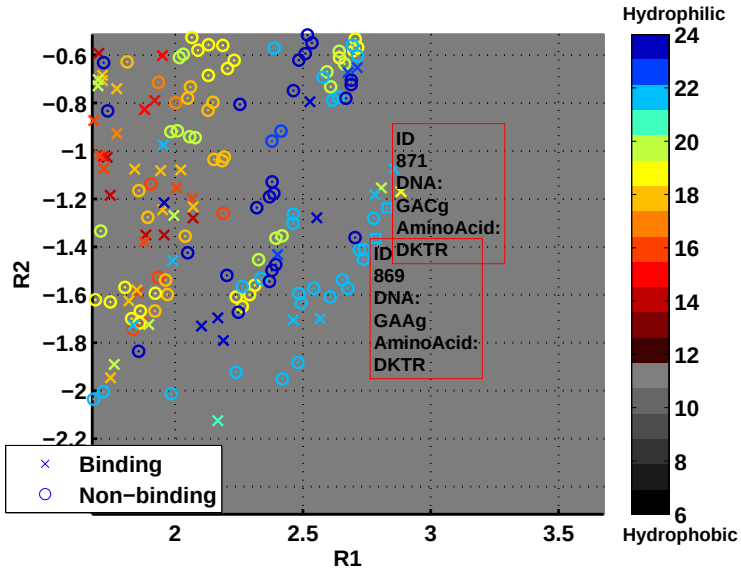
(b) In this figure, structure information of two adjacent data samples are presented. They have the same amino acids combinations (GDNV at position 2, -1, 3 and 6) and similar DNA sequences 5'-ANTg-3'.

Figure 4.14: The Euclidean metric based projection results colouring by properties of the amino acids combination. (a) is the visualisation results based on the Euclidean metric and (b) is the zoomed in visualisation result.

Compared with the visualisation result based on the Euclidean metric, the NeuroScale model using the Minkowski metric to measure the dissimilarities in the data space gives a strong different result as shown in Figure 4.15(a). Distinguished from Figure 4.14(a), the data samples are mainly represented into two clusters. However, the coloured distribution of the amino acids combinations shows irregularity. Nonetheless, by zooming in the circled area in Figure 4.15(a), the selected group of data samples still have similar structure information as illustrated in Figure 4.15(b), which means that the Minkowski metric based visualisation model can also represent the structure relationships in the high dimensional space properly, albeit with a different metric structure.



(a) This is the visualisation using the Minkowski metric in the data space to measure the dissimilarities. It is coloured by the hydrophobicity of the amino acids combination. Generally, data samples with different hydrophobicities appear in all clusters.

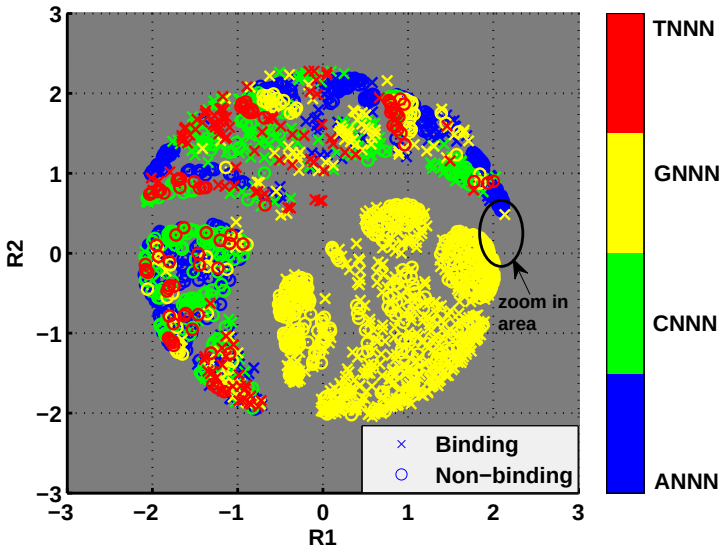


(b) In this figure, structure information of two adjacent data samples are presented. They not only have the same amino acid combination (DKRT), but also have similar information of the DNA sequence (5'-GANg-3').

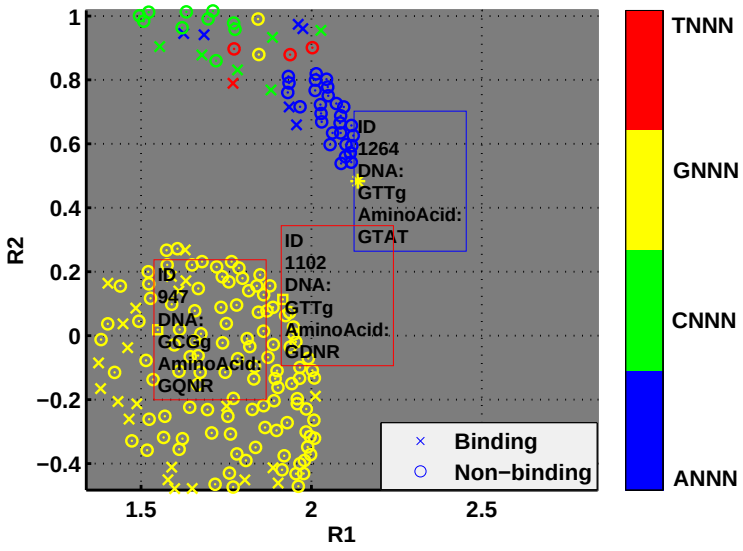
Figure 4.15: The Minkowski metric based projection results of colouring by properties of the amino acids combination. (a) is the visualisation results based on the Minkowski metric and (b) is the zoomed in visualisation result.

Visualisation results coloured based on DNA

Since the NeuroScale model is employed to represent the structure properties of the interactions between the DNA sequences and the zinc finger proteins, it is also worth investigating the coloured results based on the DNA sequences. Considering that four bases participate in the interactions, colour coding of the DNA sequence is defined from ANNN to TNNN which is in 5'-3' order, where each 'N' represents one of bases: A, C, G and T at the binding positions 6, 3 and -1. Figures 4.16 and 4.17 show the coloured projection results based on the information of the DNA sequences. It is interesting that the interactions in the bottom right area of Figure 4.16(a) and the circular 'thumbprint' of Figure 4.17(a) are to DNA sequence 5'-GNN-3' on the primary strand. By further investigating the information of the amino acids in Figures 4.16(b) and 4.17(b), the proteins in the areas containing R at the 6 position of the α-helix, also explains why some data samples with the same DNA sequences are projected into other clusters. However, the remaining data samples with other DNA sequences such as 5'-ANN-3', 5'-CNN-3' and 5'-TNN-3' are projected into the overlapping clusters, which is hard to distinguish. Moreover, the notable clustering of the Cys_2His_2 zinc finger-DNA combinations in the both visualisation results are separated by distinctive 'gaps'. These gaps are considered to reflect the lack of existence of certain types of combinations that do not naturally occur. This surmise will be verified in Subsection 4.3.7.

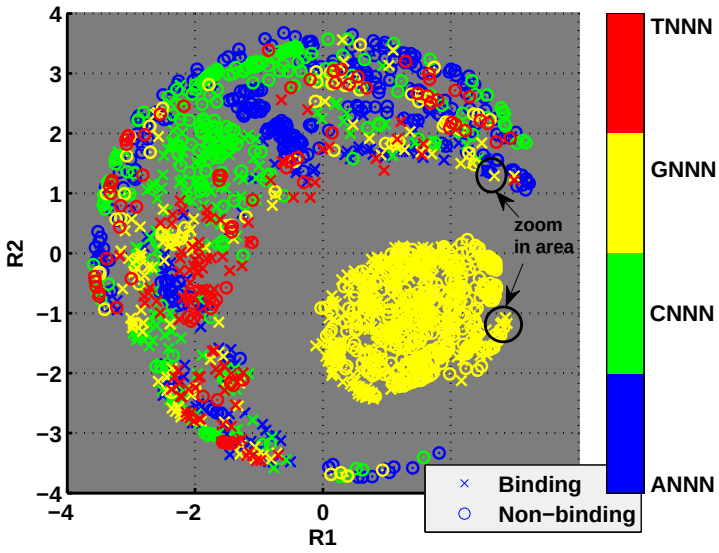


(a) This is the visualisation using the Euclidean metric in the data space to measure the dissimilarities. Except the group of clusters on the bottom right only contains the data samples with the DNA sequence '5'-GNN-3' on the primary strand, the remaining clusters have all kinds of DNA sequences.

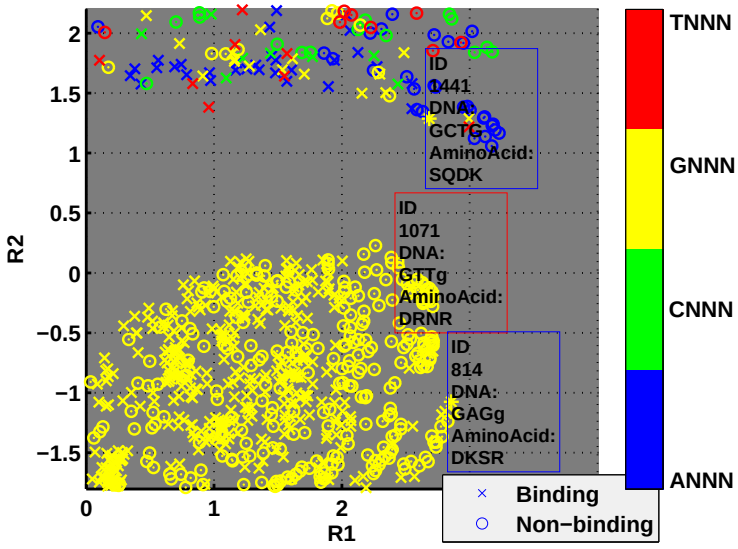


(b) In this figure, three data samples are selected to verify the structure information. Although the three samples have the same colour, even the samples: ID 1102 and ID 1264 have the same information of the DNA sequence, as the amino acid at the 6 position of the α helix are different (for ID 1102 is 'R', ID 1264 is 'T'.), they are projected into different clusters.

Figure 4.16: The Euclidean metric based projection results colouring by DNA sequences. (a) is the visualisation results based on the Euclidean metric and (b) is the zoomed in visualisation result.



(a) This is the visualisation using the Minkowski metric in the data space to measure the dissimilarities. It is coloured by the information of DNA sequence. It is notable that the smaller cluster on the bottom right only contains the data samples with the DNA sequence '5'-GNN-3' on the primary strand.

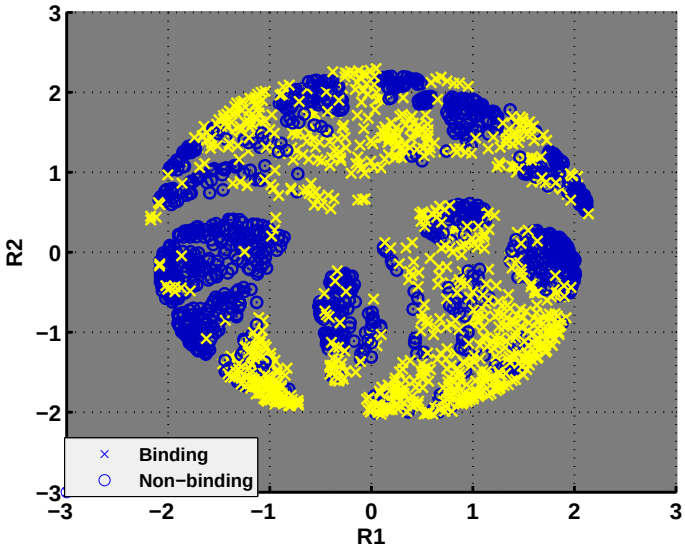


(b) In this figure, details of three selected data samples are presented. By comparing the relevant structure information, it is found that although they have same base at position 6 in DNA sequences, only the data samples which have amino acid 'R' at the 6 position of the α helix are projected into the circle cluster.

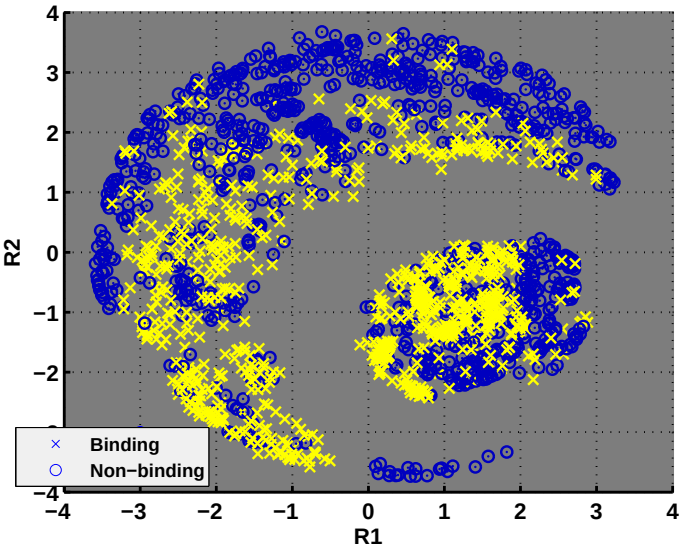
Figure 4.17: The Minkowski metric based projection results of colouring by DNA sequences. (a) is the visualisation results based on the Minkowski metric and (b) is the zoomed in visualisation result.

Visualisation results coloured based on binding status

By studying the coloured projection results based on either the properties of the amino acids combination or the information of the DNA sequence, it is proved that the NeuroScale model is able to represent the relevant structural information of the data samples from the high dimensional data space to the low dimensional feature space. The binding status as one of the key features of the interaction also needs to be studied. Figures 4.18(a) and 4.18(b) plot the representative results coloured by the binding statuses (binding/non-binding) respectively. Through comparing the distributions of the binding and non-binding data samples in the visualisation results based on the different dissimilarity metrics, the representation corresponding to the indefinite Minkowski metric shows a better separation of binding versus non-binding experiments. This discovery implies that the deployment of non positive definite metrics has some benefit in this situation. The choice of the Minkowski metric has no fundamental biological motivation, but neither does the assumption of a Euclidean or other positive definite metric defining the dissimilarity space. Therefore, searching the most appropriate metric on the dissimilarity description can be a direction for future research.



(a) The visualisation result based on Euclidean distance. In this figure, except one cluster on the right mainly including non-binding data, most of samples with either binding or non-binding status are distributed in all clusters.



(b) The visualisation result based on Minkowski distance. In the figure, the data samples with non-binding properties are mainly projected on the top of clusters with the crescent shape and bottom right of the figure, and binding data samples are represented mainly gathering in the bottom left and centre of the figure.

Figure 4.18: The NeuroScale projection results of the interaction colouring by binding statuses. (a) is the visualisation result based on classical Euclidean dissimilarity and (b) is the Minkowski metric based projection result. Comparing them, the representation corresponding to the indefinite Minkowski metric shows a better visual separation of binding versus non-binding experiments.

4.3.7 Generated synthetic data visualisation

In Subsection 4.3.5, the projection results of the DNA-binding Cys_2His_2 zinc finger interactions using the NeuroScale model was discussed. Through analysing the results, it is notable that there are some distinctive 'gaps' between the projected data. Since the total number of possible protein-DNA binding sites, for zinc fingers alone, is almost 41 million, and the current training dataset only includes 1860 data samples, these gaps indicate that certain combinations do not naturally occur. To prove this surmise, groups of synthetic data samples are randomly selected from the synthetic database(Database DB5 which was defined in Subsection 3.2.2) and visualised with the data samples from the database DB1. Figures 4.19 and 4.20 show representative results for both the classical Euclidean and Minkowski metrics. Restricted by the computing speed of the server, in Figure 4.19, a maximum of, 12,000 synthetic data are selected and visualised with the 1860 experimental data samples. In the figure, almost all gaps appearing in the visualisation results which were discussed in the previous subsection are filled by the synthetic data samples. A similar result is obtained in Figure 4.20. Since only 8,000 synthetic data samples were used to train the NeuroScale model, the gap between the two major clusters still exists, but is narrowed. On the other hand, it is found that the experimental data samples are projected into certain areas of the visualisation space. This phenomenon illustrates that the naturally occurring interactions only distribute in a specific area of the high dimensional structural space.

As explained in Subsection 3.2.2, a 1×320 vector is used to represent the structural information of the DNA-binding zinc finger proteins. The order of four sections for representing the binding pairs at four positions are randomly determined. For example, the first section, i.e., from index 1 to 80, indicates the binding of one amino acid with one nucleotide on the complementary strand at position 2. The second section (81 to 160) for position -1, the third (161 to 240) for position 3 and the fourth (241 to 320) for position 6. So the specific coding scheme used does not determine the structure in the visualisation space. Additional experiments have been carried out to verify the the order of the four sections has no effect on the visualisation result.

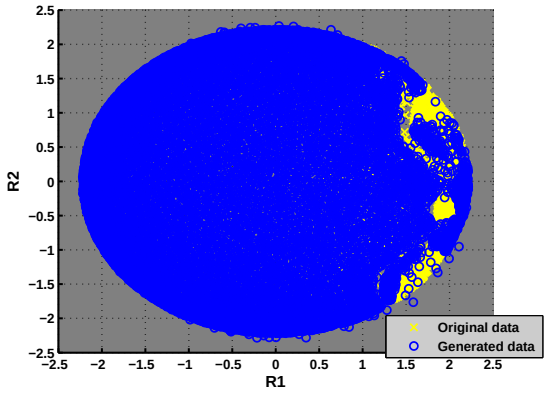


Figure 4.19: The visualisation result using synthetic data based on Euclidean distance. In the figure, the gaps discussed in Subsection 4.3.5 are now filled by the synthetic data, and the original dataset is mainly projected on the right hand side of the figure.

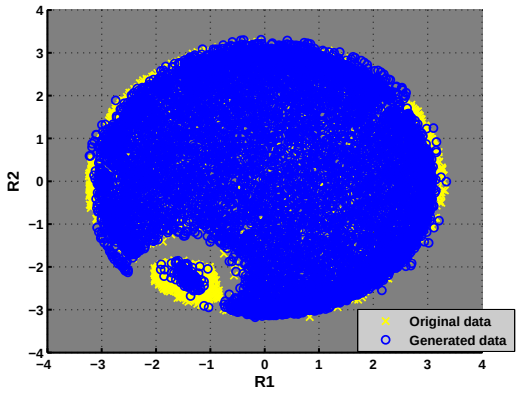


Figure 4.20: The visualisation result using synthetic data based on Minkowski distance. Similar to Figure 4.19, the gap between the two clusters is narrowed but not filled. It is considered due to the limited number of applied synthetic data.

4.4 Summary

This chapter has focussed on the mathematics of a specific topographic low-dimensional representation approach in which the input space metric need not be positive definite. Through studying and comparing various candidate non-linear dimensionality reduction methods, the NeuroScale model was exploited to represent, interpolate and project the high dimensional data under such circumstances. It was revealed that this approach can elucidate interesting structure in very high dimensional and large data problems. In the specific case here of the Cys_2His_2 zinc finger interactions with DNA it was discovered

that the dissimilarity representation using Minkowski metrics gives structural groupings of binding/non-binding examples which cluster in biologically-interesting ways. It is worth emphasising that the binding/non-binding knowledge was not used in any part of the modelling other than to label the final figures, and so this binding attribute was genuinely 'discovered' by the process. An interesting feature is the existence of 'forbidden bands' in the low-dimensional representation which are likely reflect evolutionary preferences for certain types of zinc finger-DNA complexes. This property has been investigated through visualising the created synthetic dataset. So, if the topographic visualisation space is reflecting functional properties of the DNA- protein interactions, perhaps adaptive classifiers could be constructed using the structural coding of the input data, or its projection visualisation, to predict possible binding affinity.

In order to evaluate the potential of the representation results on predicting the functional properties of the given data samples, various classifiers will be applied in Chapter 5. Besides the 2-D visualisation results, the 320-D original dataset and the PCA based reconstruction dataset will also be employed with selected prediction models. The classification results will be assessed by various quality criteria.

5 Analysis methods (II): Data prediction

CONTENTS

In the previous chapter, the geometric structure of DNA-binding Cys_2His_2 zinc finger interaction in high dimensions has been represented by NeuroScale in low dimensions based on different dissimilarity measures. Through analysing the visualisation results, several prediction approaches have been employed. This chapter will be focused on investigating various prediction models, and comparing the accuracy of each model based on Receiver Operator Characteristic (ROC). The chapter begins with the characteristics of the database which includes both high dimensional data and represented data in the low dimensional feature space. Then, the quality criteria, such as ROC curve, are introduced in the second part of Section 5.1, followed by the prediction algorithms and relevant results in Section 5.2.

5.1 Experimental Methodology

The created database DB1 (as defined in Table 3.3) forms the foundation of predicting the DNA-binding Cys_2His_2 zinc finger interaction. Moreover, as the visualisation results, illustrated in Chapter 4 cluster the data samples in biologically-interesting ways, it is worth utilising the projection data as another database to investigate various prediction models. In this section, the creation of data sets, such as training data set, test data set and validation data set, based on different databases is introduced firstly. Then, the characteristics of each database, especially those of the low dimension reconstructed databases, are discussed in Subsection 5.1.2 and 5.1.3. Finally, the quality criteria which are used to evaluate the accuracy of each prediction approach are explained.

5.1.1 Characteristics of re–organized database

In order to apply prediction techniques, the database DB1 that was created based on the published papers, is further sorted into three categories based on Monte Carlo methods: training dataset, test dataset and validation dataset. The training dataset takes 50% of the total data available, while the test dataset includes 40% and the validation dataset 10% [1].

[1] In this work, the training dataset was used to train the prediction models. Then, the test dataset was applied to determine relevant parameters such as hidden centres of neural networks and neighours of *k*-nearest neighbours model. Finally, the validation dataset is used to evaluate the performance of each prediction ap-

Moreover, in order to ensure the consistency and comprehensiveness of predictions, the 1860 binary data samples in DB1 consisting of 882 binding and 978 non-binding examples are randomly reconstituted one hundred times according to the proportion defined above. It has been ensured that, there is no overlapping between the training, test and validation data in each reconstitution. The one hundred data subsets then form a 320 dimensional (320-D) database. This new database is called '320-D original database' in this work. It will be used as a database for prediction, and to reconstruct three more databases[2] for prediction models investigation. The details of creating the 320-D original database is described in Appendix D.1.

In the 320-D original database, there are 100 groups of datasets. Each dataset is stipulated to include 933 training data samples, 737 test data samples and 190 validation data samples. In each category, the data samples are selected from the 25 data sources listed in Appendix A Table A.3 and A.4. Besides being used in the 320-D binary original database, the data samples can also be represented using the subsequent processing methods, such as NeuroScale and PCA. The binding affinity as target variable is defined as [0 1] for binding, [1 0] for non-binding in building the prediction model.

5.1.2 Two dimensional (2-D) reconstruction database

The 2-D reconstruction database in the work is composed of two groups of projection results which are obtained through applying NeuroScale based on different dissimilarity measures: Euclidean distance and Minkowski distance. As discussed in Chapter 4, the representations corresponding to the two metrics can cluster the data samples in biologically-interesting ways. Moreover, although the knowledge of binding/non-binding is not used in any part of the visualisation modelling, the dissimilarity representation using the Minkowski metric shows a better separation of binding versus non-binding samples. The observation motivates us to now analyse the accuracy of using predictive models based, not on the input patterns, but on the projected two-dimensional data as generated

proaches.

[2]The three databases include a 2-D Euclidean distance based reconstruction database, a 2-D Minkowski distance based reconstruction database and a 320-D PCA based reconstruction database.

by the NeuroScale visualisation map.

5.1.3 320-D database

Alongside the 320-D original database and the 2-D reconstruction databases aforementioned, a 320-D reconstruction database created based on PCA is also used in the prediction model training.

Although the visualisation result of the PCA model discussed in Subsection 4.2.2 is unsatisfactory, the varying trend of eigenvalues which is shown in Figure 5.1 still arouses the interest in studying the accuracy of prediction, by using the reconstruction database based on the PCA. According to Table B.1 in Appendix B.1.1, when the number of eigenvectors reaches 232, 100% variances can be represented by the model. In order to verify this finding, extra visualisation experiments have been carried out, based on NeuroScale using the data samples that are reconstructed by different number of eigenvectors. The results can be found in Appendix C.2, which confirm that the data reconstructed by the first 232 eigenvectors from PCA can describe the characteristics of the interaction as the 320-D original data. Therefore, besides the 320-D original database, another 320-D database is created to contain these reconstructed data. The same as the 2-D databases, the 320-D reconstructed vectors are represented in the form of real numbers as shown in Table 5.1.

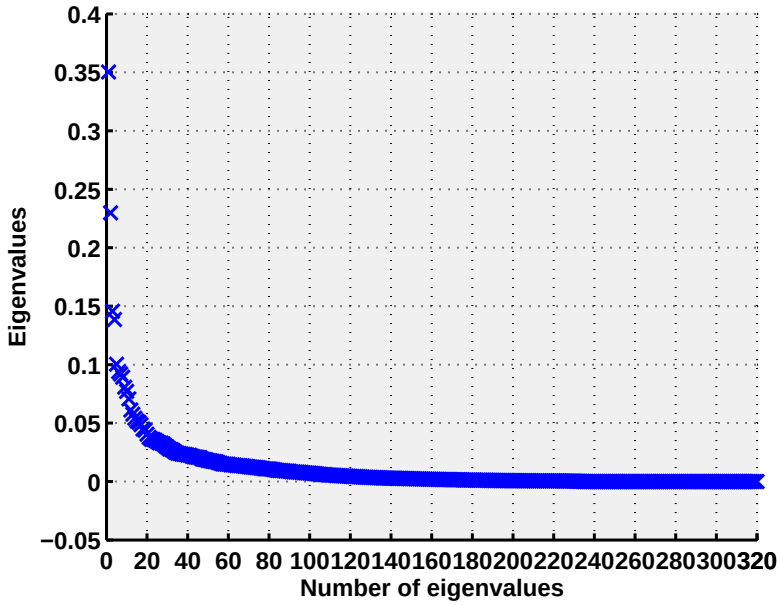


Figure 5.1: The varying trend of eigenvalues of relative component eigenvectors. As the original data is represented by 320 dimensional vector, the maximum number of component eigenvectors is 320. By plotting the eigenvalues in descending order, the number of eigenvectors which have relative important contributions for representing the original data in the feature space can be determined. Specific to this work, the first 232 eigenvectors are selected to reconstruct the database.

	1	2	3	···	319	320
1	-1.8909$\times10^{-16}$	0	2.9490$\times10^{-17}$	···	1.4452$\times10^{-16}$	0
2	7.2858$\times10^{-17}$	0	1.0443$\times10^{-15}$	···	2.4709$\times10^{-16}$	0
3	-1.4051$\times10^{-16}$	0	5.6032$\times10^{-16}$	···	-4.5439$\times10^{-16}$	0
4	8.5869$\times10^{-17}$	0	-1.9776$\times10^{-16}$	···	-2.2595$\times10^{-16}$	0

Table 5.1: Examples of 320-D reconstruction data. In this table, four 320-D reconstruction data samples are provided. Different from the original data, the reconstructed data samples are represented by continuous numbers instead of the binary format. This change may affect the performance of prediction models.

5.1.4 Quality criteria

Receiver Operating Characteristic (ROC) analysis is an evaluation technique applied in signal detection theory (Swets, 1988). In recent years, in the machine learning community, the ROC curve has been exploited to depict relative trade offs between benefits (true positive) and costs (false positive) (Fawcett, 2006). In this work, for the four different

reconstituted databases described above, the ROC graphs and relevant parameters, such as true positive rate (TPR), false positive rate (FPR) and accuracy (ACC), computed as in Appendix C.3 are used for comparing the selected prediction methods. It is generated by varying a threshold across the defined output range of a scoring model which is explained in Appendix C.3 Figure C.11. Since the ROC graph has an attractive property that it is insensitive to changes in class distribution, the changes between the one hundred subsets in each database would not affect the performance evaluation of the models. In this work, the ROC curve is computed at each subset and these curves are averaged for every selected methods by computing an average number of predicted true positives at every false positive rate. Besides comparing the two-dimensional curves of the prediction models in the same graph, an area under the ROC curve, which is abbreviated as AUC (Bradley, 1997; Hanley and McNeil, 1982), is also calculated to represent the performances of methods as a single scalar value. In general, an area of 1 represents a perfect performance of the classifier. When the area equals to 0.5, the prediction model is considered to be worthless as the classification is arbitrary. The ROC curves and relevant AUC will be provided and discussed in Section 5.2.

5.2 Prediction algorithms and results

The focus of this Section is on discussing the various prediction approaches employed in this work, and evaluating the performances of these models through plotting the ROC graphs of the classifiers and comparing their area under the ROC curve (AUC).

5.2.1 Prediction algorithms

Since the vector representation of the DNA-binding Cys_2His_2 zinc finger interaction based on the canonical binding model has a natural biochemical interpretation corresponding to the potential contacts between the bases and the amino acids, the goal of prediction is to deduce the possibilities of amino acid-nucleotide interactions in the four canonical contacts. In this work, six prediction methods: linear regression, *k*-nearest neighbours (*k*-NN), multi-layer perceptron (MLP), radial basis function (RBF), support

vector machine (SVM) and relevance vector machine (RVM) are applied to predict the binding label of the data. The principles of these models can be found in (Bishop, 2007). In the following paragraphs, specific choices of the various parameters of the models will be explained.

Linear regression

Linear regression is an approach of modelling the relationship between a scalar variable y and one or more input variables $\mathbf{x}$ (Bishop, 2007). In linear regression, data are modelled using linear functions, and unknown model parameters are estimated from the data. In this work, although polynomial functions which are applied to find the best expression between each parameter and the target binding statuses are non-linear, the parameters in the linear regression model are still determined linearly. The reason for choosing linear regression as one of the prediction method is to use it as a reference for other methods.

***k*-nearest neighbours (*k*-NN)**

k-NN is a non-parametric method for classifying objects based on closest training examples in the feature space. In this work, the nearest neighbour selection is implemented by finding out the smallest Euclidean distance between the target data samples and the surrounding reference samples (i.e., training data samples). Through changing the number of neighbours[3] and verifying the binding statuses of the selected nearest reference samples, the binding affinity of the target data can be determined. By evaluating the classification accuracy of the test dataset, the number of neighbours is able to be determined and applied to predict the binding affinities of the validation dataset.

Multi-layer perceptron (MLP)

Multi-layer perceptron (MLP) neural network is a non linear regression model originally inspired by the structure and functional aspects of biological neural networks (Rumelhart et al., 1986). In this work, it provides an optimised non linear mapping function that maps the input feature vector $\mathbf{x}$ to an output that represents the binding affinity. When building the MLP model, the training dataset is selected as the input data and the relevant

[3] In this work, the number of neighbours was changed from 1 to 11 with the interval of 2.

binding status is the target output. The network is constructed with a logistic output function which is suitable for a two-classes problem). To optimise the weights of the output function, back-propagation as a general technique for evaluating derivatives of the activation functions is applied. In this work, the main function of the MLP network was called from Natlab library. A Scaled Conjugate Gradient (SCG) algorithm is utilised to train the weights based on the training dataset by changing the number of hidden centres from 3 to 150 with the interval of 3. Due to the high dimensionality of the input data, a big number of parameters are generated which makes the model overfits the training data. To avoid overfitting, some additional techniques are necessary, such as early stopping, cross-validation, regularization etc.. Specific to this work, the test dataset is use to implement the cross-validation by comparing the normalised error which defined later in Equation 5.1. The details will be discussed in subsequent sections with prediction results.

Radial basis function (RBF)

The radial basis function (RBF) network is a non-linear functional interpolation model where the parameters of interest multiplying non linear basis functions can be determined using linear techniques (Webb, 1999). Similarly to the MLP, the RBF provides a transformation of the training dataset to a 2-D output space according to a function $\mathbf{y} = f(\mathbf{x}, \mathbf{W})$, where $\mathbf{W}$ is the weight matrix of the output layer. Different from the MLP, the activation of the hidden centres in the RBF network is given by a non-linear function which calculates the distance between the input vector and a weight vector. Specific to this work, the Matlab code based on the Netlab library selected a thin plate spline (TPS) as the basis function due to its advantage in fitting a surface through a set of points and using a roughness penalty (Meinguet, 1979). The number of hidden centres varies from 2 to 150 with the interval of 2, and is finally determined by cross-validating the normalised error with the test dataset. The values of centres and weights are initially randomly selected from the training dataset, and are optimised by the EM algorithm.

Support vector machines (SVM)

In machine learning, the support vector machine (SVM) is a learning system (Cortes

and Vapnik, 1995) that uses associated learning algorithms to analyse training data and produce an inferred function, which can be used for classification and regression analysis (Cortes and Vapnik, 1995; Cristianini and Shawe-Taylor, 2000). Given a set of training examples, each marked as belonging to one of two categories, a SVM creates a linear classifier between the two classes in a high-dimensional space of the training dataset feature vectors. A weight vector w is found by a constrained optimisation process such that a hyperplane is defined which separates positive from negative examples. In order to achieve a good separation, the selected hyperplane ought to have the largest distance from it to the nearest training data or any class. For a non-linear data space, a kernel function is needed to project the input data into a high dimensional linear feature space for the hyperplane construction. In this work, LIBSVM version 3.14 (Chang and Lin, 2011) is used to train the SVMs. Moreover, both support vector classification and regression have been used, and RBF is selected as the kernel function.

Relevance vector machine (RVM)

The relevance vector machine (RVM) is a Bayesian sparse kernel technique that uses Bayesian inference to obtain parsimonious solutions for regression and classification (Tipping, 2001; Bishop, 2007). The RVM has an identical functional form to the SVM, but provides probabilistic classification. In this work, SparseBayes version 1.1 (Tipping, 2001) is used to train the RVM. Given the training dataset as the input data examples, the relevant binding status is the target output. The kernel function is defined as Gaussian, where its default relative noise is 0.1 and the kernel length scale is adjusted according to different input datasets.

Through reviewing the employed prediction models, the specific settings of each method have been clarified. In order to keep the consistency of building prediction models, the setting is not changed when different databases are used as the input data examples. Moreover, in the classification process, the reference label which has either binding (defined as [0 1] or [1] for SVM) or non-binding (defined as [1 0] or [0] for SVM) status is used as the target output. As introduced in Subsection 3.1.1, the dissociation constant K_d which is a specific type of equilibrium constant that measures the propensity of a larger

object to separate reversibly into smaller components, is used to describe the binding preference in part of the data samples. The threshold of K_d here is set to be 200 nm. If $K_d <$ 200 nm, the status is defined as binding, and expressed as [0 1], otherwise, the status is defined as non-binding which is [1 0].

In order to obtain the most accurate results from various prediction models, adjusting parameters of the models, such as *k*-NN, MLP and RBF becomes necessary. Therefore, a normalised classification error is defined as a preliminary criterion to evaluate the performance of each model. Given a dataset of N data samples, the reference target output and the prediction result are represented by y_{target} and $y_{predict}$ respectively. The normalised error is then defined as:

$$E = \frac{1}{N}\sum_{i=1}^{N}\frac{\|y_{predict_i} - y_{target_i}\|}{\|y_{target_i} - \overline{y}\|} \tag{5.1}$$

where $\overline{y}$ is the average of the target output. The more accurate the prediction result is, the smaller the error is. In subsequent subsections, the error of each model based on different databases will be compared and discussed.

5.2.2 Prediction results based on 2-D reconstruction data

In Subsection 5.1.2, the characteristics of two 2-D reconstruction databases were introduced. The incentive of creating these two databases based on different dissimilarity metrics is that in the visualisation results the data samples are clustered in the biologically-interesting ways. Can the 2-D reconstruction databases, especially the Minkowski based database, provide more advantages than the 320-D original database for a prediction model? Will the *k*-NN model have the best performance by using the low dimensional database? In the following paragraphs the results of various prediction models based on two databases will be presented.

Prediction results based on Euclidean metric

When NeuroScale was applied to project the 320-D data samples in the 2-D Euclidean space, a classic Euclidean dissimilarity metric was used in the input space. By reconsti-

tuting the projection data samples, a 2-D database based on a Euclidean metric is reconstructed and is employed to investigate various prediction methods.

Figure 5.2 is a plot of the normalised error of the *k*-NN model computed on test and validation datasets. For the 2-D Euclidean metric based reconstruction database, the number of nearest neighbours is adjusted from 1 to 11 with the interval of 2. According to the normalised error, when 5 neighbours are used to define the target test data samples, the normalised error which is averaged by 100 groups reaches the lowest point: 0.3338. With the same number of the nearest neighbours for verifying the validation dataset, the obtained error is 0.3369.

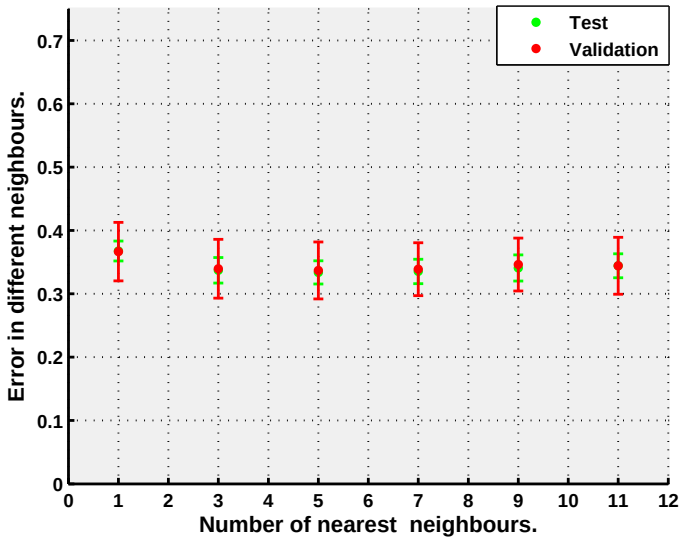


Figure 5.2: *k*-NN normalised classification error for the 2-D reconstruction datasets based on Euclidean distance. When the number of neighbours is 5, the normalised classification errors for the test and validation datasets reach the global minimum at 0.3338 and 0.3369, respectively.

Figure 5.3 is the graph of the normalised classification error of the MLP based on 2-D Euclidean metric based reconstruction database. In this graph, the number of hidden centres is changed from 3 to 150 with the interval of 3. The highest errors of the three data subsets occur at the beginning where the number of hidden centres is smallest; then, the errors diminish quickly when the number increases and remain at a low level. With the errors of both training and test datasets taken into account, 54 hidden centres are selected as the relatively lowest points to implement the prediction of validation data samples. The normalised classification error of validation dataset with 54 hidden centres is 0.4288. The

relevant ROC curve and AUC will be discussed later.

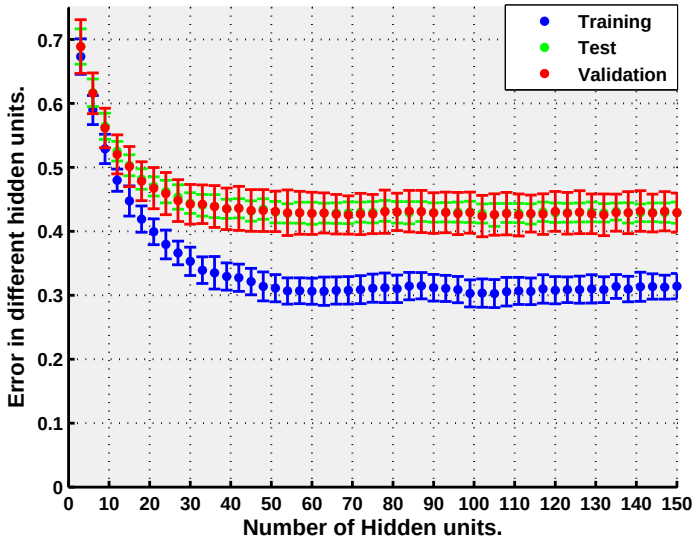


Figure 5.3: The MLP normalised classification error for the 2-D reconstruction data based on the Euclidean distance, with respect to the number of hidden centres. The normalised errors for training dataset are generally better than those of the test and validation datasets. When the hidden centres set to be 54, the error of training dataset is 0.3066, the error of test dataset is 0.4280, and the error of validation dataset is 0.4288.

Figure 5.4 shows the normalised error of the RBF model based on using a 2-D Euclidean metric reconstruction database. The number of centres is adjusted from 2 to 150 with the interval of 2. Compared with the error from the MLP algorithm, the error of the RBF is much higher and the error of the training dataset remains at a high level, decreasing slowly. When there are more than 100 hidden centres, the errors of test and validation datasets rebound, caused by over-training of the model. Using the errors of both training and test data as a reference, the number of the hidden centres is set at 80 where the error of the test data reaches the lowest value and that of the training data is also at a low level. The normalised classification error of validation dataset with 80 hidden centres is 0.5883.

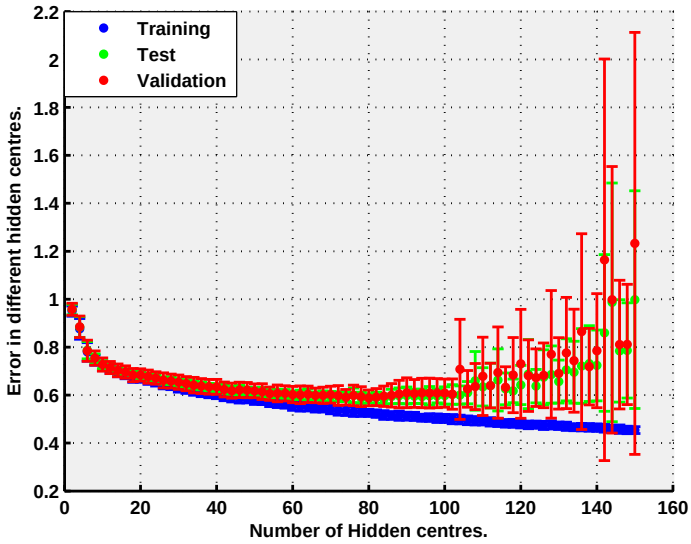


Figure 5.4: The RBF normalised classification error for the 2-D reconstruction data based on the Euclidean distance. When the number of hidden centres is smaller than 100, the normalised errors differ little between three datasets. Hereafter, due to over-training, the difference becomes rather obvious. Also, the error bars for the test and validation datasets enlarge quickly. When the hidden centres is 80, the error of test dataset has the lowest value: 0.5866, while the error of training dataset is 0.5258, and the error of validation dataset is 0.5883.

Table 5.2 is a summary of the normalised classification error of all applied prediction models. With this evaluation criteria alone, the *k*-NN model shows the best prediction performance for both test and validation data subsets followed by RVM, while linear regression shows the worst.

	Linear Reg.	**KNN**	**MLP**	**RBF**	**SVM (Clas.)**	**SVM (Reg.)**	**RVM**
Training	0.9202	——	0.3066	0.5258	——	0.6767	0.3349
Test	0.9349	**0.3338**	0.4280	0.5866	0.5463	0.7317	0.3720
Validation	0.9693	**0.3369**	0.4288	0.5883	0.5949	0.7216	0.3726

Table 5.2: The normalised classification error of the 2-D reconstruction data based on the Euclidean distance. From this table, the *k*-NN has the lowest normalised classification error for both test and validation data subsets: 0.3338 and 0.3369.

Besides the normalised classification error, the accuracy[4] of each prediction method is calculated according to the definition explained in Appendix C.3. Table 5.3 lists the accuracy of all prediction models for different data subsets. The performance of the *k*-NN is still the best one, where the accuracy of the test dataset is 0.8331 and that of the

[4]The accuracy as defined in Appendix C.3, is calculated as $accuracy = \frac{TP+TN}{P+N}$

validation dataset is 0.8323. As highlighted in this Table, the prediction accuracy of the SVM regression model is also as good as *k*-NN (test dataset: 0.8312, validation dataset: 0.8327), but the normalised error of this method is unsatisfactory (test dataset: 0.7317, validation dataset:0.7216). This is because of the characteristics of the output and because the normalised error is calculated between the prediction results and the target output. If the value of a 2-D prediction output is geometrically far from [0 1] or [1 0] (all target outputs are represented in this format), such as [0.5967 0.4030], the normalised error would be very high. Despite of this, the data example can still be classified into the correct group using the defined threshold and through considering the Euclidean distance between the target output and the prediction result. Therefore, in this work, the normalised error is not suitable to evaluate the performance of the SVM regression model. More related information on accuracy can be found in Appendix C.4 Table C.2.

	Linear Reg.	**KNN**	**MLP**	**RBF**	**SVM (Clas.)**	**SVM (Reg.)**	**RVM**
Training	0.6201	——	0.8984	0.8384	——	**0.9090**	0.8331
Test	0.6191	**0.8331**	0.8260	0.8075	0.7668	**0.8312**	0.8152
Validation	0.6118	**0.8323**	0.8249	0.8067	0.7681	**0.8327**	0.8143

Table 5.3: The accuracy of the 2-D reconstruction data based on the Euclidean distance. In this table, the *k*-NN and SVM regression model have the best prediction performance specific for the 2-D Euclidean distance based reconstruction database. The accuracy of test and validation datasets of the *k*-NN is 0.8331 and 0.8323 respectively, meanwhile for the SVM regression, the accuracy is 0.8312 and 0.8327.

The ROC curve is another quality criterion that can show the performance of the prediction models. Figure 5.5 shows the ROC curves of the training, test and validation datasets for the MLP, RBF, SVM regression and RVM methods, respectively. In these figures, the ROC curves of training dataset always have the best performance, the differences between the test and validation datasets are very small, which verifies the results of accuracy in Table 5.3. Figure 5.6 shows the ROC curves for the cross-validation analysis. For three different datasets, the MLP, SVM regression and RVM models outperform the RBF. Though the SVM regression holds the top true positive rates at same low false positive rate on all datasets, the areas under ROC curves (AUC) of MLP found in Table 5.4 are nonetheless better than that of the SVM regression and RVM models, with the RBF

trailing behind them.

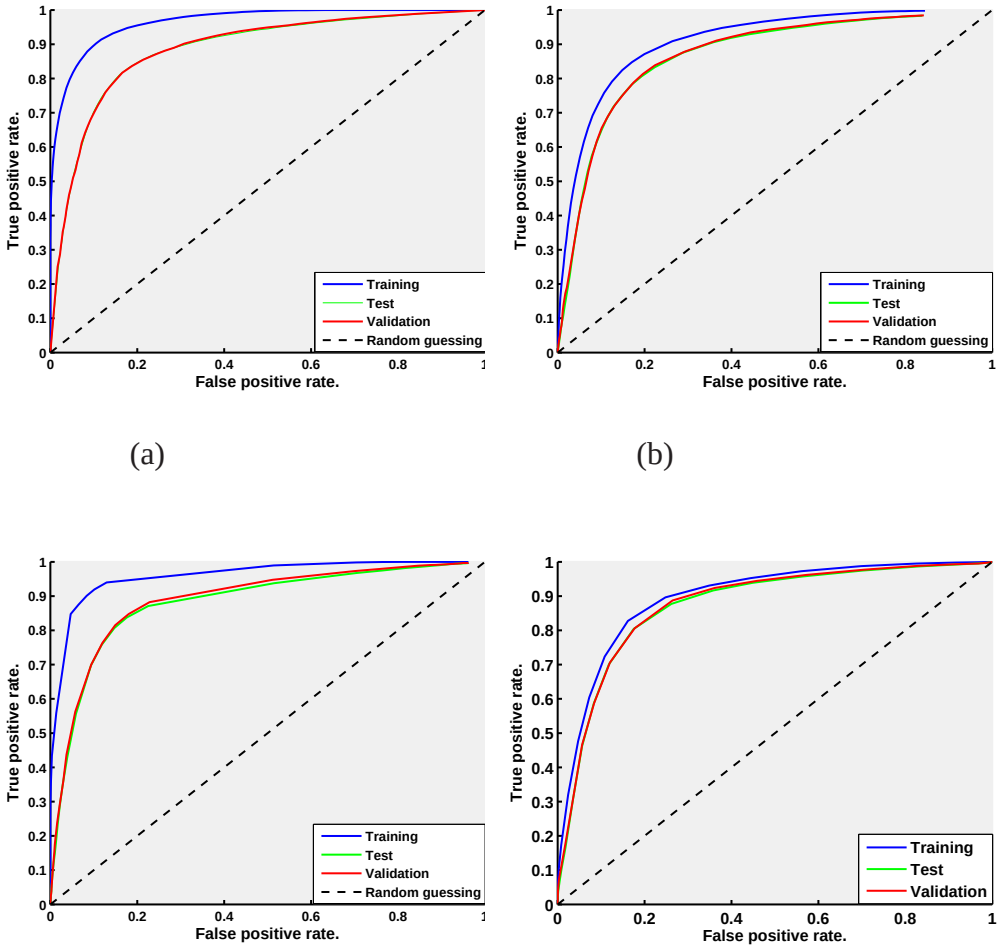
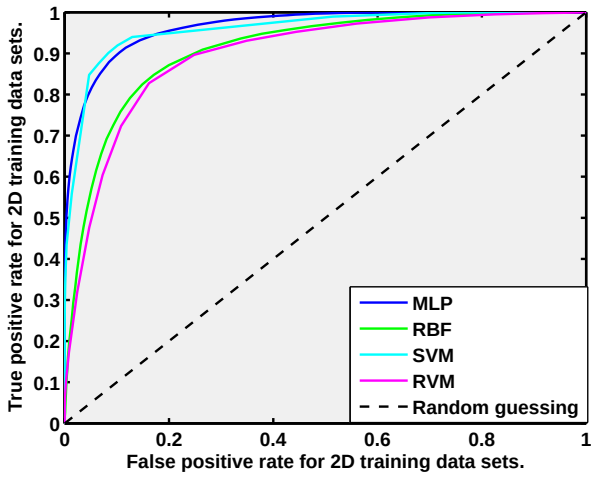
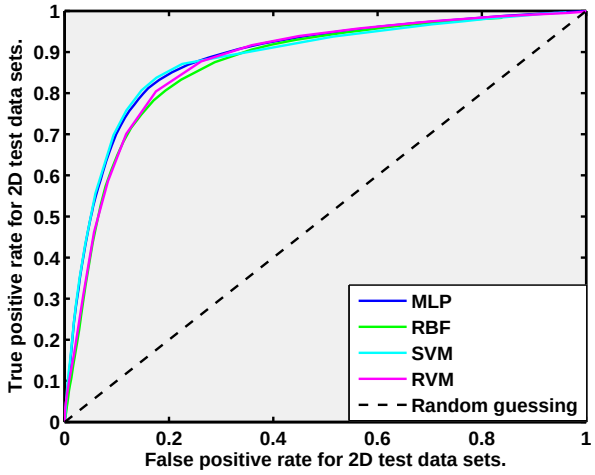


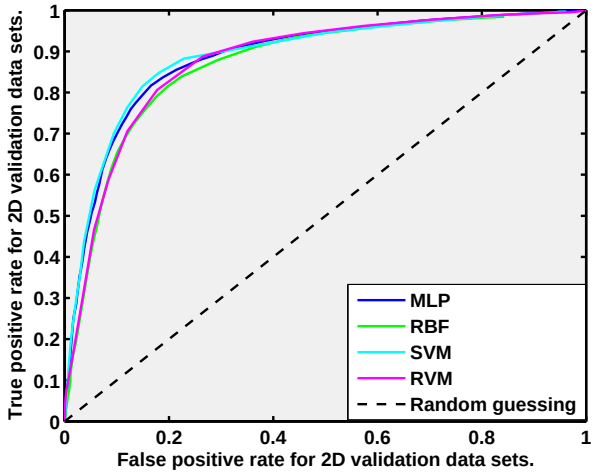
Figure 5.5: The ROC curves of different classifiers using the 2-D reconstruction datasets based on the Euclidean distance. (a) MLP classifier (AUC values: 0.9652, 0.8844 and 0.8862.); (b) RBF classifier (AUC values:0.7543, 0.7081 and 0.7134.) (c) SVM regression classifier (AUC values:0.9221, 0.8433 and 0.8509.) and (d) RVM classifier (AUC values:0.8963, 0.8734 and 0.8764.). Generally, the classifiers performs much better than random guessing (AUC: 0.5).



(a) The ROC curves of 2-D training dataset.



(b) The ROC curves of 2-D testing dataset.



(c) The ROC curves of 2-D validation dataset.

Figure 5.6: The ROC curves for the cross-validation analysis using 2-D reconstruction database on the Euclidean distance. (a) Training dataset; (b) Test dataset and (c) Validation dataset.

Subset	MLP	RBF	SVM Regression	RVM
Training	**0.9652**	0.7543	0.9221	0.8963
Test	**0.8844**	0.7081	0.8433	0.8734
Validation	**0.8862**	0.7134	0.8509	0.8764

Table 5.4: AUC values for cross validation testing on training, test and validation subsets based on Euclidean distance. The AUC values of three datasets show that the overall performance of the MLP is slight better than the SVM regression. The performance of RBF is the worst one comparing with the other two.

Prediction results based on Minkowski

Different from the Euclidean metric based database, the database that is reconstructed by NeuroScale uses the Minkowski indefinite inner product where the dimensions of the input space corresponding to the connections to the complementary DNA strand are weighted with -1 and the connections related to the primary DNA helix are weighted +1. In this part, the performances of various classifiers by using the 2-D Minkowski metric based reconstruction database are compared with the Euclidean metric based results.

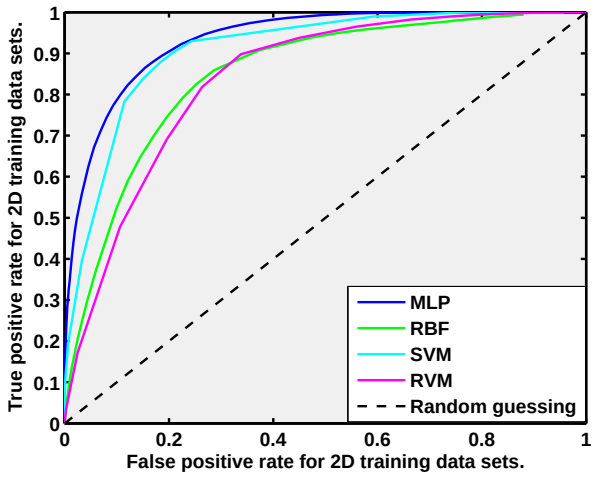
According to the normalised classification error of the *k*-NN, MLP and RBF models which are shown in Appendix C.4.1, the normalised classification errors of all classifiers using the Minkowski metric based reconstruction database listed in Table 5.5 are generally worse by comparing with Table 5.2. Although the visualisation results of NeuroScale based on the Minkowski metric shows a better separation of binding versus non-binding samples, the normalised error of the *k*-NN is inferior to that for the Euclidean metric based reconstruction database. In contrast to the *k*-NN, the performance of the RVM model is stable, as the normalised error for the test dataset is 0.3720 and the validation dataset is 0.3726. The normalised error of linear regression is still the highest one.

	Linear Reg.	KNN	MLP	RBF	SVM (Clas.)	SVM (Reg.)	RVM
Training	0.9155	——	0.4203	0.6642	——	0.6767	**0.3349**
Test	0.9158	0.4351	0.5557	0.7136	0.5826	0.7317	**0.3720**
Validation	0.9127	0.4384	0.5559	0.7133	0.5499	0.7216	**0.3726**

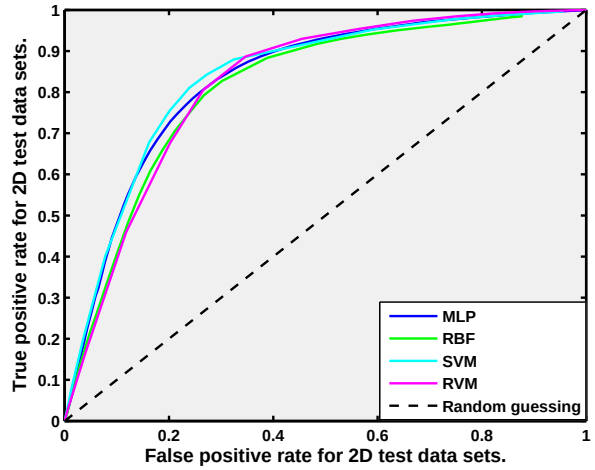
Table 5.5: The normalised classification error for the 2-D reconstruction data based on the Minkowski distance. From this table, the RVM has the lowest normalised classification error for both test and validation data subsets: 0.3720 and 0.3726.

The accuracy of all classifiers is presented in Table 5.6 where the relevant ROC curves are included in Appendix C.4.1 Figure C.15. Comparing with the results in Table 5.3, the performance of the *k*-NN becomes worse this time. The SVM regression outperforms other methods, although the normalised error of this model is much higher than that of the RVM as shown in Table 5.5. The reason for this phenomenon should be same as the discussion in the previous part: the normalised error is calculated by the differences between the prediction results and the target output; but the accuracy is calculated depending on the classification results which are obtained by comparing the defined threshold and the Euclidean distance between the target output and the prediction results.

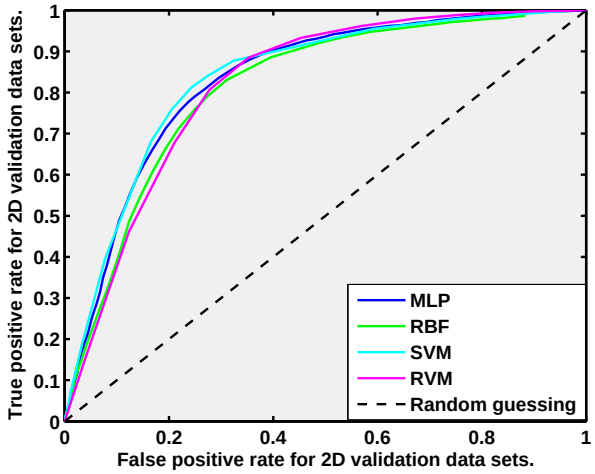
Figure 5.7 shows the ROC curves for the cross-validation analysis. For the training dataset, the MLP gives the best result with an AUC of 0.9346. Although the SVM regression produces the top true positive rate at the same lower false positive rate for the test and validation datasets, the MLP classifier still obtains the best AUC value as presented in Table 5.7.



(a) The ROC curves for the 2-D training dataset.



(b) The ROC curves for the 2-D testing dataset.



(c) The ROC curves for the 2-D validation dataset.

Figure 5.7: The ROC curves for the cross-validation analysis using 2-D reconstruction database on the Minkowski distance. (a) Training dataset; (b) Test dataset and (c) Validation dataset.

	Linear Reg.	KNN	MLP	RBF	SVM (Clas.)	SVM (Reg.)	RVM
Training	0.6442	——	0.8549	0.7844	——	**0.8458**	0.7753
Test	0.6461	0.7178	0.7680	0.7606	0.7739	**0.7840**	0.7674
Validation	0.6424	0.7287	0.7698	0.7586	0.7736	**0.7846**	0.7645

Table 5.6: The accuracy for the 2-D reconstruction data based on the Minkowski distance. In this table, the SVM regression model has the best prediction performance specific for the 2-D Minkowski distance based reconstruction database. The accuracy of test and validation datasets of the SVM regression is 0.7840 and 0.7836 respectively.

Subset	MLP	RBF	SVM Regression	RVM
Training	**0.9346**	0.7301	0.8680	0.8390
Test	**0.8310**	0.6921	0.7934	0.8212
Validation	**0.8315**	0.6965	0.7970	0.8184

Table 5.7: AUC values for cross validation testing on training, test and validation subsets based on Minkowski distance. The AUC values of three datasets show that the MLP classifier outperforms other two methods.

Discussion

In this Subsection, the prediction methods are evaluated by the 2-D reconstruction databases. Through comparing the normalised classification error and cross-validating the ROC curves and the area under ROC curves, the overall performances of the SVM regression and MLP are better than other classifiers for the 2-D databases. Meanwhile, although the visualisation results of NeuroScale indicate a better perceptual separation of binding versus non-binding samples when the database is Minkowski metric based, the Euclidean metric based database demonstrates better prediction results. In particular, the *k*-NN has the best performance in such a database.

5.2.3 Prediction results based on 320-D data

Different from the 2-D reconstruction databases, the 320-D databases retain all information of the DNA-binding Cys_2His_2 zinc finger interaction. In this Subsection, the prediction results of various classifiers based on the 320-D databases will be analysed. In particular, it is interesting to reveal the effect of the 320-D PCA based reconstruction

database on the accuracy of these methods.

Prediction results based on 320-D original data

The 320-D original database is the only database which represents the characteristics of the experimental data samples without any post-processing. The focus of this part is on discussing the prediction results of various prediction methods using the 320-D original database. Furthermore, these results can be the reference for evaluating another 320-D database which is reconstructed based on PCA.

Similar to Subsection 5.2.2, the normalised classification error of all applied prediction models based on the 320-D original database are summarised in Table 5.8, where relevant figures can be found in Appendix C.4.2 Figure C.16, C.17 and C.18. As shown in the table, the overfitting occurred when the training dataset is applied to build the MLP model. By confirming the global minimum error of the test dataset, the number of hidden centres of the prediction model can be defined and the MLP model has the lowest normalised classification error for the validation dataset followed by RVM.

	Linear Reg.	**KNN**	**MLP**	**RBF**	**SVM (Clas.)**	**SVM (Reg.)**	**RVM**
Training	0.5584	——	7.7504×10^{-13}	0.4685	——	0.5540	0.0128
Test	0.6723	0.3425	**0.2104**	0.5597	0.3753	0.6073	0.2269
Validation	0.6625	0.3710	**0.2144**	0.5616	0.3327	0.6063	0.2257

Table 5.8: The normalised classification error for the 320-D original data. From this table, the MLP has the lowest normalised classification error for both test and validation data subsets: 0.2104 and 0.2144.

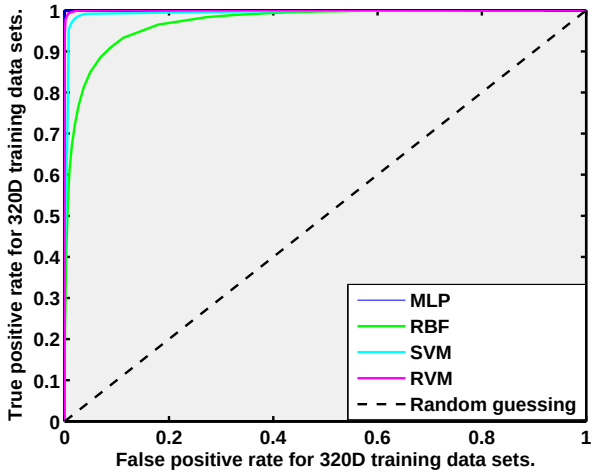
The accuracy of all prediction models are listed in Table 5.9. Although the MLP has the best normalised error result, the prediction accuracy of the SVM classification outperform the MLP and other methods (i.e. test dataset: 0.9132, validation dataset: 0.9168). The reason for the MLP having the lowest normalised error but without the best prediction accuracy is due to how the binding status of the data samples is defined. Particularly in this case, though the prediction results are very close to the target outputs, the final predicted binding status still depends on the selection of the threshold in the ROC curve. Thus, since the SVM classification model separates data samples into one of two classes directly, it is able to offer better prediction accuracy, despite that the false

classified samples may increase the normalised error.

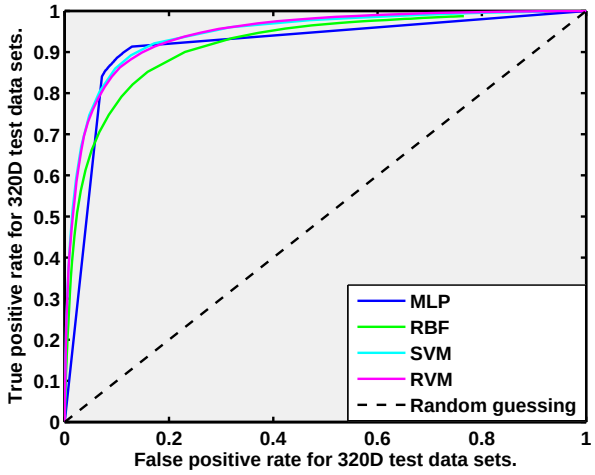
	Linear Reg.	KNN	MLP	RBF	SVM (Clas.)	SVM (Reg.)	RVM
Training	0.9295	——	**1**	0.9138	——	0.9802	0.9936
Test	0.8683	0.8221	0.8923	0.8482	**0.9132**	0.8828	0.8879
Validation	0.8656	0.8157	0.8896	0.8509	**0.9168**	0.8801	0.8876

Table 5.9: The accuracy for the 320-D original data. The SVM classification model has the best prediction performance followed by the MLP, RVM and SVM regression classifiers. The accuracy of test and validation datasets of the SVM classification is 0.9132 and 0.9168 respectively.

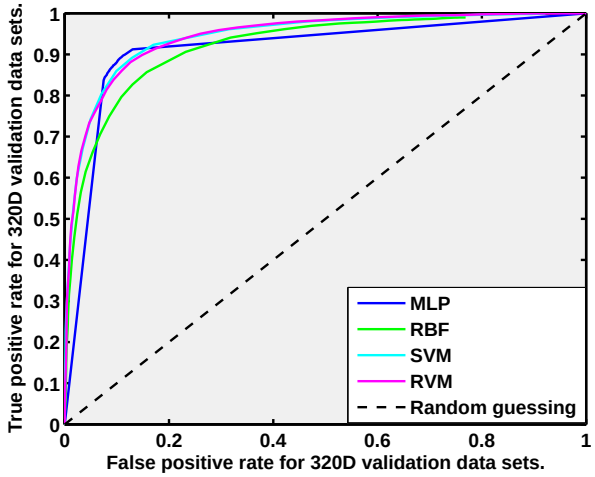
Figure 5.8 are the ROC curves of the selected models based on the 320-D original database. Similar to Figure 5.6 or Figure 5.7, Figure 5.8 shows the ROC curves for the cross-validation analysis. As seen in Figure 5.8(a), due to the over-training of the high dimensional input, the MLP model obtains the best AUC result of unity for the training dataset. Moreover, in Figure 5.8(b) and 5.8(c), the MLP curve also contains the top true positive rates at the same low false positive rates for the test and validation datasets. However, by comparing the general performances of the ROC curves, for the test and validation datasets, the RVM model obtains the best result which is proved by the related AUC results as presented in Table 5.10.



(a) The ROC curves for the 320-D training dataset.



(b) The ROC curves for the 320-D testing dataset.



(c) The ROC curves for the 320-D validation dataset.

Figure 5.8: The ROC curves for the cross-validation analysis using 320-D original database. (a) Training dataset; (b) Test dataset and (c) Validation dataset.

Subset	MLP	RBF	SVM Regression	RVM
Training	**1**	0.7357	0.9142	0.9997
Test	0.9140	0.6837	0.8633	**0.9440**
Validation	0.9118	0.6904	0.8665	**0.9439**

Table 5.10: AUC values for cross validation testing on 320-D original training, test and validation subsets. The AUC values of three datasets show that the RVM classifier outperforms other methods for the test and validation datasets with the AUC values: 0.9440 and 0.9439.

Prediction results based on 320-D reconstruction data

In Subsection 5.1.3, the 320-reconstruction database has been introduced. The purpose of creating this database and applying it in the prediction models is to verify the little effect of the removed eigenvectors on describing the characteristics of the DNA-binding Cys_2His_2 zinc finger interaction. In this part, the related results will be presented and discussed.

In Table 5.11, the normalised classification error of all applied prediction models based on the 320-D reconstruction database are summarised. Same as the 320-D original database, the MLP shows the lowest normalised error for both test and validation data subsets followed by the RVM. Details of the normalised classification error for the selected methods are represented in Appendix C.4.3.

	Linear Reg.	KNN	MLP	RBF	SVM (Clas.)	SVM (Reg.)	RVM
Training	0.5584	——	7.5916×10^{-13}	0.4685	——	0.5540	0.0128
Test	0.6723	0.2693	**0.2116**	0.5571	0.3753	0.6073	0.2269
Validation	0.6625	0.2835	**0.2133**	0.5607	0.3327	0.6063	0.2257

Table 5.11: The normalised classification error for the 320-D reconstruction data. The MLP has the lowest normalised classification error for both test and validation data subsets: 0.2116 and 0.2133.

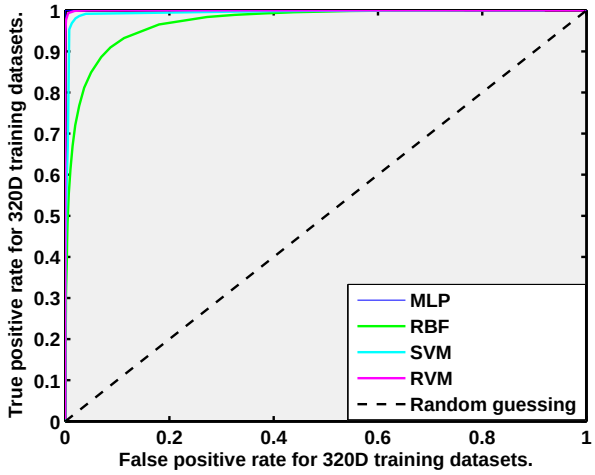
Shown in Table 5.12, the SVM classification has the best prediction accuracy for both the test and validation data sets. Moreover, the prediction accuracy of some classifiers, such as the SVM classification, SVM regression, RVM and linear regression, is identical to that of the 320-D original database. It can be inferred that the first 233 eigenvectors obtained by PCA are adequate to represent the characteristics of the DNA-binding Cys_2His_2

zinc finger interaction. For other prediction models, the accuracy is slightly better than that in Table 5.9. This can be understood that the unused eigenvectors in reconstructing the 320-D database are relatively irrelevant to describe the characteristics of the DNA-binding protein interaction.

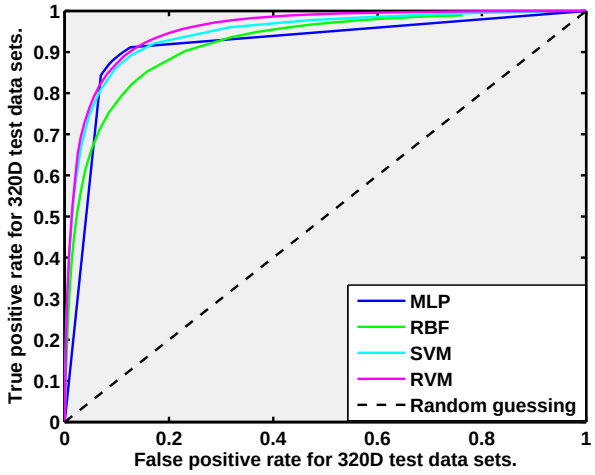
	Linear Reg.	**KNN**	**MLP**	**RBF**	**SVM (Clas.)**	**SVM (Reg.)**	**RVM**
Training	0.9295	——	**1**	0.9115	——	0.9802	0.9936
Test	0.8683	0.8626	0.8944	0.8481	**0.9132**	0.8828	0.8879
Validation	0.8656	0.8581	0.8941	0.8473	**0.9168**	0.8801	0.8876

Table 5.12: The accuracy for the 320-D reconstruction data. The SVM classification model have the best prediction performance again, and followed by the MLP, RVM and SVM regression classifiers. The accuracy of test and validation datasets of the SVM classification is 0.9132 and 0.9168 respectively.

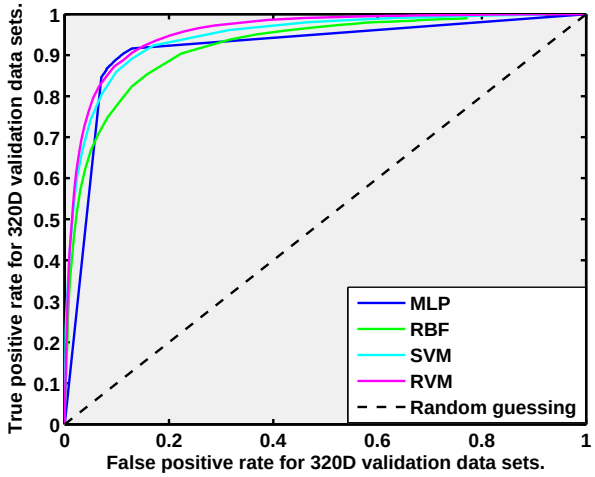
In Figure 5.9, the ROC curves for the cross-validation analysis are plotted using the 320-D reconstruction database, respectively. Since the information represented by the reconstructed database is almost identical to the original database, the ROC curves illustrated in Figure 5.9 have similar trends as with Figure 5.8. The RVM approach still has the best performance in the ROC curves and the highest AUC results which are listed in Table 5.13.



(a) The ROC curves for the 320-D training dataset.



(b) The ROC curves for the 320-D testing dataset.



(c) The ROC curves for the 320-D validation dataset.

Figure 5.9: The ROC curves for the cross-validation analysis using 320-D reconstruction database. (a) Training dataset; (b) Test dataset and (c) Validation dataset.

Subset	MLP	RBF	SVM Regression	RVM
Training	**1**	0.7367	0.9142	0.9983
Test	0.9145	0.6838	0.8633	**0.9543**
Validation	0.9162	0.6944	0.8665	**0.9542**

Table 5.13: AUC values for cross validation testing on 320-D reconstructed training, test and validation subsets. Due to the overfitting, the MLP model shows the best fitting of the training dataset. However, through comparing the AUC values of the test and validation datasets, the RVM classifier outperforms other methods for the test and validation datasets with the AUC values: 0.9983, 0.9543 and 0.9542, respectively.

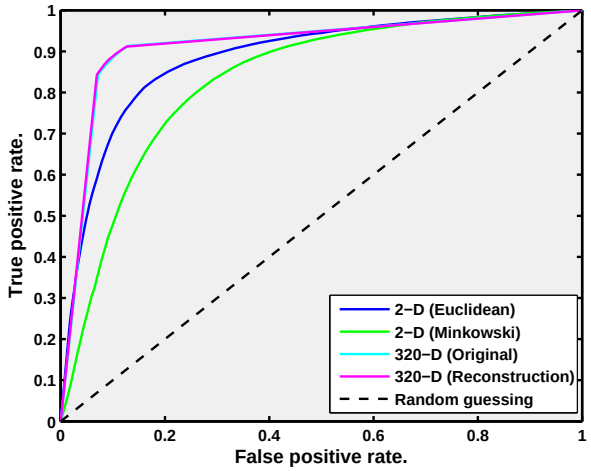
Discussion

In this Subsection, the prediction methods were evaluated through using the 320-D original and reconstruction databases. Identified through the normalised classification error and cross validating the ROC curves and the area under ROC curves, the SVM classification has the best performance on predicting the data examples from the 320-D databases. Meanwhile, based on the accuracy and the ROC curves, the PCA based 320-D reconstruction database can be confirmed to have the ability of representing the integral information of the DNA-binding Cys_2His_2 zinc finger interaction.

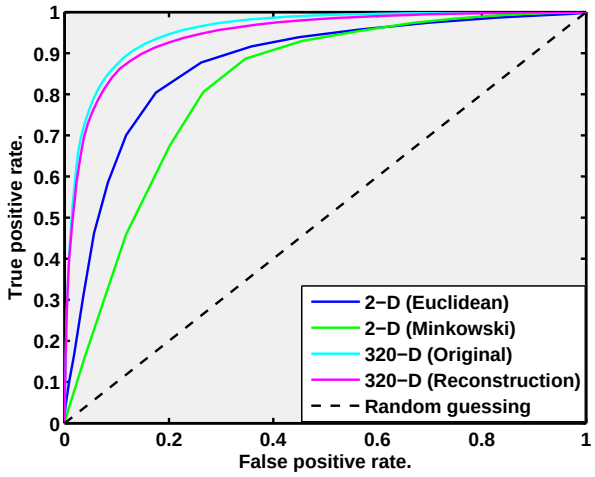
5.2.4 Discussion

In Subsection 5.2.2 and 5.2.3, the prediction results of various selected classifiers based on both the 2-D reconstruction databases and the 320-D databases have been analysed and discussed, respectively. For the 2-D reconstruction databases, the performance of the prediction models using the 2-D Euclidean based reconstruction database as input is generally better than using the 2-D Minkowski based reconstruction database. Moreover, the SVM regression classifier outperforms other prediction methods when adopting the 2-D reconstruction databases as inputs. Different from the results for 2-D reconstruction databases, the performance of the classifiers using either the 320-D original database or the 320-D reconstruction database is similar. The performance of the SVM classification model is the best for the 320-D databases. The purpose of this Subsection is to cross validate the results depend on these four databases.

Figure 5.10 and Figure 5.11 show the ROC curves of the MLP and RVM models which use the test and validation datasets from different databases as the inputs. The related area under the ROC curves (AUC) are presented in Table 5.14 and 5.15 respectively. Through comparing the ROC curves in Figure 5.10 and 5.11 and considering the AUC results as the reference, it is discovered that the 320-D databases have the superiority of building the classification models over the 2-D reconstruction databases. Furthermore, the performance of the MLP classifier is better than the RVM model. This conclusion provides guidance on predicting the binding status of the novel data samples in the next chapter. Table 5.16 summarises the accuracy of each prediction method for the databases, where the SVM classification model as the best classifier has been highlighted.



(a) The AUC values for the ROC curves for different datasets as listed in the figure are 0.8844, 0.8310, 0.9140 and 0.9145 respectively.

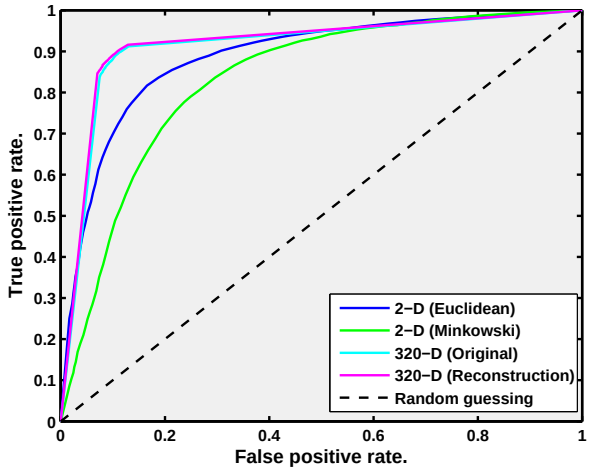


(b) The AUC values for the ROC curves for different datasets as listed in the figure are 0.8734, 0.8212, 0.9543 and 0.9440 respectively.

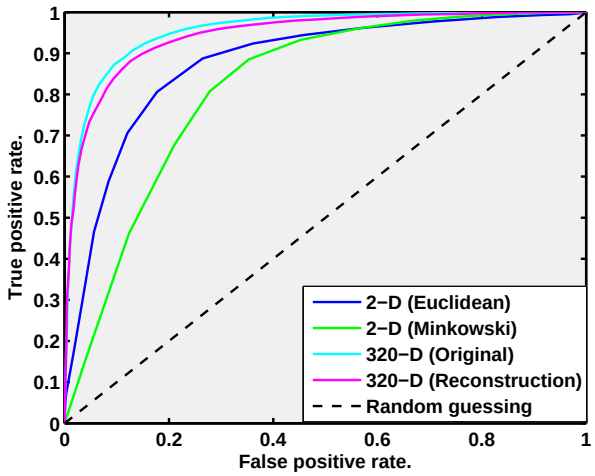
Figure 5.10: The ROC curves of test datasets. (a) The MLP and (b) The RVM. Using the 320-D test datasets helps to produce better prediction results than using the 2-D reconstruction datasets.

Subset	MLP	RVM
2-D (Euclidean)	**0.8844**	0.8734
2-D (Minkowski)	**0.8310**	0.8212
320-D (Original)	0.9140	**0.9543**
320-D (Reconstruction)	0.9145	**0.9440**

Table 5.14: The AUC values for the cross validation testing on test data subsets. As shown in the table, for the 2-D reconstruction datasets, the MLP and RVM models have similar AUC values, where the MLP is slightly better than the RVM. However, for the 320-D datasets, the RVM classifier has better AUC results than the MLP.



(a) The AUC values for the ROC curves for different datasets as listed in the figure are 0.8862, 0.8315, 0.9118 and 0.9162



(b) The AUC values for the ROC curves for different datasets as listed in the figure are 0.8764, 0.8184, 0.9542 and 0.9439

Figure 5.11: The ROC curves for the validation datasets. (a) The MLP and (b) The RVM. Similar to Figure 5.10, using the 320-D validation datasets helps to produce better prediction results than using the 2-D reconstruction datasets and the RVM model outperforms other classifiers.

Subset	MLP	RVM Regression
2-D (Euclidean)	**0.8862**	0.8764
2-D (Minkowski)	**0.8315**	0.8184
320-D (Original)	0.9118	**0.9542**
320-D (Reconstruction)	0.9162	**0.9439**

Table 5.15: The AUC values for the cross validation testing on the validation data subsets. The 320-D validation datasets have better prediction results than the 2-D reconstruction datasets depend on the selected classifiers.

	Subset	Linear Reg.	KNN	MLP	RBF	SVM (Clas.)	SVM (Reg.)	RVM
Test	**2-D (Euclidean)**	0.6191	0.8331	0.8260	0.8075	0.7668	0.8312	0.8152
	2-D (Minkowski)	0.6461	0.7178	0.7680	0.7606	0.7739	0.7840	0.7674
	320-D (Original)	0.8683	0.8221	0.8923	0.8482	**0.9132**	0.8828	0.8879
	320-D (Reconstruction)	0.8683	0.8626	0.8944	0.8481	**0.9132**	0.8828	0.8879
Validation	**2-D (Euclidean)**	0.6118	0.8323	0.8249	0.8067	0.7681	0.8327	0.8143
	2-D (Minkowski)	0.6424	0.7287	0.7698	0.7586	0.7736	0.7846	0.7645
	320-D (Original)	0.8656	0.8581	0.8941	0.8509	**0.9168**	0.8801	0.8876
	320-D (Reconstruction)	0.8656	0.8157	0.8896	0.8473	**0.9168**	0.8801	0.8876

Table 5.16: The accuracy for the cross validation testing on different databases. The SVM classification outperforms other classifiers with the accuracy of test datasets: 0.9132 and validation datasets: 0.9168.

5.3 Synthetic data study

In Section 5.2, various prediction models have been verified by using both 2-D projection datasets and 320-D reconstruction and original datasets. Through evaluating the performance of the models, the SVM, RVM and MLP classifiers produce better performance. The purpose of this section is to apply the best three models to predict the binding affinities of a set of synthetic data samples. As there is no quantitative binding information for the synthetic data, the visualisation model is used to project the synthetic data in the 2-D feature space and colour the projected samples binding affinities according to the different

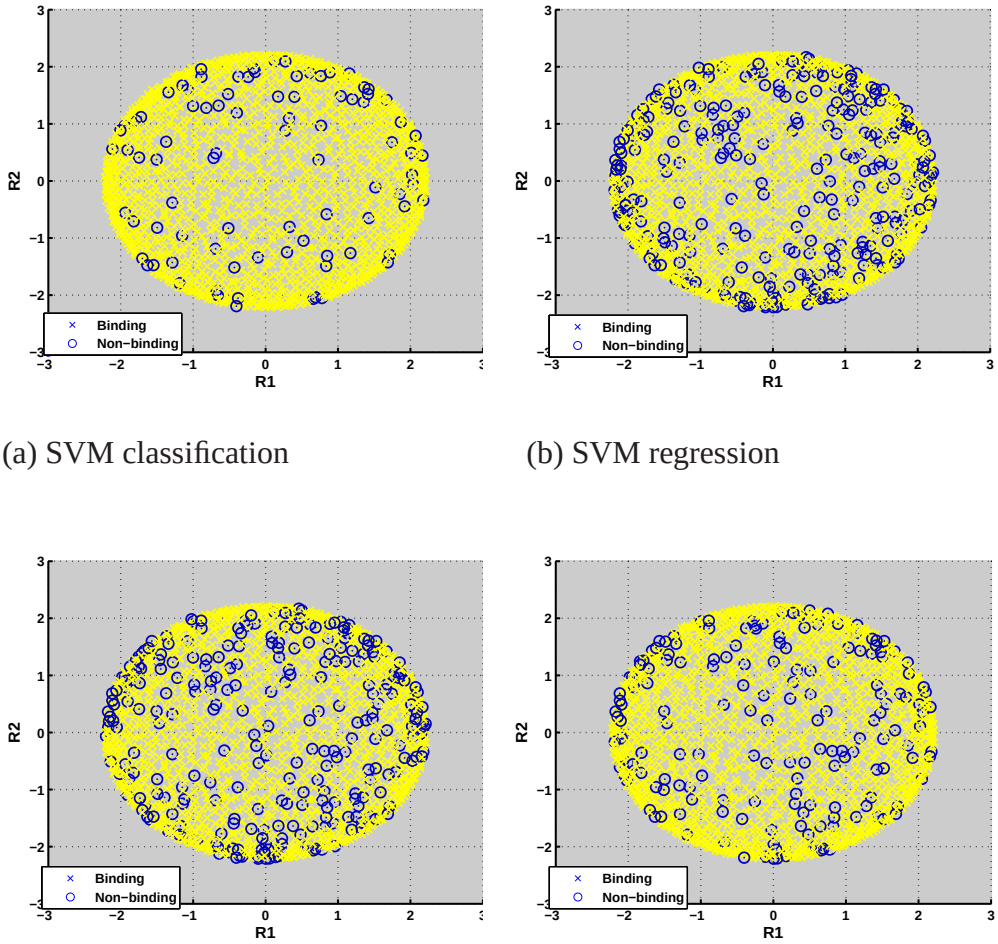
prediction models trained on real data from DB1.

Since only 1860 data sample (DB1 database) are provided with the binding status, to optimise the selected classifiers, 460 data examples are randomly removed from the DB1 database to constitute a know-target test dataset, and the remaining 1400 data samples form the training dataset for the prediction models. Moreover, 1600 synthetic data samples are selected using the Monte Carlo method as the validation dataset.

Figure 5.12 shows the visualisation results of the selected synthetic data samples based on the NeuroScale model. The projected data samples are coloured based on the predicted binding status (binding vs. non-binding). Although the SVM classification model outperforms the SVM regression and RVM models according to the prediction accuracies of the test dataset shown in Table 5.17, through comparing the visualisation results, it is easy to notice that the SVM classification model provides significantly different distribution patterns of prediction results where the predicted number of non-binding examples is far smaller than that of other classifiers. As there is no binding affinity provided for the synthetic data, it is difficult to verify the accuracy, which is only possible through experiments in the laboratory. The visualisation is used as an intuitive method to represent the relationships between data samples depending on their relative neighbours in the visualisation space.

	SVM Clas.	SVM Reg.	RVM	MLP
Accuracy of testset	0.9244	0.8935	0.9109	0.9261

Table 5.17: Accuracy of test dataset based on the selected prediction models. According to the table, the MLP model outperforms other classifiers with the prediction accuracy: 0.9261 followed by the SVM classification model.



(a) SVM classification (b) SVM regression

(c) MLP (d) RVM

Figure 5.12: Visualisation results of the synthetic dataset based on NeuroScale model. (a) SVM classification: 87 non-binding examples; (b) SVM regression: 273 non-binding examples; (c) MLP prediction result: 271 non-binding examples; and (d)RVM regression result 169 non-binding examples. Generally, comparing with the SVM classification model, the prediction results of the SVM regression, RVM and MLP model are similar.

5.4 Summary

This chapter has revealed that, without using explicit biochemical information, there is sufficient information in the geometric knowledge to allow the construction of semi-parametric prediction models capable of predicting binding status with high accuracy. The methods of reconstituting the databases and the characteristics of each database were introduced. In this work, the database DB1 which was created using the canonical binding model was selected to constitute the 320-D original database. Moreover, based on the visualisation results presented in the previous chapter, the 2-D databases which are based on the dissimilarity metrics and the 320-D database which depends on the PCA model were reconstructed. In order to evaluate the performance of the selected predication methods, the defined normalised classification error and the ROC curve were introduced as the quality criteria. To implement the prediction of the binding status, six prediction models, i.e., linear regression, *k*-NN, MLP, RBF, SVM and RVM have been employed. Based on the cross validation analysis on the the prediction results, the prediction models using the 320-D databases as the inputs perform better than using the 2-D reconstruction databases. Among all the models, the SVM produces the best performance for both the 2-D reconstruction databases and the 320-D databases, followed by the RVM and MLP classifiers. In the last section, a group of synthetic data samples were applied to verify the classifiers which are selected based on the performances. Moreover, the NeuroScale is used to project relationships between the predicted binding and non-binding examples.

In the next chapter, novel data samples will be introduced. Before applying the prediction models, the data samples will be projected and analysed using NeuroScale based on different dissimilarity metrics. Hereafter, using the findings in this chapter, the binding status of the novel data samples will be predicted and verified.

6 Analysis methods (III): New experimental data study

CONTENTS

In the previous chapters the approaches which are specific to visualise and predict the interaction between DNA and the Cys_2His_2 zinc finger protein have been discussed. NeuroScale, as a topographic visualisation method, has a demonstrated capability of projecting the geometric structure of such protein-DNA interactions from a high dimensional space into a low dimensional feature space by preserving the structure of data samples. The focus of this chapter is to utilise the properties of the previously developed non-linear topographic visualisation techniques to infer characteristics of novel data collected in different experiments. Although the potential binding behaviour between the Cys_2His_2 zinc finger protein and DNA sequence will be explored, other characteristics including physicochemical properties such as hydrophobicity and hydrophilicity will also be explored. In addition, we also investigate classification models superimposed on the visualisation space to aid with prediction of potential behaviour.

In Section 6.1 the characteristics of the new experimental data are explained. Then, the visualisation results of the novel data are presented in Section 6.2, which is followed by the discussion of the results.

6.1 Database of new experimental data

Data, collected from numerous experiments, are the basis of studying and understanding the principles of DNA binding zinc finger interactions. In the previous chapters, the training data set which was named as the DB1 database, was exploited to study structural relationships by visualisation techniques and build prediction models for binding status determination. In this chapter, alongside the training data set, another three data sets which were mentioned in Chapter 3 and listed in Table 3.3, will be applied as novel sources of data to investigate the utility of topographic visualisation as a tool for high dimensional data analytics in the DNA-protein interaction domain.

As defined in Table 3.3, the DB1 database is still utilised as the training data set in this chapter. Moreover, the DB2 database which contains 673 comparative examples is selected as the test data set for cross validating the visualisation results of NeuroScale.

Different from the DB1 and DB2 databases, the validation data sets used in this chap-

ter are created based on two totally different data sources. The 31 data examples in the DB4 database are obtained by filtering out the duplicate data samples from the original data set which is comprised of the experimental data in the publications.

On the other hand, the validation data set defined as DB3 in Table 3.3 is generated based on a combinatorial randomized protein library from real laboratory experiments. As described in Subsection 3.1.2, in the data set, the binding pair at position 2 is fixed as 'GD'. The three binding pairs at position -1, 3 and 6 are selected regarding certain specific data examples in the training data set. Since the binding pair at position 2 is fixed, the published data with different binding pairs at position 2 are filtered out at the beginning[1]. Then, by fixing any two binding pairs in the primary DNA chain, the remaining principal DNA-contacting residue is varied with all 20 amino acids. For example, the interacting bases in the DNA sequence of a data example can be 'gAAA' (3'-5') as the binding pair at position 2 is fixed. The corresponding amino acids at position 2, -1, 3 and 6 are 'DRHW'. If the bases in the DNA sequence and the corresponding amino acids at the first three positions: 2, -1 and 3 are fixed, 19 new data examples can be generated by altering the amino acid at position 6. The 'DRHW' is excluded due to data duplication. Therefore, for each data sample in the DB3 database, only one binding pair is different from the certain specific data examples in the DB1 database.

6.2 New data visualisation

Protein-DNA interactions as introduced in Chapter 2, are an essential feature in the genetic activities of life. The aim of this work is to use predictive modelling to forecast the properties of engineered zinc-finger proteins, such as specificity and efficacy. In the previous chapters, visualisation and prediction methods have been applied to study the DNA-binding Cys_2His_2 zinc finger interactions using published data examples. The purpose of this section is to investigate the ability to potentially predict the characteristics of new data samples using the methods based on the existing conclusions from Chapter 4

[1]There are 593 data samples (31.88%) in the training data set containing binding pair 'GD' at position 2.

and Chapter 5.

Topographic visualisation techniques, especially NeuroScale, are employed to obtain in-sights into the relative distributions of the protein-DNA combinations that exist in nature. In Chapter 4, the structure distributions of the training data set have been analysed. In this subsection, the study is focused on using NeuroScale to evaluate how new data is mapped conditional on previous trained models. Moreover, the visualisation results which represent the similarities between the data samples are expected to provide some clues for the possibility of studying the potential binding behaviour and properties of new data.

The schematic of the NeuroScale model is aimed at projecting the geometric structures of data samples from high-dimensional data space into low-dimensional feature space through preserving the dissimilarities. To visualise the novel data $\mathbf{x}_{\text{new}}$ using a previous trained model based on the training data set $\mathbf{x}_{\text{old}}$, the distance d^*_{newold} [2] between them in the original space is measured by the dissimilarity metric. In this thesis, the Euclidean and Minkowski metrics are exploited. Then, through applying the non-linear transformation f which is effected by the RBF model with well trained parameters, the novel data can be projected in the feature space as $\mathbf{y}_{new}$[3].

6.2.1 Modifications to the projection model

NeuroScale, as discussed in Subsection 4.3.1, employs a non-linear transformation to preserve geometric structure while mapping the data from the original configuration space into the feature space. In other words, it is trying to preserve the relative 'dissimilarities' between data samples when projecting the data from the original space into the transformed space where 'dissimilarity' can be chosen to reflect biological knowledge. Although the training database is created based on data derived from 25 publications, the number of data samples is still limited compared with the 41 million possible configurations in the theoretical space of all combinations. Therefore, it is worth studying the most

[2] where $d^*_{newold} = d^*(\mathbf{x}_{new}, \mathbf{x}_{old})$.

[3] Where $\mathbf{y}_{new} = f(\mathbf{x}_{new}; \mathbf{W})$.

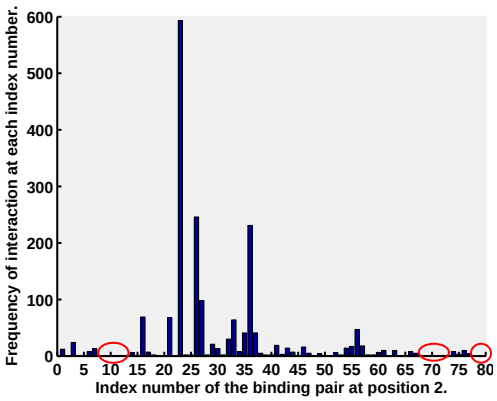
extreme situation where data samples are completely different from the training data set.

By counting the number of occurrences of each possible binding pair[4] at four binding positions (position 2, -1, 3 and 6) based on the training data set, a group of statistical histograms are plotted in Figure 6.1. According to the statistical histograms, some binding pairs have been found which never happened at the positions in the training data set as highlighting in the figure. Through listing these binding pairs, nine data samples[5] which are completely different from the training data set are generated by randomly selecting the never occurring binding pairs at each position.

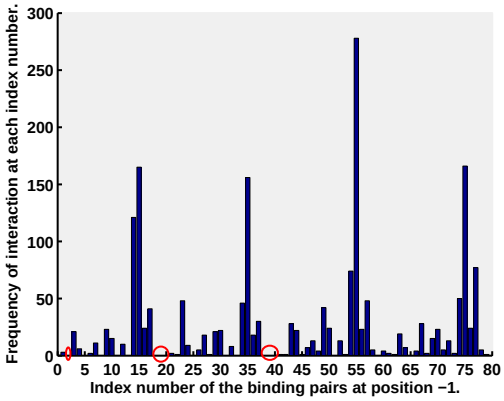
Applying NeuroScale to project the structure relationship between the training data samples and the generated data samples from the 320 dimensional space into the feature space, the visualisation result is shown in Figure 6.2. It is unexpected that the nine generated data samples are projected to the same position as highlighted. By checking the dissimilarities between the new data and the training data set, it was discovered that all generated data samples have the same Euclidean distances to the training data set in the original space. This discovery can explain why the projected new data samples overlap at the same position. The reason for having the same distances in the high-dimensional space is because of the representation model. As described in Subsection 3.2.2, each data sample is represented as a sparse, binary 1×320 vector in which only four '1's are present for the specific interaction positions and the rest are all '0's. Moreover, when using NeuroScale to implement the visualisation, the hidden centres of the RBF model are randomly selected from the training data set. Since the generated data samples are completely different from the training data set, the dissimilarities measured by the Euclidean metric between the hidden centres and the generated data in the original space always have the same maximum value at 2.828. Therefore, when projecting the preserved structural relationships in the feature space, the generated data samples are plotted at the same positions as the black stars in Figure 6.2.

[4] According to the canonical binding model, at each binding position, there are 80 possible binding pairs (20 amino acids $\times$ 4 bases).

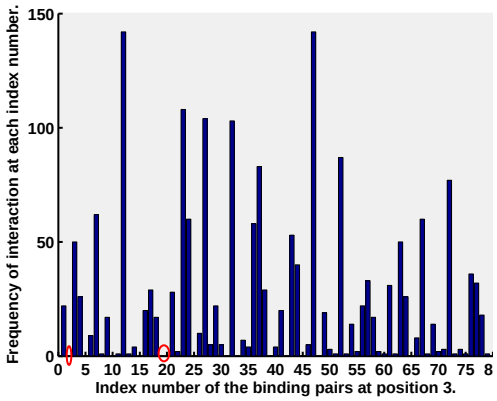
[5] Detailed information of the nine data samples are listed in Appendix D Table D.1.



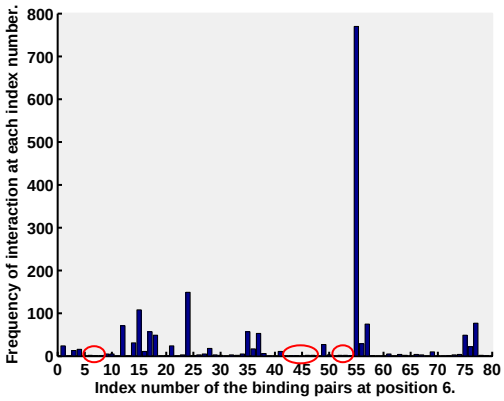
(a) Position -1.



(b) Position 2.



(c) Position 3.



(d) Position 6.

Figure 6.1: Statistical histogram of the interaction frequency at different binding positions. According to the histograms, the binding pairs which do not exist in the training data set at the four positions are indicated by the highlight bubbles. By randomly selecting the highlighted binding pair from each position and combine them together, the completely different protein-DNA interaction data sample is generated.

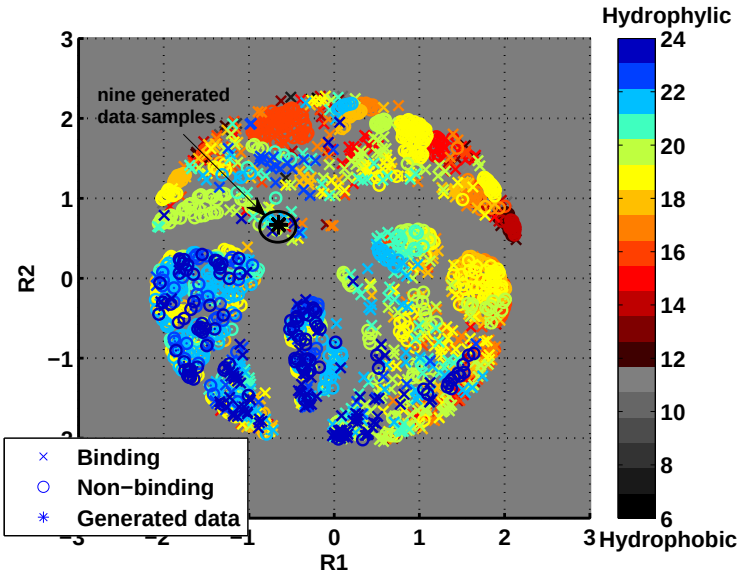


Figure 6.2: The visualisation result of nine generated data samples using NeuroScale. As pointed out in the figure, nine data samples overlap at the same position which means the 'dissimilarities' between these generated data samples and the training data set are same in the high-dimensional structure space. This phenomenon is caused by the special sparse, binary expression of the data samples. In addition, the visualisation result is coloured according to the hydrophilicity and hydrophobicity of the proteins.

To overcome this problem, each RBF centre is treated as a continuous point in the 320 dimensional space, sampled from a 4-component Gaussian Mixture model. Four normal distributions[6] are considered to substitute for the '1's at the four specific interaction positions. For example, given a hidden centre, the positions of four '1's can be confirmed from the 1×320 vector. The mean μ is the position of the '1', and the value of x can change between 1 and 320. To obtain the best substitution effect, different choices of the standard deviation σ were investigated. When $\sigma = 1$, the *Sammon stress* achieved the lowest stress value. Then, four generated normal distributions are combined together and normalised. Figure 6.3 shows an example of the generated hidden centres. In this figure, the standard deviation σ is changed from 1 to 5, and the relevant results are plotted in different colours. Using the optimised hidden centres, the generated data samples can now be plotted separately in Figure 6.4, and the created validation data set will be visualised in

[6] The normal distribution is defined as $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ where μ is the mean and σ is the standard deviation.

the same way. In Figure 6.4, some generated data samples are projected outside the main area of the training data set. It illustrates that the training and the validation data sets have significant dissimilarities, and for these data samples, it is hard to infer the binding properties.

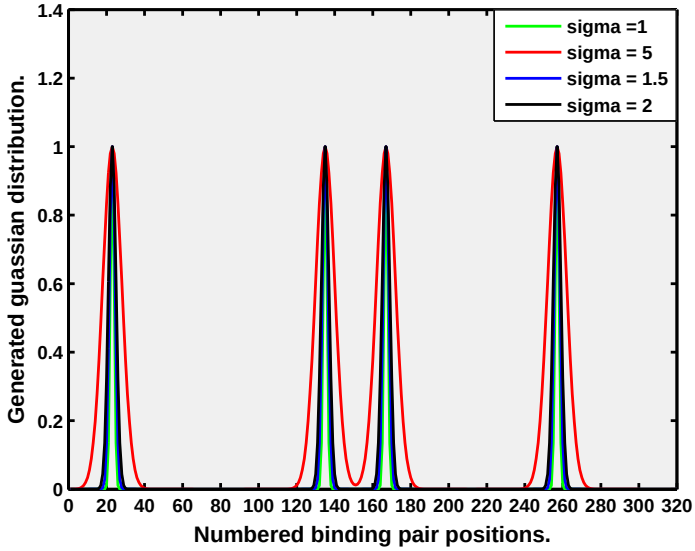


Figure 6.3: The generated hidden centre using the normal distribution. When the standard deviation σ equals to 1, the generated hidden centre is closer to the original one. Increasing the value of σ, the overlapping between each specific interaction position becomes more and more significant.

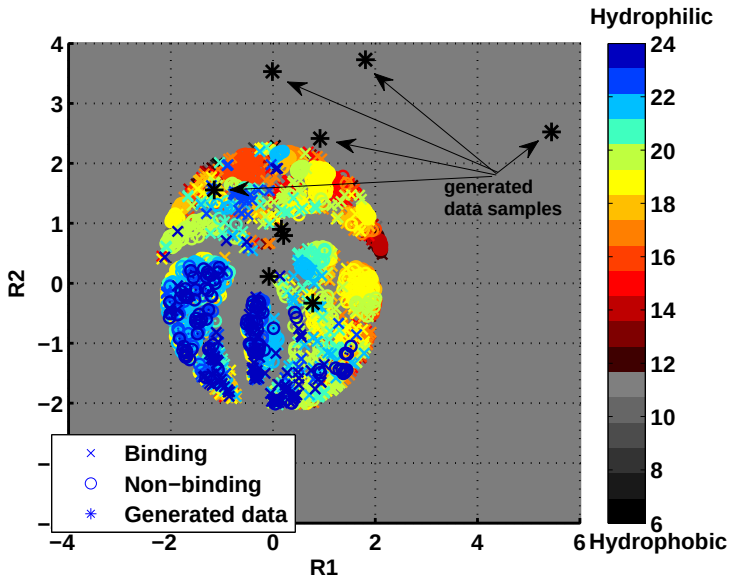


Figure 6.4: The visualisation result of nine generated data samples using NeuroScale based on the optimised hidden centres superimposed on the training data projections colour coded by the hydrophilicity and hydrophobicity of proteins. Comparing with Figure 6.2, the positions of the nine generated data samples are separated in the feature space. As pointed out in the Figure, four data samples are projected external to the main area, which are therefore considered to have significant distances from the training data set.

6.2.2 Visualisation results of the new data

Through applying the normal distribution to define the selected hidden centres of the RBF model, specific binary data samples that are significantly dissimilar can be projected into low dimensional feature space by NeuroScale. This circumvents the problem of constraining the centres to be located on the discrete grid which led to the anomalies observed in Figure 6.2. In this subsection, the visualisation results of the selected test and validation data sets based on the Euclidean metric as the dissimilarity measurement in high dimensional space are represented in Figures 6.6, 6.7 and 6.9 respectively [7]. Moreover, as a benchmark of the visualisation study, the PCA model is employed to project the test data into the feature space, shown in Figure 6.5. Similar to the visualisation result of the training data set in Subsection 4.2.2, the represented test data samples with different structure properties are not separated appropriately, due to the linear nature of PCA.

[7] The visualisation results of the selected test and validation data sets based on the Minkowski metric are represented in Appendix D.

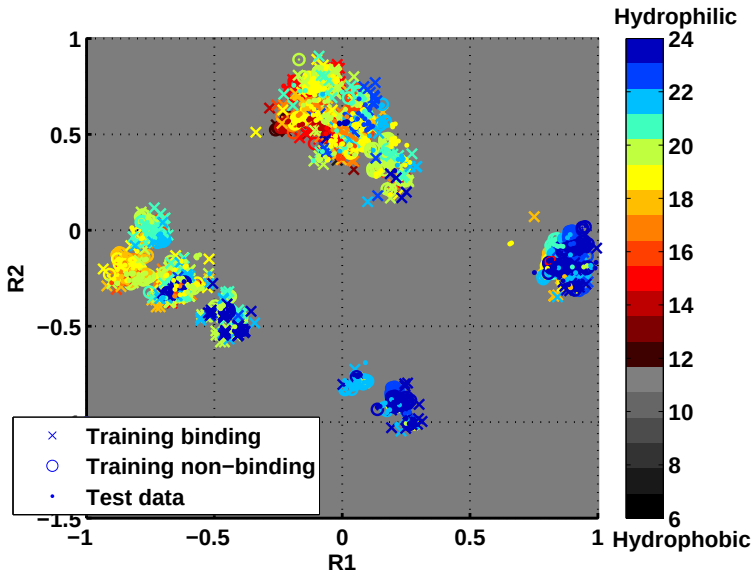
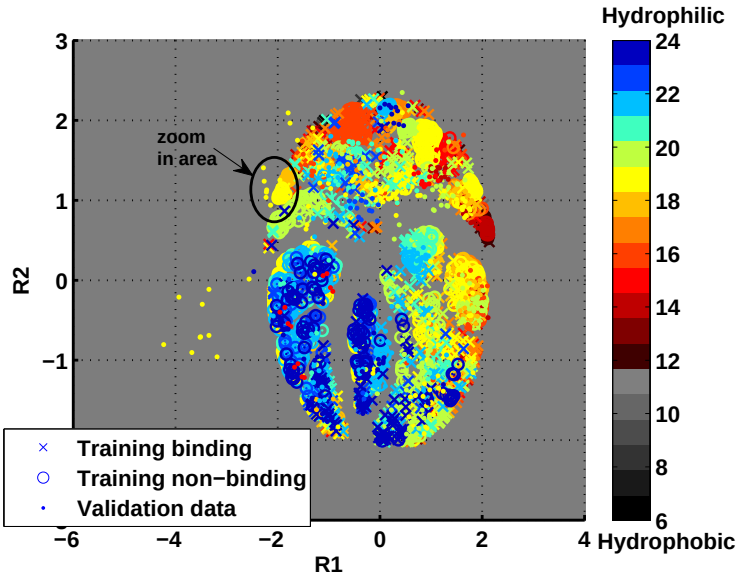
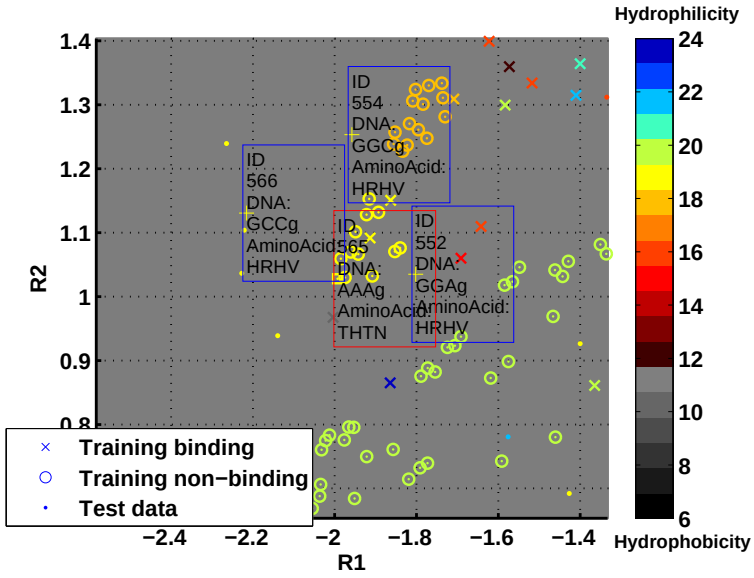


Figure 6.5: Visualisation result of the test data set based on the PCA model. Both training and test data projections colour coded by the hydrophilicity and hydrophobicity of amino acids. Although in the feature space, all test data can be projected into the training data clusters, it is hard to distinguish the properties of the interactions such as binding status and biochemical structures.

Figure 6.6(a) shows the projection result of the test data set (673 data samples from the DB2 database) where only the DB1 database (1860 data samples) is used to train the NeuroScale network. Compared with Figure 6.7(a) where the network is trained based on both the training and test data sets, several of the test data samples in Figure 6.6(a) are projected external to the main visualisation area. When all data samples (both training and test data sets) are applied to train the model, the test data set can be clustered more appropriately. As NeuroScale implements the high-dimensional data visualisation by preserving the geometric structure from the original configuration space into the feature space, the data samples with similar structures can be grouped together. Specific to Figure 6.6(a), the data samples which are projected away from the majority groups illustrate that the structures of these data ought to be more different compared with others. However, through studying the biochemical information of these data as shown in Figure 6.6(b), there is no significant structural difference between the test data samples which are projected external and internal to the main visualisation area. In addition, although the selected test data in Figure 6.6(a) have the same amino acid colour coding with the surrounding training data, the structural information is relatively independent.



(a) In this figure, a few test data samples are projected external to the main visualisation area. In order to find out the reason for this phenomenon, a small range of data samples are selected and magnified.

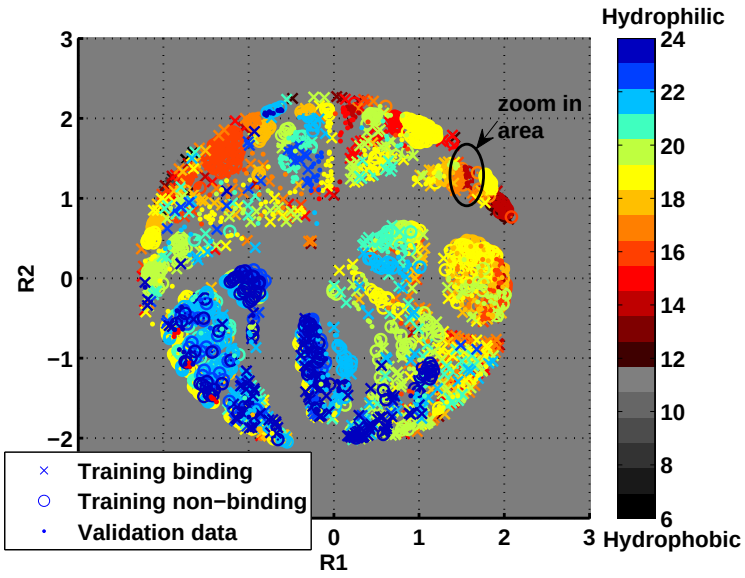


(b) In the figure, the structural information of the selected data samples are represented. The data points with IDs 554, 552 and 566 are selected from the test data set. They have the same amino acid combinations 'HRHV'. The training data sample with ID 565 in the middle has a completely different structure information where the DNA sequence is 'AAAg' in the order from 5 to 3, and the amino acid list is 'THTN' with position order 2, -1, 3, 6.

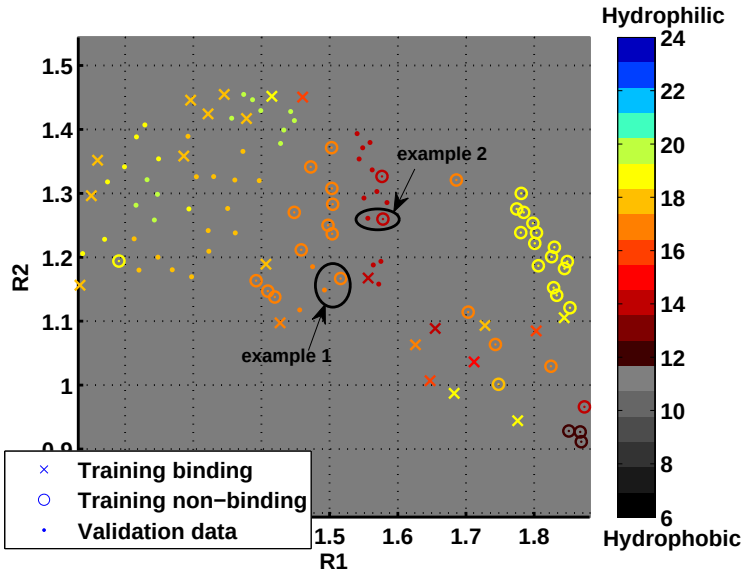
Figure 6.6: Visualisation result of the test data set in which only the training data set has been trained. (a) is the result using only the training data set to train the visualisation model. (b) is the magnified visualisation result with the detailed structural informations of the selected data samples.

In order to understand the phenomenon in Figure 6.6(a), the histograms of the dissimilarities [8] in the high-dimensional space are represented in Figure 6.8. The structures of the outside data samples are rather similar. Compared with Figure 6.8(b), the range of the averaged dissimilarities between the test and training data samples in Figure 6.8(a) varies slightly with a similar distribution. There is a smaller proportion of the test data samples with distances between 2.55 and 2.65 than the statistical result on which the histogram of the average distance between the training data samples is based. Moreover, given the visualisation result in Figure 6.6(a), it can be seen that the test data samples which are projected away from the main area are not bound to have the largest distance from the training data sets. The reason for this phenomenon is still under investigation. It may be caused by the selection of the dissimilarity measure, or due to the quality criteria considered or an issue of extrapolation by the RBF model. Therefore, as a possible direction for future work, a reliable and robust method that can evaluate the biochemical similarity between a novel data sample and the existing database needs to be identified.

[8] Different from the histogram in Subsection 4.3.4, the histograms plot in Figure 6.8 are based on the averaged distance from each test data samples to all training data samples or the averaged distance from each training data samples to all other training data points.

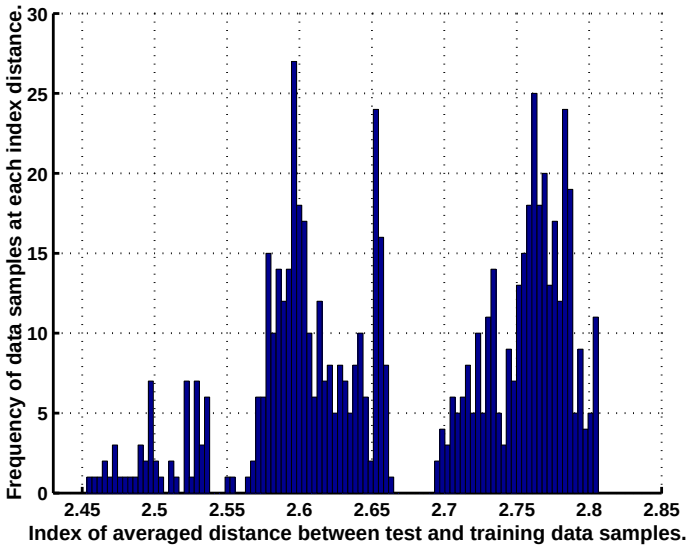


(a) When both training and test data sets are used to train the NeuroScale model, all data samples can be projected into the main visualisation area. Most of test data are projected to the clusters which have the same amino acid colour coding. To verify their structural relationships, a small range of data samples are selected and plotted in Sub-Figure b.

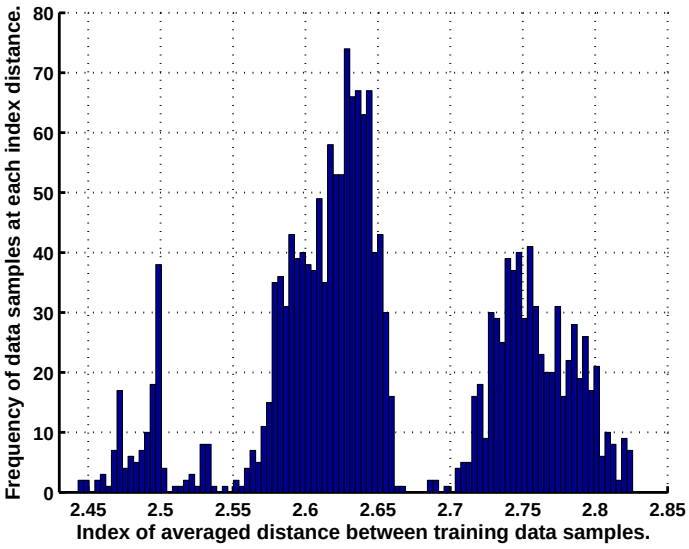


(b) In the zoomed in area, four data samples compose two examples. In example 1 the information of the two data samples are CTTg-GTNE (training non-binding) and CGTg-GTNE (test data); the information of example 2 are CTAg-GTAE (training non-binding) and CCAg-GTAE (test data). In the two examples, the DNA sequences have similar information, and the amino acids have the same hydrophobicity and combinations.

Figure 6.7: Visualisation result of the test data set in which both the training and the test data set were used to optimise the mapping. (a) is the result only use the training data set to train the visualisation model. (b) is the visualisation result in which both the training data set and the validation data set have been used to create the mapping.



(a) Histogram of the averaged Euclidean distance between each test data samples and the training data set in the feature space.

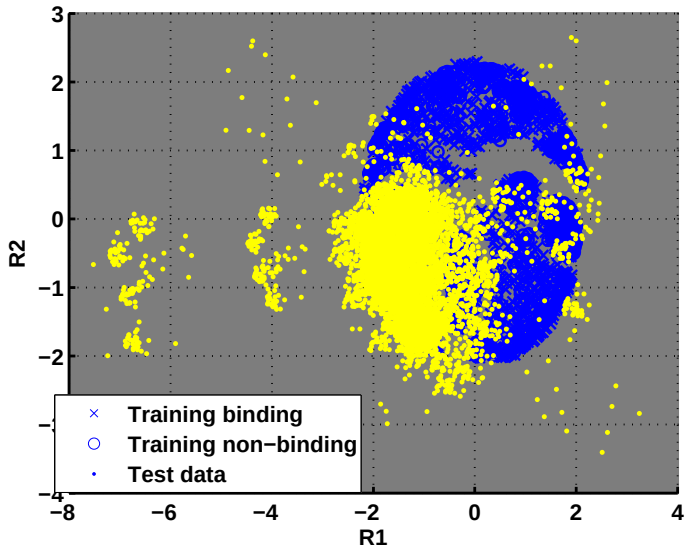


(b) Histogram of the averaged Euclidean distance between the training samples in the feature space.

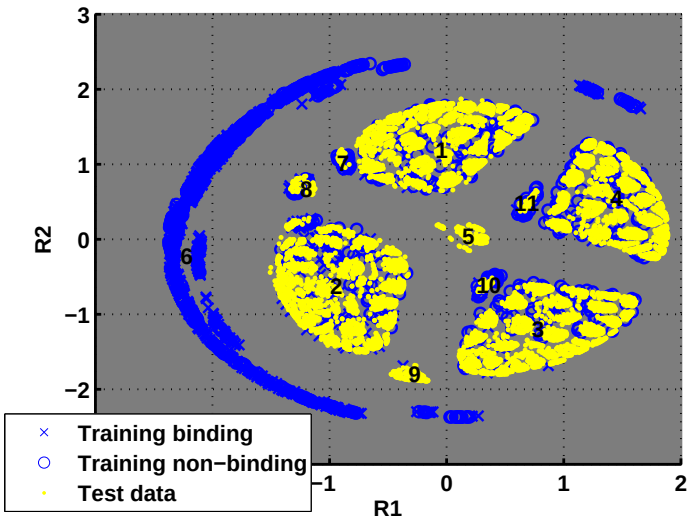
Figure 6.8: Histogram of the averaged Euclidean distance between different data sets. The distance changes from 2.4 to 2.85, and data samples are mainly in three distance ranges: 2.45 to 2.5, 2.55 to 2.65 and 2.7 to 2.8. (a) is the histogram of the averaged distance between each test data samples and the training data set. (b) as a reference plots the histogram of the averaged distance between each test data sample and the training data set.

Different from Figure 6.6, when the novel data samples are also applied to train the model, the visualisation results shown in Figure 6.7(b) can reflect more accurate structural relationships which are verified by two selected examples. In the two examples, the

selected data samples not only have the same hydrophobicity properties (same colouring code of the amino acid combinations), but also contain similar structural information in DNA sequences. To further confirm it, a validation data set (7615 data samples in database DB3) was employed, and the relevant projection results are shown in Figure 6.9. Figure 6.9(a) represents the visualisation result which is based on only the 1860 training data samples. Although the validation data set is much larger than the training data set, as each validation data sample only has one position different from the training data set, most of the validation data samples still can be projected into the main visualisation area in the feature space, while a few of data samples are represented outside. The relevant histograms of the dissimilarities in the high-dimensional space are shown in Appendix D Figure D.4. Similar to Figure 6.8, it is hard to find any clue directly only depending on the dissimilarities to explain this phenomenon. Contrast to it, the NeuroScale model trained using both training and validation data sets provides a very different result as shown in Figure 6.9(b). In this figure, the training data samples containing completely different structural informations are projected into cluster 6, where all of the validation data samples and the relevant training data samples are projected into the main visualisation area. Through studying the data samples of the clusters, it is illustrated again that the visualisation capability of the NeuroScale model trained using all data samples can reflect the structural relationships between the data samples, and the properties of the new data samples can be inferred according to the well known neighbours (training data samples), such as binding specificity, hydrophobicity and hydrophilicity of the amino acid combinations, structural features of the interactions and polarisation. The visualisation results of both the test and validation data sets by using the Minkowski inner product to measure the dissimilarities in the data space are plotted in Appendix D Figures D.2 and D.3, respectively. The relevant histograms are shown in Appendix D Figure D.4.



(a) In this Figure, most of validation data samples are projected onto the left side of the main visualisation area. However, by checking the structural information, we found that the validation data samples projected in this area do not have similar amino acid combinations as the training data samples. Moreover, there is no significant difference between the data samples which are projected external and internal to the main visualisation area.



(b) In this Figure, the clusters are numbered from 1 to 11. Except for cluster 6 which has completely different interaction structures, most clusters are projected based on the DNA sequence, as the amino acids have the similar combination information: 'DRXX' following the binding position 2, -1, 3, and 6. In cluster 1 and 7, the DNA sequences are 'XXAg'; Cluster 2 and 8 have the DNA sequence: 'XXGg' and 'XXCg' appear in cluster 4 and 11. In cluster 5, the DNA sequence is 'GXXg' and the amino acids is 'DKXX'. Cluster 9 has DNA pattern 'GXXg' and the amino acids 'DTXX'.

Figure 6.9: Visualisation results of validation data set (database DB3). (a) is the result only using the training data set to create the visualisation model; (b) is the result of using both the training and validation data sets to optimise the NeuroScale model.

6.3 Summary

Different from the databases which were applied to investigate the visualisation and prediction methods in the previous chapters, the validation data sets using in this chapter are created based on two totally different data sources: one is created based on the duplicated data samples from the same database as the training data set; the other is generated depending on the combinatorial randomized protein library and with only one binding pair selected different from the specific data examples in the training data set. Since the training data set is small compared with the theoretical database, the most extreme situation where the data samples are completely different from this data set were considered. A Gaussian distribution method was exploited to pre-process the hidden centres of the RBF model when using NeuroScale to implement the visualisation. Although the validation data samples are similar to the training data set, there are samples projected away from the majority of clusters in the visualisation results without retraining. The reason for the phenomenon is still under investigation, but the selection of the dissimilarity metric and the stress function are considered to be possible factors. On the contrary, when the NeuroScale model was re-trained on the whole data set, the representation results for the binding status prediction become much improved. Through the visualisation results, the characteristics of novel data samples, such as binding specificity, hydrophobicity and hydrophilicity of the amino acid combinations, the structure features of the interactions can be inferred based on the neighbouring data samples. According to the visualisation results, there is sufficient information in the binary coding space for DNA-binding Cys_2His_2 zinc finger interactions. Moreover, it also proves that the properties of novel interactions are predictable based on neighbour properties in the map which is obtained by the visualisation model.

7 Conclusions

CONTENTS

Cys_2His_2 zinc finger binding DNA interaction as one of the typical protein-DNA interactions has been widely studied in the past two decades due to its ability of recognising specific DNA sequences. The characteristics of the interaction can be represented by high-dimensional data. Topographic visualisation, as one of the visual informatics methods, can be used to map the data from a high dimensional space into a low dimensional space by preserving the structure of the data. In this thesis an analysis system has been developed to indicate properties of novel DNA-binding Cys_2His_2 zinc finger interaction through exploiting the topographic visualisation method to represent the high-dimensional structure properties of the well known interactions into low-dimensional feature space.

This chapter aims to review the findings which have been discussed in the previous chapters following the four study phases mentioned in Section 1.2, and indicate directions of future research of the analysis system.

7.1 Summary of the Thesis

To create an analysis system that can indicate properties of novel DNA-binding protein activity, the main contributions of the thesis include the introduction of a dimension-reducing, topographic transformation of the reconstructed binary data to investigate structural relationship in high-dimensional space which define the Cys_2His_2 zinc finger interaction space. There are various techniques available for dimensionality reduction. As the purpose of this work was to study the structural properties of the interaction, the models which can implement topographic visualisation were considered such as GTM, SNE, LLE and Sammon mapping approaches. Moreover, PCA as a classic projection model was also employed as a benchmark. Through applying the created database to investigate the performances of such methods, it was discovered that most of them are not suitable for this work. For the probabilistic based GTM and SNE models, they have advantages on mapping the continuous data samples, but not the sparse binary dataset. The PCA method as a linear model is not good at visualising the high dimensional non-linear data samples. According to the results, the Sammon mapping has the best performance, but can not vi-

sualise new unseen data directly. Due to the reasons mentioned above, the NeuroScale model as a Sammon mapping related algorithm was exploited in this work. NeuroScale is a topographic feature extraction method that can be used to implement lower-dimensional topographic mapping representations for high-dimensional data visualisation. To the author's best knowledge, this thesis is the first to apply NeuroScale and consider different dissimilarity measurements in studying the structural relationship of the partially published DNA-binding Cys_2His_2 zinc finger interactions (partially due to the fact that most of the theoretical interactions do not exist in nature.). Based on the visualisation results, various prediction methods were considered to investigate the possible advantages of using the projection data, rather than the original or reconstructed high-dimensional data as the input to implement the interaction prediction.

Before applying various analysis methods to study the properties of the interactions, an appropriate representation model was essential to convert the original data to sparse binary vectors. As analysed in Section 3.1, the number of the available data samples of the DNA-binding zinc finger protein interaction in this thesis was restricted by laboratory experiments and affects the subsequent analysis and model-building. A canonical binding model that is only based on the information of the DNA sequence and Cys_2His_2 zinc finger was employed to describe features of the selected data samples as explained in Section 3.2. Moreover, a synthetic database which contained all the 41 million possible interactions was created as a reference.

The investigation of dimensionality reduction methods for visualising the converted 320-dimensional database on real experimental data as the second step of this work was discussed in Chapter 4. As the database was formed by 320-dimensional sparse binary vectors and the visualisation models are expected to project the novel data only based on the prior data samples, most of the reduction methods such as Sammon's mapping, were abandoned since they are intrinsically inappropriate to be applied to novel data without retraining the full model [1]. NeuroScale, as the primary visualisation approach was finally exploited. The principle of the model is to preserve geometric structures with a non-linear

[1] The relevant visualisation results of various visualisation models, such as PCA, GTM, LLE, SNE and Sammon's mapping were plotted in Subsection 4.2.2.

transformation while mapping the data from the original configuration space into the feature space. The geometric structure is described by relative 'dissimilarities' which is the 'distance' between data points in the original and transformed spaces, respectively. Since the contributions of the four binding positions in the interaction are different, although the application of the commonly used Euclidean metric can represent the dissimilarities very well, this thesis also explored other metrics such as the Minkowski indefinite inner product[2] to measure the 'dissimilarities'. The *Sammon stress metric* was discussed in Subsection 4.3.2 as the approach to measure and control the quality of the visualisation results, especially when different dissimilarity metrics are applied. The details of the employed dissimilarity metrics were introduced in Subsection 4.3.3. The representation corresponding to the indefinite Minkowski metric, shown in Subsection 4.3.6, revealed a better separation of binding versus non-binding samples, implying that the deployment of non positive definite metrics has some benefits in this situation, compared with the visualisation result based on the Euclidean metric. This interesting discovery indicates that ideally a more biologically-plausible dissimilarity measure should be explored, which is worth further investigations in the future.

To investigate this inference, in Chapter 5, some typical classification models were applied, such as linear regression, *k*-NN, MLP, RBF, SVM and RVM, where*k*-NN was specially selected for the 2-dimensional projective data. ROC curves and relevant AUC figures were also used to evaluate the performance of each method by using different input data. Although the visualisation result shows that the Minkowski metric based model can obtain a better separation result, the related prediction result was worse than those from the classifiers that were re-trained using the 2-dimensional Euclidean metric based projective dataset. Moreover, through comparing the normalised error of each classification model and cross validating the ROC curves and relevant parameters such as accuracy (sensitivity) and the value of AUC, the classifiers re-trained by the 320-dimensional dataset has greater advantage. Furthermore, regarding the ROC curves and the value of AUC, the MLP, SVM and RVM models outperform traditional classifiers such as the linear regres-

[2]The Minkowski metric is defined by the dimensions of the input space corresponding to connections to the complementary DNA strand (the 2^{nd} position) are weighted with -1 and the connections related to connections to the primary DNA helix are weighted +1.

sion and the RBF models.

The final stage of the work investigated the possibility of indicating the properties of the novel DNA-binding protein interactions based on limited experimental data sources. As discussed in Subsection 6.2.1, the NeuroScale model was exploited to represent the generated novel data based on the combinatorial randomized protein library. Although the NeuroScale model has the capability to visualise new data samples without re-training the model, some data samples having similar structure properties to other samples were still projected external to the main visualisation area, which requires further investigation. On the contrary, when using the whole dataset (both training and novel datasets) to re-train the model, the structure relationships between the data samples were reflected very well. Thus, one can infer the properties of the new data samples, with respect to the known neighbours. In Subsection 6.2.2, the abilities of the selected prediction models (MLP, SVM and RVM) were investigated on the performance of predicting the new data samples using the 320-dimensional datasets for the training. Affected by the class-imbalanced validation dataset, there is no difference in the prediction accuracy between the selected classifiers. Moreover, through verifying the predicted binding statuses of the validation data samples, all models show the same classification results, which may need further verification through laboratory experiments.

7.2 Directions for future work

The future work discussed below aims to improve the existing analysis approach introduced in the thesis to implement the prediction of the interaction between DNA and mutated engineered zinc fingers by:

- **Developing the approach on evaluating the dissimilarities between the existing experimental data and the novel data.** In Chapter 6, the visualisation results reflected that given some new data similar to the training database in the high dimensional space, it is possible that the projected position of the new data is outside the main data group. Therefore, developing an approach by considering the mathematical and biochemical knowledge to evaluate the similarity between the new data and the training dataset will be helpful to obtain clues about where the novel data will be projected in the feature space before the visualisation. On the other hand,

the compatibility of the model can also be improved by merging the synthetic data which have different characteristics from the experimental data.

- **Further investigation of the prediction methods on predicting the DNA-binding activity proteins with limited data sources.** In this thesis, only six classic and commonly used prediction models were employed to investigate the four different databases, where good performance has been observed. There are some other prediction methods that may implement better prediction for this problem domain. Moreover, the performance of the selected models in this thesis could be improved by optimizing parameters, through extending the validation database. In Chapter 4, it has been proved that the visualisation models have the superiority to discover the properties of the DNA-binding protein interactions. Instead of applying the visualisation results as the input data of the prediction model, integrating the visualisation results from various visualisation approaches and applying it as a prediction factor could be another option.

- **Introducing data fusion methods to the predictive system according to the selected prediction models.** As proved in Chapter 5 and 6, according to the ROC curves and relative AUC, more than one prediction model can provide a more accurate prediction for the given data. Applying data fusion methods to integrate the prediction results of the best prediction models could improve the accuracy of the prediction for the DNA-binding activity proteins.

- **Searching for the most appropriate metric on the dissimilarity description.** Topographic projection uses a 'distance' function in input space, a 'distance' function in output space, and a 'distance' function to evaluate performance (Sammon *STRESS*). Each of these distances could be modified. In this thesis we explored, briefly, the effect of modifying the input space metric. Recently others have explored different metrics for performance evaluation (eg. the Bregman divergence (Sun, 2011)) and used likelihood functions to map similarities between distributions (Lee, 1999) rather than isolated data points. A future investigation should try and reflect explicit biological knowledge into constructing more appropriate metrics.

- **Generalise the obtained results to a wider class of protein-DNA interactions.** In the thesis, only the interactions between the Cys_2His_2 zinc finger proteins and DNA sequences have been explored. Whilst this can be justified on the grounds that zinc fingers represent one of the most common types of DNA-binding domains found in the majority of eukaryotic genomes, extending the work to other proteins requires a modification of the coding scheme used to encode the data. Different approaches need to be explored in the future to encode more general proteins.

Bibliography

Alder, B. and Wainwright, T. (1959). Studies in Molecular Dynamics. I. GGeneral Method. *J. Chem. Phys.*, 31(2):459.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185.

AP, D., NM, L., and DB, R. (1977). Maximum likelihood from imcomplete data via the EM algorithm. *J R Stat Soc Ser B*, 39:1–38.

Bae, K.-H., Kwon, Y. D., Shin, H.-C., Hwang, M.-S., Ryu, E.-H., Park, K.-S., Yang, H.-Y., ki Lee, D., Lee, Y., Park, J., Kwon, H. S., Kim, H.-W., Yeh, B.-I., Lee, H.-W., Sohn, S. H., Yoon, J., Seol, W., and Kim, J.-S. (2003). Human zinc fingers as building blocks in the construction of artificial transcription factors. *Nature Biotechnology*, 21:275 – 280.

Belkin, M. and Niyogi, P. (2001). Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering. In *Advances in Neural Information Processing Systems 14*.

Ben-Naim, A. Y. (1980). *Hydrophobic interaction*. Plenum Press, New York.

Benos, P. V., Lapedes, A. S., and Stormo, G. D. (2002). Probabilistic Code for DNA Recognition by Proteins of the EGR Family. *Journal of Molecular Biology*, 323(4):701 – 727.

Berg, J. M., Rodgers, J. R., and Stryer, L. (2006). *Biochemistry*. W. H. Freeman and Company.

Berg, O. and von Hippel, P. (1987). Selection of dna binding sites by regulatory proteins. *J. Mol*, 193:723–750.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res*, 28(1):235–242.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1978). The protein data bank: A computer-based archival file for macromolecular structures. *Archives of Biochemistry and Biophysics*, 185:584 – 591.

Bishop, C. M. (2007). *Pattern Recognition and Machine Learning*. Springer Science+Business Media, LLC, USA.

Bishop, C. M., Svensen, M., and Williams, C. K. I. (1998). Developments of the generative topographic mapping. *Neurocomputing*, 21(1-3):203 – 224.

Blancafort, P., Manenat, L., and Barbas, C. F. (2003). Scanning the human genome with combinatorial transcription factor libraries. *Nature Biotechnology*, 21(3):269–274.

Bohm, H. (1998). Prediction of binding constants of protein ligands: a fast method for the prioritization of hits obtained from de novo design or 3d database search programs. *J. Comput. Aided Mol. Des.*, 12(4):309–23.

Bradley, A. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30 (7):1145–1159.

Bulyk, M. L., Johnson, P. L. F., and Church, G. M. (2002). Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res*, 30(5):1255–1261.

CA, B., AM, G., and GM, C. (1998). Minor groove-binding architectural proteins: structure, function, and dna recognition. *Annu Rev Biophys Biomol Struct.*, 27:105–31.

Carl (1984). Protein-dna reco. *Ann. Rev. Biochem*, 53:293–321.

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: a library for support vector machine. *ACM transactions on intelligent systems and technology*, 2:27:1–27:27.

Choo, Y. and Klug, A. (1994a). Selection of DNA binding sites for zinc fingers using rationally randomized DNA reveals coded interactions. *Proc Natl Acad Sci U S A*, 91(23):11168–11172.

Choo, Y. and Klug, A. (1994b). Toward a code for the interactions of zinc fingers with DNA: selection of randomized fingers displayed on phage. *Proc Natl Acad Sci U S A*, 91(23):11163–1167.

Cook, T., Gebelein, B., and Urrutia, R. (1999). Sp1 and Its Likes: Biochemical and Functional Predictions for a Growing Family of Zinc Finger Transcription Factors. *Annals of the New York Academy of Sciences*, 880:94–102.

Cortes, C. and Vapnik, V. N. (1995). Support-Vector Networks. *Machine Learning*, 20-3:273–297.

Cristianini, N. and Shawe-Taylor, J. (2000). *An introduction to support vector machine and other kernel-based learning method*. Cambridge University Press, New York.

Desjarlais, J. R. and Berg, J. M. (1992). Toward rules relating zinc finger protein sequences and DNA binding site preferences. *Proc Natl Acad Sci U S A*, 89(16):7345–7349.

Desjarlais, J. R. and Berg, J. M. (1993). Use of a zinc-finger consensus sequence framework and specificity rules to design specific DNA binding proteins. *Proc Natl Acad Sci U S A*, 90(6):2256–2260.

Dreier, B., Beerli, R. R., Segal, D. J., Flippin, J. D., and Barbas, C. F. (2001). Development of Zinc Finger Domains for Recognition of the 5'-ANN-3' Family of DNA Sequences and Their Use in the Construction of Artificial Transcription Factors. *Biological Chemistry*, 276-31:29466–29478.

Dreier, B., Fuller, R., Segal, D. J., Lund, C., Blancafort, P., Huber, A., Koksch, B., and Barbas, C. F. (2005). Development of Zinc finger domains for recognition of the 5'-CNN-3' family DNA sequences and their use in the construction of artificial transcription factors. *Biological Chemistry*, 280(42):35588–35597.

Dreier, B., Segal, D. J., and Barbas, C. F. (2000). Insights into the molecular recognition of the 5'-GNN-3' family of DNA sequences by zinc finger domains. *Molecular Biology*, 303(4):489–502.

Elrod-Erickson, M., Rould, M. A., Nekludova, L., and Pabo, C. O. (1996). Zif268 protein-DNA complex refined at 1.6 A: a model system for understanding zinc finger-DNA interactions. *Structure*, 4:1171–1180.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27:861–874.

Feig, M., Onufriev, A., Lee, M., Im, W., Case, D., and Brooks, C. (2004). Performance comparison of genegeneral born and Poisson methods in the calculation of electrostatic solvation energies for protein structures. *Journal of Computational Chemistry*, 25(2):265–84.

G.E.Hinton and Roweis, S. (2002). Stochastic Neighbor Embedding. *Neural Information Processing System: Natural and Synthetic*, 15:833–840.

Gerland, U., Moroz, J. D., and Hwa, T. (2002). Physical constraints and functional ccharacteristic of transcription factor-dna interaction. *Proc. Natl. Acad. Sci.*, 99:12015–12020.

Goldman, B. and Wipke, W. (2000). Qsd quadratic shape descriptors. 2. molecular docking using quadratic shape descriptors (qsdock). *Proteins*, 38(1):79–94.

Greisman, H. A. and Pabo, C. O. (1997). A General Strategy for Selecting High-Affinity Zinc Finger Proteins for Diverse DNA Target Sites. *Science*, 275:657–661.

H, A., H, N., N, H., M, O., M, N., T, F., E, T., S, S., and Y, N. (1995). Molecular cloning, characterization, and chromosomal mapping of a novel human gene (GTF3A)

that is highly homologous to Xenopus transcription factor IIIA. *Cytogenet Cell Genet*, 70(3-4):235–8.

Hames, B. D. (2000). *Instant notes in biochemistry*. Oxford : BIOS Scientific Publishers.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143:29–36.

Holm, L. and Sander, C. (1996). Mapping the Protein Universe. *Science*, 273:595–602.

Hughes, M. D., Nagel, D., Santos, A., Sutherland, A., and Hine, A. (2003). Removing the redundancy from randomized gene libraries. *J. Mol. Biol.*, 331:973–979.

Hughes, M. D., Zhang, Z.-R., Sutherland, A. J., Santos, A. F., and Hine, A. V. (2005). Discovery of active proteins directly from combinatorial randomized protein libraries without display, purification or sequencing: identification of novel zinc finger proteins. *Nucleic Acids Research*, 33:e32.

Isalan, M., Choo, Y., and Klug, A. (1997). Synergy between adjacent zinc fingers in sequence-specific DNA recognition. *Proc Natl Acad Sci U S A*, 94(11):5617–5621.

Isalan, M., Klug, A., and Choo, Y. (2001). A rapid, generally applicable method to engineer zinc fingers illustrated by targeting the HIV-1 promoter. *Nature Biotechnology*, 19(7):656–60.

Iuchi, S. (2001). Three clclass of c2h2 zinc finger proteins. *Cell Mol Life Sci.*, 58(4):625–35.

Jain, A. (2006). Scoring function for protein-ligand docking. *Curr. Protein Pept. Sci.*, 7(5):407–20.

Jamieson, A. C., Kim, S.-H., and Wells, J. A. (1994). In vitro selection of zinc fingers with altered DNA-binding specificity. *Biochemistry*, 33(19):5689–5695. PMID: 8180194.

Janin, J. and Wodak, S. (1985). Reaction pathway for the quaternary structure change in hemoglobin. *Biopolymers*, 24:509–526.

Jollife, I. (2002). *Principal Component Analysis*. Science+Business Media, LLC, USA.

JP, M. and M., C. (1998). Zinc fingers are sticking together. *Trends Biochem. Sci.*, 23:1–4.

Kaplan, T., Friedman, N., and Margalit, H. (2005). Ab initio prediction of transcription factor targets using structural knowledge. *PLoS Comput Biol*, 1:1.

Karin, M. (1990). Too many transcription factors: positive and negative interactions. *New Biol.*, 2(2):126–31.

Karplus, M. and McCammon, J. A. (2002). Molecular dynamics simulations of biomolecules. *Nature structural biology*, 9(9):646–652.

Laity, J. H. (2006). *Handbook of metalloproteins*. John Wiley & Sons, Ltd.

Latchman, D. (1997). Transcription factors: an overview. *Int. J. Biochem. Cell Biol.*, 29(12):1305–12.

Lee, L. (1999). Measures of distributional similarity. In *In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 25–32.

Liu, Q., Xia, Z., and Case, C. C. (2002). Validated Zinc Finger Protein Designs for All 16 GNN DNA Triplet Targets. *J. Biol. Chem.*, 277(6):3850–3856.

Lowe, D. and Tipping, M. E. (1997). NeuroScale: Novel topographic feature extraction using RBF networks.

McCall, K. A., Huang, C.-c., and Fierke, C. A. (2000). Function and Mechanism of Zinc Metalloenzymes. *The Journal of Nutrition*, 130:1437S–1446S.

McCammon, J. A., Gelin, B., and Karplus, M. (1977). Dynamics of folded proteins. *Nature*, 267:585–590.

McCulloch, W. and Pitts, W. (1943). A Logical Calculus of Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics*, 5(4):115–133.

McNaught, A. D. and Wilkinson, A., editors (1997). *Compendium of Chemical Terminology*. Blackwell Scientific Publications, Oxford.

Meinguet, J. (1979). Multivariate interpolation at arbitrary points made simple. *Journal of Applied Mathematics and Physics*, 61:19–28.

Mendez, R., Leplae, R., Maria, L. D., and Wodak, S. J. (2003). Assessment of Blind Predictions of Protein-Protein Interactions: Current Status of Docking Methods. *PROTEINS: Structure, Function, and Genetics*, 52:51–57.

Meng, E., Shoichet, B., and Kuntz, I. (2004). Automated docking with grid-based energy evaluation. *Journal of Computational Chemistry.*, 13(4):505–524.

Morozov, A. V., Havranek, J. J., Baker, D., and Siggia, E. D. (2005). Protein-DNA binding specificity predictions with structural model. *Nucleic Acids Resarch*, 33:5781–5798.

Morris, G., Goodsell, D., Halliday, R., Huey, R., Hart, W., Belew, R., and Olson, A. (1998). Automated docking using a lamarckian gengene algorithm and an empirical binding free energy function. *Journal of Computational Chemistry*, 19(14):1639–1662.

MS, L., GP, G., KV, S., DA, C., and PE, W. (1989). Three-dimensional solution structure of a single zinc finger dna-binding domain. *Science*, 245(4918):635–7.

Nabney, I. T. (2002). *Netlab Algorithms for Pattern Recognition*. Springer Science+Business Media, LLC, USA.

Nakata, K. (1995). Prediction of zinc finger DNA binding protein. *Computer applications in the biosciences : CABIOS*, 11:125–131.

Nardelli, J., Gibson, T., and Charnay, P. (1992). Zinc finger-DNA recognition: analysis of base specificity by site-directed mutagenesis. *Nucleic Acids Res*, 20(16):4137–4144.

Passerini, A., Andreini, C., Menchetti, S., Rosato, A., and Frasconi, P. (2007). Predicting zinc binding at the proteome level. *BMC Bioinformatics*, 8:39.

Pavletich, N. P. and Pabo, C. O. (1991). Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. *Science*, 252(5007):809–17.

Pearson, K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*, 2:559–572.

Persikov, A. V., Osada, R., and Singh, M. (2008). Predicting DNA recognition by Cys2His2 zinc finger proteins. *Bioinformatics*, 25, Issue1:22–29.

Petkov, V. (2010). *Minkowski Spacetime: A Hundred Years Later*. Springer.

Puzyn, T., Leszczynski, J., and Cronin, M. T., editors (2010). *Recent Advances in QSAR Studies: Methods and Applications*. Springer Dordrecht Heidelberg London New York.

Rahman, A. (1964). Correlations in the Motion of Atoms in Liquid Argon. *Phys Rev*, 136(2A):A405A411.

Rebar, E. J. and Pabo, C. O. (1994). Zinc Finger Phage: Affinity Selection of Fingers with New DNA-Binding Specificities. *Science*, 263(5147):671–3.

Reynolds, L., Ullman, C., Moore, M., Isalan, M., West, M. J., Clapham, P., Klug, A., and Choo, Y. (2003). Repression of the HIV-1 5' LTR promoter and inhibition of HIV-1 replication by using engineered zinc-finger transcription factors. *Proc Natl Acad Sci U S A*, 100(4):1615–1620.

Roweis, S. and Saul., L. (2000). Nonlinear dimensionality reduction by Locally linear embedding. *Science*, 290(5500):2323–2326.

Rumelhart, D., Hinton, G., and Williams., R. (1986). *Learning internal representation by error propagation*. MIT Press.

Sammon, J. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, 18:401–409.

Saul, L. and Roweis, S. (2003). Think globally, fit locally: Unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research*, 4:119–155.

Segal, D. J., Dreier, B., Beerli, R. R., and Barbas, C. F. (1999). Toward controlling gene expression at will: Selection and design of zinc finger domains recognizing each of the 5'-GNN-3' DNA target sequences. *Biological Sciences - Biochemistry*, 96(6):2758–2763.

Shastry, B. (1996). Transcription factor IIIA (TFIIIA) in the second decade. *J. Cell. Sci.*, 109:535–39.

Shu, N., Zhou, T., and HovmÃűller, S. (2008). Prediction of zinc-binding sites in proteins from sequence. *Bioinformatics*, 24:775–782.

Siggers, T. W. and Honig, B. (2007). Structure-based prediction of C2H2 zinc-finger binding specificity: sensitivity to docking geometry. *Nucleic Acids Research*, 35:1085–1097.

Silva, V. D. and Tenenbaum, J. B. (2003). Global versus local methods in nonlinear dimensionality reduction. In *Advances in Neural Information Processing Systems 15*, pages 705–712. MIT Press.

Sims, G. and Sommers (1985). Degradation of pyridine derivatives in soil. *J. Environmental Quality.*, 14:580–584.

Sivaraksa, M. (2008). *Uncertainty and Topographic Visualisations*. PhD thesis, Aston University.

Sivaraksa, M. and Lowe, D. (2008). Predictive gene lists for breast cancer prognosis: A topographic visualisation study. *BMC Med Genomics.*, 1:1–8.

Stormo, G. D., Schneider, T. D., Gold, L., and Ehrenfeucht, A. (1982). Use of the 'Perceptro' algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Research*, 10(9):2997–3011.

Sun, J. (2011). *Extending metric multidimensional scaling with bregman divergences*. PhD thesis, University of the West Scotland.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, 240:1285–1293.

Tachikawa, K. and Briggs., S. P. (2006). Tartarget the human genome. *Current Opinion in Biotechnology*, 17:659–665.

Tenenbaum, J., de Silva, V., and Langford, J. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323.

Thiesen, H.-j. and Bach, C. (1990). Target Detection Assay (TDA): a versatile procedure to determine DNA blinding sites as demonstrated on SP1 protein. *Nucleic Acids Res*, 18(11):3203–3209.

Tipping, M., House, S. G., Street, G., and Nh, C. C. (1999). Probabilistic Visualisation of High-dimensional Binary Data.

Tipping, M. E. (1996). *Topographic Mappings and Feed-Forward Neural Networks*. PhD thesis, Aston University.

Tipping, M. E. (2001). Sparse bayesian Learning and the relevance vector machine. *Journal of Machine Learning Research*, pages 211–244.

Tsuchiya, Y., Kinoshita, K., and Nakamura, H. (2004). Structure-based prediction of DNA-binding sites on preteins using the empirical preference of electrostatic potential and the shape of molecular surfaces. *Proteins: Structure*, 55:885–894.

Vallee, B. L. and Auld, D. S. (1993). Cocatalytic zinc motifs in enzyme catalysis. *Proc. Natl. Acad. Sci.*, 90:2715–2718.

Webb, A. (1999). *Statistical pattern recognition*. Oxford University Press Inc.

Wingender, E., Chen, X., Fricke, E., Geffers, R., Hehl, R., Liebich, I., Krull, M., Matys, V., Michael, H., OhnhÃd'user, R., PrÃijÃ§, M., Schacherer, F., Thiele, S., and Urbach, S. (2001). The TRANSFAC system on gene expression regulation. *Nucleic Acids Research*, 29(1):281–283.

Wodak, S. and Janin, J. (1978). Computer analysis of protein-protein interaction. *J. Mol. Biol.*, 124:323–342.

Wold, S., Sjostrom, M., and Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2):109–130.

Wolfe, S. A., and, L. N., and Pabo, C. O. (June 2000). DNA recognition by Cys2His2 zinc finger proteins. *Annual Review of Biophysics and Biomolecular Structure*, 29:183–212.

Wolfe, S. A., Greisman, H. A., Ramm, E. I., and Pabo, C. O. (1999). Analysis of zinc fingers optimized via phage display: evaluating the utility of a recognition code. *Journal of Molecular Biology*, 285(5):1917–1934.

Wu, H., Yang, W. P., and Barbas, C. F. (1995). Building zinc fingers by selection: toward a therapeutic application. *Proc Natl Acad Sci U S A*, 92(2):344–348.

Yamasaki, S., Terada, T., Kono, H., Shimizu, K., and Sarai, A. (2012). A new method for evaluating the specificity of indirect readout in protein-dna recognition. *Nucleic Acids Res.*, 40:e129.

Zheng, M., Liu, Z., Xue, C., Zhu, W., Chen, K., Luo, X., and Jiang, H. (2006). Mutagenic probability estimation of chemical compounds by a novel molecular electrophilicity vector and support vector machine. *Bioinformatics*, 22:2099–2106.

A Dataset

A.1 Published data list

Data samples which consist of the 26 data sources contain five parts: paper title of the source, DNA sequence, number of binding zinc fingers, the amino acids with the correct positions [-1, 2, 3, 6] in zinc finger regions with the pattern $CX_{2-6}CX_{12}HX_{2-6}H$, and a dissociation constant K_d [1] which is used to identify the binding affinity.

Table A.1 lists 26 published data sources.

[1] A dissociation constant K_d is a specific type of equilibrium constant that measures the propensity of a larger object to separate reversibly into smaller components. Units of K_d is nM, when K_d < 200 nM, the data sample can be considered as a positive binding example.

Number	Source	Reference
1	DBSFB01	Dreier et al. (2001)
2	DFSLBHKB05	Dreier et al. (2005)
3	SDBB99	Segal et al. (1999)
4	DSB00	Dreier et al. (2000)
5	BMB03	Blancafort et al. (2003)
6	BFS02	Benos et al. (2002)
7	BJC02	Bulyk et al. (2002)
8	BKSHRP03	Bae et al. (2003)
9	CGU99	Cook et al. (1999)
10	CK94a	Choo and Klug (1994b)
11	CK94b	Choo and Klug (1994a)
12	DB92	Desjarlais and Berg (1992)
13	DB93	Desjarlais and Berg (1993)
14	GP97	Greisman and Pabo (1997)
15	ICK97	Isalan et al. (1997)
16	IKC01	Isalan et al. (2001)
17	JKW94	Jamieson et al. (1994)
18	KFM05	Kaplan et al. (2005)
19	LXC02	Liu et al. (2002)
20	NGC92	Nardelli et al. (1992)
21	PDB	Berman et al. (2000)
22	RP94	Rebar and Pabo (1994)
23	RUMIWCKC03	Reynolds et al. (2003)
24	TB90	Thiesen and Bach (1990)
25	WGRP99	Wolfe et al. (1999)
26	WYB95	Wu et al. (1995)

Table A.1: Cited published sources list.

A.2 Structure of original published data

Table A.2 represents the structures information of data samples in the original database.

Source	DNA	No. of zf.	f1	f2	f3	ex	Kd
DBSFB01	ctcgcgGGGgcggcc	3	KSADLKRHIRI	RSDHLTTHIRT	RSDERKRHTKI	Kd	0.5
CK94a	tatatagcgGTGgcgtatata	3	RSDELTRHIRI	REDVLIRHGKT	RSDERKRHTKI	+	—
BJC02	tatatagcgTTGgcgtatata	3	RSDELTRHIRI	KASNLVSHIRT	RSDERKRHTKI	Kd	0.001106475
DSB00	cccgcgGCCgcgtcc	3	KSADLKRHIRI	QSSNLVRHIRT	RSDERKRHTKI	-	—
SDBB99	cccgcgGGGgcgtcc	3	KSADLKRHIRI	RSDKLVRHIRT	RSDERKRHTKI	Kd	6.0
BFS02	-gcgtgggagt	3	rsdeltrhir	rsdhlthir	rsderkrhtk	+	—
KFM05	-GCGTGGGTG-	3	RSDELTRHIRI	SDHLTTHIRTT	RSDERKRHTKI	+	—
JKW94	gatccgcgtggGTTctgca	3	ESRALTRHIRI	RSDHLTTHIRT	RSDERKRHTKI	Kd	2.1
WYB95	cctgcgtggTGTccc	3	RSDELTRHIRI	RSDHLTTHIRT	RSDERKRHTKI	Kd	81.8
NGC92	gtacgcgAGGgcggtta	3	RSDELTRHIRI	QSSHLTRHIRT	RSDERKRHTKI	Kd	>20
PDB	agcgtgggacc	3	DSSNLTR	RSDHLTT	RSDERKR	+	—
WGRP99	—ggctataaaag—	3	QKTNLDTHIRI	QQASLNAHIRT	TLHTRTRHTKI	Kd	120
...	...	...	...	...	...	...	...

Table A.2: This table summarises the source of the data as a citation reference, the DNA usded in the experiments that number of zinc fingers considered, followed by each zinc finger, then followed by its binding status if known ('+' = binding, '-' = nonbinding, K_d= a K_d value is provided which is given in the final column).

A.3 List of published data

In this section, the published data sources are divided into three groups according to the structure information provided. Table A.3 lists all data sources which only provide the information of the 2nd zinc finger. In Table A.4, the data sources which include the information of three zinc fingers are listed. Table A.5 is the list of the data sources which have compared DNA sequences without binding status information.

Source	**No. of Finger**	**No. of Samples**	**No. of Selected Samples**	**Ref.**
DBSFB01	1	322	264	1
DFSLBHKB05	1	220	214	2
SDBB99	1	201	43	3
DSB00	1	458	326	4
BJC02	1	320	288	7
BKSHRP03	1	33	31	8
CK94a	1	67	0	10
CK94b	1	19	11	11
ICK97	1	4	1	15
JKW94	1	18	18	17
NGC92	1	140	110	20
RP94	1	12	10	22
WYB95	1	26	20	26
Total	1336			

Table A.3: List1 (one zinc finger selected)

Source	No. of Finger	No. of Samples	No. of Selected Samples	Ref.
BMB03	3	44×3	52	5
BFS02	3	1005×3	385	6
CGU99	3	5×3	0	9
DB93	3	9×3	6	13
GP97	3	24×3	9	14
IKC01	3	7×3	0	16
KFM05	3	40×3	13	18
LXC02	3	32×3	18	19
PDB	3	14×3	5	21
RUMIWCKC03	3	8×3	19	23
TB90	3	11×3	14	24
WGRP99	3	6×3	3	25
Total	524			

Table A.4: List 2(three zinc finger selected)

Source	No. of used zf.	No. of Samples	No. of selected samples	Ref.
DBSFB01	1	124	77	1
DFSLBHKB05	1	220	197	2
DSB00	1	410	247	4
SDBB99	1	320	45	3
DB92	1	12	0	12
IKC01	3	50×3	107	16
Total	673			

Table A.5: List 3 (Comparing data)

A.4 Database generation explain

The original data samples which represent the ZF-DNA binding interaction can be understood with respect to the 'canonical structural model' where each zinc finger contacts DNA in an antiparallel manner Pavletich and Pabo (1991); Elrod-Erickson et al. (1996). Once the data sources are determined, the interacting bases in the DNA sequences can be selected and stored in the 3'-5' order in the database of the sorted data sources, as shown in Table A.6. Also included in the database are the specific amino acid labels at position 2, -1, 3 and 6, their binding status and the K_d value where it has been reported.

No.	DNA (3'-5')	a_2	a_{-1}	a_3	a_6	Binding	K_d
253	gGCG	D	R	H	T	—	206.7
1275	gTAG	G	T	N	R	'+'	—
554	gACA	G	T	N	V	'-'	—
1143	cCCG	D	R	E	R	—	3
1245	gGGG	D	R	Q	R	—	<2.5
⋮	⋮		⋮	⋮	⋮	⋮	⋮

Table A.6: Examples of categorised data samples. In this table, the order of the DNA sequence is arranged from 5'-3' to 3'-5'. The amino acid at the contacting positions 2, -1, 3, 6 are retained. The number in the first column can be used to reference back to the original experiment. For example, DNA sequence gTAG in the second row was GATg before rearranged. Amino acids G, T, N and R would bind the DNA sequence at position 2(GG), -1(TT), 3(AN) and 6(GR) separately. From the record of the experiment, the binding status is defined as binding (+).

Table A.7 illustrates the databases regarding the canonical structural model number.

No.	a_2	a_{-1}	a_3	a_6	K_d/Binding
253	23	135	267	297	206.7
1275	26	157	172	295	+
554	26	97	192	258	-
1143	43	115	184	295	3
1245	23	135	214	295	<2.5
⋮		⋮	⋮	⋮	⋮

Table A.7: Converted data source with reference number. In this table, the binding pairs at 2, -1, 3 and 6 position are represented by a number according to the concept of the canonical structural model. Using the model numbers, the 320-Dimensional vector for each data sample can be created as shown in Table 3.2. Using a binding pair '01gD' in the first row in Table A.6 as an example, '01' indicates that the index of this binding pair should be between 1 and 80. Considering the base '*g*' is on the primary DNA strand, the interaction happens on the complementary strand, '*g*' should be replaced by '*c*' when define the index of this binding pair. According to Table A.9 in Appendix A, the order of nucleotides was defined as: A, C, G, T. Therefore, there should be 20 possibilities for each nucleotide by binding with 20 amino acids. Moreover, since 'D' was numbered as the 3^{rd} amino acids in this study, the index of '01gD' was defined as '23' finally.

Using Table A.9 as a look-up dictionary, it is quite straightforward to find the binding pairs between nucleotides and amino acids in different positions by checking the indices of non-zero elements in these vectors. Table 3.2 in subsection 3.2.2 shows an example where the value of '1' appear at locations 23, 135, 267 and 297. Table A.9 equates the a_{-1} position to nucleotide '*g*' and amino acid '*R*'.

Table A.8 shows the transformed database that integrates all the information of each binding pair. In the table, the original number corresponds to the location of each pair in the database illustrated in Table A.7. For example, '253' refers to the 253^{rd} row in Table A.7. This makes it possible to track the original citation of the datum. The order of the DNA sequence in this database is reversed back to 5'-3'. The order of the amino acids is set to be a_6, a_3, a_{-1} and a_2 for reading convenience. K_d and the binding status are merged into one column, which can also be used for information checking, and helps understand the distribution of data samples. For example, in a data visualization process, each data sample was plotted following the order which is shown in the 'Current No.' column, and labelled based on the value in 'K_d/Binding' column. Then, if any relative information of selected points is requested to be shown, by linking the 'Current No.' columns, the

corresponding information can be found in DNA and amino acid columns.

Original No.	Current No.	DNA (5'-3')	a_6, a_3, a_{-1}, a_2	K_d/Binding
070013	253	GCGg	T, H, R, D	206.7
031784	1275	GATg	R, N, T, G	+
010506	554	ACAg	V, N, T, G	-
171449	1143	GCCc	R, E, R, D	3
201516	1245	GGGg	R, Q, R, D	<2.5
⋮	⋮	⋮	⋮	⋮

Table A.8: Binding status and related information database. The first column represents the address of the selected data sample in the original citation. The first two numbers represent the location in the list of data sources used. The last four numbers denotes the index of the selected data sample in the data source. For example, based on '171449' in the fourth row, the index of this original data sample is 1449^{th} which could be found from the 17^{th}citation.

Table A.9 defines all possible binding pairs at each binding position, which is created base on the canonical binding model.

01 - between amino acid $a_2$and nucleiotide b_4'
02 - between amino acid a_{-1} and nucleotide b_1
03 - between amino acid a_3 and nucleotide b_2
04 - between amino acid a_6 and nucleotide b_3

Index	a_2-b_4'	**Index**	a_{-1}-b_1	**Index**	a_3-b_2	**Index**	a_6-b_3
1	01aA	81	02aA	161	03aA	241	04aA
2	01aC	82	02aC	162	03aC	242	04aC
3	01aD	83	02aD	163	03aD	243	04aD
4	01aE	84	02aE	164	03aE	244	04aE
5	01aF	85	02aF	165	03aF	245	04aF
6	01aG	86	02aG	166	03aG	246	04aG
7	01aH	87	02aH	167	03aH	247	04aH
8	01aI	88	02aI	168	03aI	248	04aI
9	01aK	89	02aK	169	03aK	249	04aK
10	01aL	90	02aL	170	03aL	250	04aL

11	01aM	91	02aM	171	03aM	251	04aM
12	01aN	92	02aN	172	03aN	252	04aN
13	01aP	93	02aP	173	03aP	253	04aP
14	01aQ	94	02aQ	174	03aQ	254	04aQ
15	01aR	95	02aR	175	03aR	255	04aR
16	01aS	96	02aS	176	03aS	256	04aS
17	01aT	97	02aT	177	03aT	257	04aT
18	01aV	98	02aV	178	03aV	258	04aV
19	01aW	99	02aW	179	03aW	259	04aW
20	01aY	100	02aY	180	03aY	260	04aY
21	01cA	101	02cA	181	03cA	261	04cA
22	01cC	102	02cC	182	03cC	262	04cC
23	01cD	103	02cD	183	03cD	263	04cD
24	01cE	104	02cE	184	03cE	264	04cE
25	01cF	105	02cF	185	03cF	265	04cF
26	01cG	106	02cG	186	03cG	266	04cG
27	01cH	107	02cH	187	03cH	267	04cH
28	01cI	108	02cI	188	03cI	268	04cI
29	01cK	109	02cK	189	03cK	269	04cK
30	01cL	110	02cL	190	03cL	270	04cL
31	01cM	111	02cM	191	03cM	271	04cM
32	01cN	112	02cN	192	03cN	272	04cN
33	01cP	113	02cP	193	03cP	273	04cP
34	01cQ	114	02cQ	194	03cQ	274	04cQ
35	01cR	115	02cR	195	03cR	275	04cR
36	01cS	116	02cS	196	03cS	276	04cS
37	01cT	117	02cT	197	03cT	277	04cT
38	01cV	118	02cV	198	03cV	278	04cV

39	01cW	119	02cW	199	03cW	279	04cW
40	01cY	120	02cY	200	03cY	280	04cY
41	01gA	121	02gA	201	03gA	281	04gA
42	01gC	122	02gC	202	03gC	282	04gC
43	01gD	123	02gD	203	03gD	283	04gD
44	01gE	124	02gE	204	03gE	284	04gE
45	01gF	125	02gF	205	03gF	285	04gF
46	01gG	126	02gG	206	03gG	286	04gG
47	01gH	127	02gH	207	03gH	287	04gH
48	01gI	128	02gI	208	03gI	288	04gI
49	01gK	129	02gK	209	03gK	289	04gK
50	01gL	130	02gL	210	03gL	290	04gL
51	01gM	131	02gM	211	03gM	291	04gM
52	01gN	132	02gN	212	03gN	292	04gN
53	01gP	133	02gP	213	03gP	293	04gP
54	01gQ	134	02gQ	214	03gQ	294	04gQ
55	01gR	135	02gR	215	03gR	295	04gR
56	01gS	136	02gS	216	03gS	296	04gS
57	01gT	137	02gT	217	03gT	297	04gT
58	01gV	138	02gV	218	03gV	298	04gV
59	01gW	139	02gW	219	03gW	299	04gW
60	01gY	140	02gY	220	03gY	300	04gY
61	01tA	141	02tA	221	03tA	301	04tA
62	01tC	142	02tC	222	03tC	302	04tC
63	01tD	143	02tD	223	03tD	303	04tD
64	01tE	144	02tE	224	03tE	304	04tE
65	01tF	145	02tF	225	03tF	305	04tF
66	01tG	146	02tG	226	03tG	306	04tG

67	01tH	147	02tH	227	03tH	307	04tH
68	01tI	148	02tI	228	03tI	308	04tI
69	01tK	149	02tK	229	03tK	309	04tK
70	01tL	150	02tL	230	03tL	310	04tL
71	01tM	151	02tM	231	03tM	311	04tM
72	01tN	152	02tN	232	03tN	312	04tN
73	01tP	153	02tP	233	03tP	313	04tP
74	01tQ	154	02tQ	234	03tQ	314	04tQ
75	01tR	155	02tR	235	03tR	315	04tR
76	01tS	156	02tS	236	03tS	316	04tS
77	01tT	157	02tT	237	03tT	317	04tT
78	01tV	158	02tV	238	03tV	318	04tV
79	01tW	159	02tW	239	03tW	319	04tW
80	01tY	160	02tY	240	03tY	320	04tY

Table A.9: Reference vector information

Table A.10 lists the 1860 example included in the DB1 database. In the table, each example consists of the bases information in the DNA sequence, the amino acids in the zinc finger and the quantitative information for the binding affinity.

No.	**DNA**	**Amino acid**	**binding affinity**	**No.**	**DNA**	**Amino acid**	**binding affinity**
1	AAGg	DRHT	194.2505	931	GATg	GQHR	-
2	TGAg	DRHT	197.2296	932	GACg	GQHR	-
3	AAAg	DRHT	202.1278	933	GTGg	GQHR	-
4	AGAg	DRHT	222.4485	934	GTAg	GQHR	-
5	CCAg	DRHT	>250.1176	935	GTTg	GQHR	-
6	ATAg	DRHT	>250.1325	936	GTCg	GQHR	-

7	TAAg	DRHT	>250.1717	937	GCGg	GQHR	-
8	CAAg	DRHT	>250.2686	938	GCAg	GQHR	-
9	AACg	DRHT	>250.3442	939	GGGg	GQNR	-
10	ATCg	DRHT	>250.3446	940	GGAg	GQNR	-
11	TCAg	DRHT	>250.4797	941	GGTg	GQNR	-
12	CATg	DRHT	>250.6928	942	GGCg	GQNR	-
13	ACAg	DRHT	>250.6975	943	GTGg	GQNR	-
14	TCCg	DRHT	>250.7384	944	GTAg	GQNR	-
15	CGAg	DRHT	>250.7482	945	GTTg	GQNR	-
16	CACg	DRHT	>250.7573	946	GTCg	GQNR	-
17	ATGg	DRHT	>250.8497	947	GCGg	GQNR	-
18	TGCg	DRHT	>250.8562	948	GCAg	GQNR	-
19	CTCg	DRHT	>250.8937	949	GCCg	GQNR	-
20	AGCg	DRHT	>250.9051	950	GAAg	GDNR	-
21	CGTg	DRHT	>250.9306	951	GGGg	GDAR	-
22	AATg	DRHT	>250.9499	952	GGTg	GDAR	-
23	TGTg	DRHT	>250.9651	953	GGCg	GDAR	-
24	ATTg	DRHT	>251.0043	954	GAGg	GDAR	-
25	ACCg	DRHT	>251.0085	955	GATg	GDAR	-
26	CTGg	DRHT	>251.0416	956	GACg	GDAR	-
27	ACTg	DRHT	>251.0466	957	GTGg	GDAR	-
28	AGTg	DRHT	>251.1115	958	GCAg	GDHR	-
29	CCTg	DRHT	>251.1308	959	GCTg	GDHR	-
30	CTTg	DRHT	>251.1341	960	GAGg	SEKR	-
31	CTAg	DRHT	>251.1786	961	GTTg	SEKR	-
32	TATg	HLNT	6.3953	962	GCAg	SEKR	-
33	GATg	HLNT	71.7559	963	GCTg	SEKR	-
34	TGTg	HLNT	121.1643	964	GCCg	SEKR	-

35	CGTg	HLNT	158.8128	965	GCGg	SDKR	-
36	TACg	HLNT	178.8053	966	GTGg	GDHR	-
37	TAGg	HLNT	192.0235	967	GAAg	GDER	-
38	ATTg	HLNT	215.4926	968	GTGg	GDER	-
39	GGTg	HLNT	243.5163	969	GTAg	GDER	-
40	AACg	HLNT	>429.2377	970	GCGg	GDER	-
41	AAGg	HLNT	>429.2377	971	GGGg	RDDR	-
42	ACAg	HLNT	>429.2377	972	GGAg	RDDR	-
43	ACCg	HLNT	>429.2377	973	GGTg	RDDR	-
44	ACGg	HLNT	>429.2377	974	GGCg	RDDR	-
45	AGGg	HLNT	>429.2377	975	GAGg	RDDR	-
46	ATAg	HLNT	>429.2377	976	GATg	RDDR	-
47	ATCg	HLNT	>429.2377	977	GACg	RDDR	-
48	ATGg	HLNT	>429.2377	978	GTAg	RDDR	-
49	CAAg	HLNT	>429.2377	979	GTCg	RDDR	-
50	CACg	HLNT	>429.2377	980	GCGg	RDDR	-
51	CAGg	HLNT	>429.2377	981	GCAg	RDDR	-
52	CCAg	HLNT	>429.2377	982	GGGg	RGER	-
53	CCCg	HLNT	>429.2377	983	GAAg	RGER	-
54	CCGg	HLNT	>429.2377	984	GCGg	RGER	-
55	CGCg	HLNT	>429.2377	985	GGAg	GTNR	-
56	CGGg	HLNT	>429.2377	986	GGTg	GTNR	-
57	CTAg	HLNT	>429.2377	987	GGCg	GTNR	-
58	CTCg	HLNT	>429.2377	988	GACg	GTNR	-
59	GAAg	HLNT	>429.2377	989	GTGg	GTNR	-
60	GACg	HLNT	>429.2377	990	GCCg	GTNR	-
61	GACg	HLNT	>429.2377	991	GACg	GTER	-
62	GAGg	HLNT	>429.2377	992	GCGg	GTSR	-

63	GCAg	HLNT	>429.2377	993	GGGg	GTDR	-
64	GCCg	HLNT	>429.2377	994	GGAg	GTDR	-
65	GCGg	HLNT	>429.2377	995	GAAg	GTDR	-
66	GCTg	HLNT	>429.2377	996	GTAg	GTDR	-
67	GGAg	HLNT	>429.2377	997	GTTg	GTDR	-
68	GGCg	HLNT	>429.2377	998	GTCg	GTDR	-
69	GGGg	HLNT	>429.2377	999	GCGg	QTTR	-
70	GTAg	HLNT	>429.2377	1000	GCGg	ARKR	-
71	GTCg	HLNT	>429.2377	1001	GCAg	ARKR	-
72	GTGg	HLNT	>429.2377	1002	GCGg	AQKR	-
73	GTTg	HLNT	>429.2377	1003	GCGg	GTHR	-
74	TAAg	HLNT	>429.2377	1004	GCGg	GTKR	-
75	TCAg	HLNT	>429.2377	1005	GTGg	DTHR	-
76	TCCg	HLNT	>429.2377	1006	GCGg	DTHR	-
77	TCGg	HLNT	>429.2377	1007	GGGg	SQNR	-
78	TCTg	HLNT	>429.2377	1008	GGTg	GTSR	+
79	TGAg	HLNT	>429.2377	1009	GTTg	DRKR	-
80	TGCg	HLNT	>429.2377	1010	GCGg	SQNR	-
81	TGGg	HLNT	>429.2377	1011	GGAg	GQHR	+
82	TTAg	HLNT	>429.2377	1012	GCCg	RGER	+
83	TTGg	HLNT	>429.2377	1013	GAAg	DRKR	-
84	CCGg	PRDR	325.7161	1014	GGAg	DRKR	-
85	GCAg	PRDR	515.1555	1015	GATg	DRKR	-
86	GCCg	PRDR	528.3086	1016	GACg	DRKR	-
87	GAGg	PRDR	672.2521	1017	GCGg	DRKR	-
88	TCTg	PRDR	688.0570	1018	GCAg	DRKR	-
89	ACGg	PRDR	698.5050	1019	GCTg	DRNR	-
90	ATGg	PRDR	720.4991	1020	GCCg	DRER	-

91	TTTg	PRDR	748.0965	1021	GGCg	SQNR	-
92	CTCg	PRDR	755.8037	1022	GTGg	SQNR	-
93	CCCg	PRDR	764.1514	1023	GTTg	SQNR	-
94	GGGg	PRDR	768.9106	1024	GCAg	AQHR	-
95	CAGg	PRDR	810.1641	1025	GGCg	SEKR	+
96	AGGg	PRDR	826.8736	1026	GGGg	DTKR	+
97	ACAg	PRDR	831.6538	1027	GCTg	QSTR	+
98	AAAg	PRDR	>842.9475	1028	GCCg	DRHR	-
99	AACg	PRDR	>842.9475	1029	GTAg	DRKR	-
100	AAGg	PRDR	>842.9475	1030	GTCg	DRKR	-
101	AATg	PRDR	>842.9475	1031	GCCg	DRKR	-
102	ACTg	PRDR	>842.9475	1032	GGTg	SQNR	-
103	AGAg	PRDR	>842.9475	1033	GTCg	SQNR	-
104	AGCg	PRDR	>842.9475	1034	GACg	AQHR	-
105	AGTg	PRDR	>842.9475	1035	GTGg	AQHR	-
106	ATAg	PRDR	>842.9475	1036	GTCg	AQHR	-
107	ATCg	PRDR	>842.9475	1037	GCGg	AQHR	-
108	ATTg	PRDR	>842.9475	1038	GCTg	AQHR	-
109	CAAg	PRDR	>842.9475	1039	GCCg	AQHR	-
110	CACg	PRDR	>842.9475	1040	GCGg	SQHR	-
111	CATg	PRDR	>842.9475	1041	GGCg	GQDR	-
112	CCAg	PRDR	>842.9475	1042	GTCg	GQDR	-
113	CCTg	PRDR	>842.9475	1043	GCGg	GQDR	-
114	CGAg	PRDR	>842.9475	1044	GTGg	GDNR	-
115	CGCg	PRDR	>842.9475	1045	GCGg	GDNR	-
116	CGGg	PRDR	>842.9475	1046	GCAg	GDNR	-
117	CGTg	PRDR	>842.9475	1047	GCGg	GDHR	-
118	CTAg	PRDR	>842.9475	1048	GCGg	GTNR	-

119	CTGg	PRDR	>842.9475	1049	GAAg	DTKR	-
120	CTTg	PRDR	>842.9475	1050	GTGg	DTKR	-
121	GAAg	PRDR	>842.9475	1051	GTTg	DTKR	-
122	GACg	PRDR	>842.9475	1052	GCGg	DTKR	-
123	GATg	PRDR	>842.9475	1053	GCAg	DTKR	-
124	GGAg	PRDR	>842.9475	1054	GCTg	DTKR	-
125	GGCg	PRDR	>842.9475	1055	GCCg	DTKR	-
126	GGTg	PRDR	>842.9475	1056	GAGg	DRSR	+
127	GTAg	PRDR	>842.9475	1057	GAGg	DRTR	+
128	GTCg	PRDR	>842.9475	1058	GCCg	AKER	+
129	GTTg	PRDR	>842.9475	1059	GAAg	DRHR	-
130	TAAg	PRDR	>842.9475	1060	GATg	DRHR	-
131	TACg	PRDR	>842.9475	1061	GACg	DRHR	-
132	TAGg	PRDR	>842.9475	1062	GTAg	DRHR	-
133	TATg	PRDR	>842.9475	1063	GTTg	DRHR	-
134	TCAg	PRDR	>842.9475	1064	GCAg	DRHR	-
135	TCCg	PRDR	>842.9475	1065	GCTg	DRHR	-
136	TCGg	PRDR	>842.9475	1066	GCTg	DRKR	-
137	TGAg	PRDR	>842.9475	1067	GTCg	DRNR	-
138	TGCg	PRDR	>842.9475	1068	GGAg	DRNR	-
139	TGGg	PRDR	>842.9475	1069	GGTg	DRNR	-
140	TGTg	PRDR	>842.9475	1070	GGCg	DRNR	-
141	TTAg	PRDR	>842.9475	1071	GTTg	DRNR	-
142	TTCg	PRDR	>842.9475	1072	GCAg	DRNR	-
143	TTGg	PRDR	>842.9475	1073	GCCg	DRNR	-
144	ACCg	PRDR	>866.6677	1074	GGAg	DRDR	-
145	GAGg	DRVR	794.8918	1075	GGTg	DRDR	-
146	CCGg	DRVR	898.2653	1076	GGCg	DRDR	-

147	GGGg	DRVR	1248.9258	1077	GAAg	DRDR	-
148	GCCg	DRVR	1595.0438	1078	GATg	DRDR	-
149	TCTg	DRVR	1608.0865	1079	GACg	DRDR	-
150	ATGg	DRVR	1690.9878	1080	GGAg	DRTR	-
151	ACGg	DRVR	1831.6623	1081	GAAg	DRTR	-
152	CCCg	DRVR	1842.2048	1082	GATg	DRTR	-
153	ATAg	DRVR	1843.9296	1083	GACg	DRTR	-
154	CTCg	DRVR	1888.1037	1084	GTTg	DRTR	-
155	CAGg	DRVR	1902.3728	1085	GGAg	AKER	-
156	AAAg	DRVR	>2535.9460	1086	GAAg	AKER	-
157	AACg	DRVR	>2535.9460	1087	GCGg	AKER	-
158	AAGg	DRVR	>2535.9460	1088	GCCg	SQNR	-
159	AATg	DRVR	>2535.9460	1089	GGAg	SQNR	-
160	ACAg	DRVR	>2535.9460	1090	GACg	SQSR	-
161	ACCg	DRVR	>2535.9460	1091	GCGg	SQSR	-
162	ACTg	DRVR	>2535.9460	1092	GAGg	AQHR	-
163	AGAg	DRVR	>2535.9460	1093	GTTg	AQHR	-
164	AGCg	DRVR	>2535.9460	1094	GTGg	SQHR	-
165	AGGg	DRVR	>2535.9460	1095	GGGg	GQDR	-
166	AGTg	DRVR	>2535.9460	1096	GGTg	GQDR	-
167	ATCg	DRVR	>2535.9460	1097	GACg	AKER	-
168	ATTg	DRVR	>2535.9460	1098	GTGg	GQDR	-
169	CAAg	DRVR	>2535.9460	1099	GTAg	GQDR	-
170	CACg	DRVR	>2535.9460	1100	GTTg	GQDR	-
171	CATg	DRVR	>2535.9460	1101	GTAg	GDNR	-
172	CCAg	DRVR	>2535.9460	1102	GTTg	GDNR	-
173	CCTg	DRVR	>2535.9460	1103	GTCg	GDNR	-
174	CGAg	DRVR	>2535.9460	1104	GGGg	GDNR	-

175	CGCg	DRVR	>2535.9460	1105	GGAg	GDNR	-
176	CGGg	DRVR	>2535.9460	1106	GGTg	GDNR	-
177	CGTg	DRVR	>2535.9460	1107	GGCg	GDNR	-
178	CTAg	DRVR	>2535.9460	1108	GGAg	GDAR	-
179	CTGg	DRVR	>2535.9460	1109	GAAg	GDAR	-
180	CTTg	DRVR	>2535.9460	1110	GCGg	GDAR	-
181	TAAg	DRVR	>2535.9460	1111	GCAg	GDAR	-
182	TACg	DRVR	>2535.9460	1112	GAAg	GDHR	-
183	TAGg	DRVR	>2535.9460	1113	GTAg	GDHR	-
184	TATg	DRVR	>2535.9460	1114	GTTg	GDHR	-
185	TCAg	DRVR	>2535.9460	1115	GAAg	SEKR	-
186	TCCg	DRVR	>2535.9460	1116	GACg	SEKR	-
187	TCGg	DRVR	>2535.9460	1117	GTGg	SEKR	-
188	TGAg	DRVR	>2535.9460	1118	GTAg	SEKR	-
189	TGCg	DRVR	>2535.9460	1119	GCGg	SEKR	-
190	TGGg	DRVR	>2535.9460	1120	GAAg	RDDR	-
191	TGTg	DRVR	>2535.9460	1121	GTGg	RDDR	-
192	TTAg	DRVR	>2535.9460	1122	GGGg	GTNR	-
193	TTCg	DRVR	>2535.9460	1123	GTAg	GTNR	-
194	TTGg	DRVR	>2535.9460	1124	GTTg	GTNR	-
195	TTTg	DRVR	>2535.9460	1125	GTCg	GTNR	-
196	ATTg	SKNS	273.0889	1126	GCAg	GTNR	-
197	TATg	SKNS	289.6233	1127	GACg	DTKR	-
198	CGTg	SKNS	334.1258	1128	GGGg	GTER	-
199	TGTg	SKNS	360.3204	1129	GGAg	GTER	-
200	TAGg	SKNS	417.7175	1130	GGCg	GTER	-
201	CGCg	SKNS	420.8763	1131	GAGg	GTER	-
202	GAAg	SKNS	519.4723	1132	GAAg	GTER	-

203	GACg	SKNS	540.2045	1133	GCGg	GTER	-
204	GGCg	SKNS	572.9782	1134	GGAg	QSTR	-
205	CGGg	SKNS	577.8656	1135	GTGg	DRVR	1.3
206	GGGg	SKNS	585.3109	1136	GCGc	DRER	0.5
207	GGTg	SKNS	593.1785	1137	GAGc	DRER	2.8
208	GAGg	SKNS	645.5238	1138	GCAc	DRER	2.4
209	ATGg	SKNS	691.3845	1139	GCGc	DRDR	0.4
210	TTAg	SKNS	728.1302	1140	GTGc	REAR	0.6
211	ACCg	SKNS	744.6446	1141	GGGc	DRER	5.6
212	CACg	SKNS	754.9179	1142	GTGc	DRER	3.4
213	TACg	SKNS	789.5803	1143	GCCc	DRER	3
214	GCTg	SKNS	793.4383	1144	GCTc	DRER	3.7
215	GATg	SKNS	809.5039	1145	GAGc	DRDR	3
216	TCTg	SKNS	819.0844	1146	GGGc	DRDR	3.7
217	TCCg	SKNS	840.1106	1147	GTGc	DRDR	4
218	ACGg	SKNS	872.5159	1148	GAGc	REAR	1.5
219	TCAg	SKNS	882.3649	1149	GCGc	REAR	1.5
220	CAAg	SKNS	882.9749	1150	GGGc	REAR	1.8
221	TAAg	SKNS	910.2854	1151	GTAc	REAR	1.7
222	GTTg	SKNS	916.6901	1152	GTCc	REAR	1.8
223	TCGg	SKNS	938.3442	1153	GTTc	REAR	2.1
224	TTGg	SKNS	941.0204	1154	CGGg	DRER	>20
225	CTCg	SKNS	948.5505	1155	AGGg	DRER	>20
226	GTAg	SKNS	950.6004	1156	GGAg	DRER	>20
227	GTCg	SKNS	958.0037	1157	GGTg	DRER	>20
228	TGCg	SKNS	967.5087	1158	CGCg	DRER	>20
229	AGGg	SKNS	976.4161	1159	CCCg	DRER	>20
230	GCGg	SKNS	1001.5048	1160	CGGg	DRHR	>20

231	CCCg	SKNS	1004.2405	1161	GTGg	DRHR	>20
232	GCAg	SKNS	1031.4143	1162	GGCg	DRHR	>20
233	GTGg	SKNS	1038.0891	1163	GGAg	DRHR	>20
234	CAGg	SKNS	1050.7944	1164	GGTg	DRHR	>20
235	TGGg	SKNS	1069.5372	1165	CGCg	DRHR	>20
236	AAGg	SKNS	1087.5174	1166	CCCg	DRHR	>20
237	CTGg	SKNS	1115.3425	1167	GGGg	DRET	>20
238	GGAg	SKNS	1122.7580	1168	CGGg	DRET	>20
239	CTAg	SKNS	1127.0108	1169	AGGg	DRET	>20
240	GCCg	SKNS	1161.9639	1170	GGCg	DRET	>20
241	TGAg	SKNS	1170.0327	1171	GGAg	DRET	>20
242	CCGg	SKNS	1172.9459	1172	GGTg	DRET	>20
243	AACg	SKNS	1184.7393	1173	CCGg	DRET	>20
244	ATCg	SKNS	1207.4539	1174	CGCg	DRET	>20
245	ACAg	SKNS	1219.8222	1175	GCCg	DRET	>20
246	CCAg	SKNS	1247.5460	1176	CCCg	DRET	>20
247	ATAg	SKNS	1325.2193	1177	CGGg	DRQR	>20
248	TCGg	DRHT	175.1083	1178	AGGg	DRQR	>20
249	TAGg	DRHT	9.4150	1179	GGCg	DRQR	>20
250	CGGg	DRHT	38.4801	1180	GGAg	DRQR	>20
251	GCGg	PRDR	17.5549	1181	GGTg	DRQR	>20
252	AGGg	DRHT	52.6027	1182	CCGg	DRQR	>20
253	GCGg	DRHT	206.7075	1183	CGCg	DRQR	>20
254	GTGg	PRDR	608.5536	1184	GCCg	DRQR	>20
255	GAGg	DRHT	72.0834	1185	CCCg	DRQR	>20
256	CAGg	DRHT	188.0843	1186	CGGg	TRNR	>20
257	GAAg	DRHT	>250.3034	1187	AGGg	TRNR	>20
258	GCCg	DRHT	>250.6847	1188	TGGg	TRNR	>20

259	GTTg	DRHT	>250.7146	1189	GGCg	TRNR	>20
260	TTCg	DRHT	212.3228	1190	GGAg	TRNR	>20
261	GATg	DRHT	>250.0091	1191	GGTg	TRNR	>20
262	GTAg	DRHT	>250.3676	1192	CCGg	TRNR	>20
263	GCTg	DRHT	>250.3855	1193	CGCg	TRNR	>20
264	TTAg	DRHT	>250.4347	1194	GCCg	TRNR	>20
265	GACg	DRHT	>250.5304	1195	CCCg	TRNR	>20
266	TATg	DRHT	>250.5614	1196	CGGg	SQHR	>20
267	GTCg	DRHT	>250.5934	1197	AGGg	SQHR	>20
268	TCTg	DRHT	>250.7132	1198	TGGg	SQHR	>20
269	TACg	DRHT	>250.7840	1199	CCGg	SQHR	>20
270	TTTg	DRHT	>250.9452	1200	CGCg	SQHR	>20
271	CCCg	DRHT	>251.1349	1201	GCCg	SQHR	>20
272	CGCg	DRHT	>251.2141	1202	GGGg	SQHT	>20
273	AATg	HLNT	54.7124	1203	CGGg	SQHT	>20
274	AGTg	HLNT	96.9239	1204	AGGg	SQHT	>20
275	CATg	HLNT	110.6445	1205	GCGg	SQHT	>20
276	ACTg	HLNT	156.8135	1206	GAGg	SQHT	>20
277	CCTg	HLNT	190.3357	1207	GTGg	SQHT	>20
278	GCAg	DRVR	1373.9895	1208	GGCg	SQHT	>20
279	GCTg	DRVR	1673.2345	1209	GGAg	SQHT	>20
280	GAAg	DRVR	>2535.9460	1210	GGTg	SQHT	>20
281	GACg	DRVR	>2535.9460	1211	CCGg	SQHT	>20
282	GATg	DRVR	>2535.9460	1212	CGCg	SQHT	>20
283	GGAg	DRVR	>2535.9460	1213	GCCg	SQHT	>20
284	GGCg	DRVR	>2535.9460	1214	CCCg	SQHT	>20
285	GGTg	DRVR	>2535.9460	1215	GGGg	SEHT	>20
286	GTAg	DRVR	>2535.9460	1216	CGGg	SEHT	>20

287	GTCg	DRVR	>2535.9460	1217	AGGg	SEHT	>20
288	GTTg	DRVR	>2535.9460	1218	GCGg	SEHT	>20
289	GTCt	SDAR	0.021	1219	GAGg	SEHT	>20
290	GCCt	SDCR	0.22	1220	GTGg	SEHT	>20
291	GTTt	SHSR	0.043	1221	GGCg	SEHT	>20
292	GAGt	SKNR	0.094	1222	GGTg	SEHT	>20
293	GGAt	AQHR	0.47	1223	CCGg	SEHT	>20
294	GAGt	FQNR	3.7	1224	CGCg	SEHT	>20
295	GAAt	GQNR	0.069	1225	GCCg	SEHT	>20
296	GGAt	SQHR	0.11	1226	CCCg	SEHT	>20
297	CGAt	SQHV	3.6	1227	GGGg	SLHT	>20
298	CAAt	SQNI	2.9	1228	CGGg	SLHT	>20
299	GAAt	SQNK	0.15	1229	AGGg	SLHT	>20
300	GTAt	SQTR	0.051	1230	TGGg	SLHT	>20
301	GGAt	TQHR	0.089	1231	GCGg	SLHT	>20
302	GGGt	DRHR	0.049	1232	GAGg	SLHT	>20
303	GGGt	DRKR	0.25	1233	GTGg	SLHT	>20
304	GAGt	SSNR	0.075	1234	GGCg	SLHT	>20
305	GGTt	SWNR	0.073	1235	GGAg	SLHT	>20
306	GACt	SCNR	0.13	1236	GGTg	SLHT	>20
307	GACt	SHNK	10	1237	CCGg	SLHT	>20
308	GATt	SINR	0.0062	1238	CGCg	SLHT	>20
309	AGAt	SQHT	0.96	1239	GCCg	SLHT	>20
310	CAAt	SQNV	0.23	1240	CCCg	SLHT	>20
311	GTAt	SQSR	0.046	1241	GGGc	SQHR	<2.5
312	CGAt	TQHQ	0.034	1242	GAGg	DRER	<2.5
313	AGGt	DRHT	0.38	1243	GGTg	DRHT	5,20
314	GGGt	SRHR	0.01	1244	GAGg	DRHR	2.5,5

315	GAGt	SRNR	0.04	1245	GGGg	DRQR	<2.5
316	AATt	SVNV	0.14	1246	GCGg	DRQR	<2.5
317	GTGt	SVSR	0.12	1247	TGGg	DRER	5,20
318	GCTt	SVTR	0.81	1248	AGGg	DRHR	5,20
319	GCGt	DRER	0.056	1249	TGGg	DRHR	2.5,5
320	GCTg	GNNR	-	1250	TGGg	DRET	2.5,5
321	GCTg	AQSS	-	1251	GAGg	DRET	2.5,5
322	GTCg	AQSS	-	1252	TGGg	DRQR	5,20
323	GACg	SDNR	2.6	1253	GAGg	DRQR	<2.5
324	TTGg	DRHT	3	1254	GTGg	DRQR	<2.5
325	GTGg	DRAS	8.9	1255	GGGg	TRNR	<2.5
326	GATg	GNNR	15.6	1256	GCGg	TRNR	2.5,5
327	GTAg	AQSS	8	1257	GAGg	TRNR	<2.5
328	GATg	SDNR	35	1258	GTGg	TRNR	2.5,5
329	TGGg	DRAS	10.8	1259	GGGg	SQHR	<2.5
330	GCAg	AQSS	56.6	1260	GAGg	SQHR	2.5,5
331	AAAg	AQNA	+	1261	TGGg	SQHT	5,20
332	AACg	GDNV	+	1262	TGGg	SEHT	5,20
333	AAGg	DRTN	+	1263	GGAg	SEHT	5,20
334	ACAg	ASDR	+	1264	GTTg	GTAT	+
335	ACCg	KDDR	+	1265	GGGg	DRTR	+
336	ACGg	DRTD	+	1266	GTGg	DRTR	+
337	ACTg	LTDR	+	1267	GCTg	DRTR	-
338	AGAg	AQHA	+	1268	GGCg	GTHR	>2400
339	AGGg	DRHE	+	1269	GGGg	DRNR	45
340	AATg	THTN	+	1270	GCCg	GDNR	90
341	ATTg	NHAN	+	1271	GCCg	GDAR	>4400
342	ACAg	SSDR	+	1272	GCTg	GQDR	10

343	ACAg	GNER	+	1273	TGGg	DRHT	0.5
344	ACTg	KSDR	+	1274	GGGg	DRHR	0.4
345	AGGg	DRHN	+	1275	GATg	GTNR	3
346	ATTg	HTGT	+	1276	GCAg	GQDR	2
347	AAGg	DRNQ	+	1277	GTGg	DRER	15
348	AGGg	DRHQ	+	1278	GCGg	DRTR	+
349	AGTg	THTN	+	1279	GTGg	DRSR	3
350	ATGg	DREV	+	1280	GAGg	DRNR	1
351	AAGg	GDNV	-	1281	GGTg	GTGR	+
352	ACAg	GDNV	-	1282	GAAg	SQNR	0.5
353	ACCg	GDNV	-	1283	GGTg	GTHR	15
354	ACGg	GDNV	-	1284	GCAg	GQTR	+
355	ACTg	GDNV	-	1285	GGAg	AQHR	3
356	AGAg	GDNV	-	1286	GACg	GDNR	3
357	AGCg	GDNV	-	1287	GTTg	GTSR	5
358	AGGg	GDNV	-	1288	GTAg	SQSR	25
359	AGTg	GDNV	-	1289	GTCg	GDAR	40
360	ATAg	GDNV	-	1290	GCCg	RDDR	80
361	ATCg	GDNV	-	1291	GTAg	GQSR	+
362	ATGg	GDNV	-	1292	GAGg	DRDR	6
363	ATTg	GDNV	-	1293	GGCg	GDHR	40
364	AAAg	DRTN	-	1294	GTGg	DRKR	>1400
365	AACg	DRTN	-	1295	GCGg	DRDR	9
366	AATg	DRTN	-	1296	GCTg	GTER	65
367	ACAg	DRTN	-	1297	GTGg	SQSR	>1000
368	ACCg	DRTN	-	1298	GGTg	DTKR	+
369	ACTg	DRTN	-	1299	GGTg	STSR	+
370	AGAg	DRTN	-	1300	GACg	RAMQ	+

371	AGCg	DRTN	-	1301	GTAg	SESR	+
372	AGTg	DRTN	-	1302	GTAg	SQGR	+
373	ATAg	DRTN	-	1303	GTTg	WEMR	+
374	ATCg	DRTN	-	1304	GTCg	GETR	+
375	ATTg	DRTN	-	1305	GCTg	SRDR	+
376	AGTg	GTNV	-	1306	GCCg	KGDR	+
377	AAAg	ASDR	-	1307	GCGc	AKDR	0.5
378	AACg	ASDR	-	1308	GCGc	CKVR	6.5
379	AAGg	ASDR	-	1309	GCGc	QKLT	25
380	AATg	ASDR	-	1310	TGTc	TQAA	29.7
381	ACCg	ASDR	-	1311	TGTc	TPHT	41.6
382	ACTg	ASDR	-	1312	TGTc	DRER	81.8
383	AGAg	ASDR	-	1313	TGTc	AKDR	54.4
384	AGCg	ASDR	-	1314	TGTc	QKLT	46.7
385	AGGg	ASDR	-	1315	GCGc	TQAA	108.3
386	AGTg	ASDR	-	1316	GCGc	TPHT	188.9
387	ATAg	ASDR	-	1317	GACc	SDNR	0.019
388	ATCg	ASDR	-	1318	GCAc	ARDR	0.068
389	ATGg	ASDR	-	1319	GCAc	GQSR	0.055
390	ATTg	ASDR	-	1320	GACc	ARDR	9.3
391	AAAg	DRTD	-	1321	TTGg	MVQT	15.9
392	AACg	DRTD	-	1322	TTGg	VEST	6.4
393	AATg	DRTD	-	1323	TTGg	RRTT	27.5
394	ACAg	DRTD	-	1324	TTGg	GRNT	4.6
395	ACCg	DRTD	-	1325	CTGt	DRER	101
396	ACTg	DRTD	-	1326	GCGt	GSQR	13.1
397	AGAg	DRTD	-	1327	TGGg	MVQT	22.2
398	AGCg	DRTD	-	1328	TGGg	VEST	22.8

399	AGTg	DRTD	-	1329	TGGg	RRTT	47.9
400	ATAg	DRTD	-	1330	TGGg	GRNT	20
401	ATTg	DRTD	-	1331	GCAc	SDNR	2.5
402	AAAg	LTDR	-	1332	GCGc	SDNR	1.8
403	AACg	LTDR	-	1333	GCGc	ARDR	0.035
404	AAGg	LTDR	-	1334	GACc	GQSR	1.8
405	AATg	LTDR	-	1335	GCGc	GQSR	0.54
406	AGAg	LTDR	-	1336	GACc	DRER	33
407	AGGg	LTDR	-	1337	GATa	GTNR	+
408	ATAg	LTDR	-	1338	TAGg	DRKR	+
409	ATCg	LTDR	-	1339	TGAg	GQHS	+
410	ATGg	LTDR	-	1340	GGTt	GTHR	+
411	ATTg	LTDR	-	1341	GATa	DRKR	-
412	AAAg	DRHE	-	1342	GAAa	DRKR	-
413	AACg	DRHE	-	1343	GTTg	DRNT	-
414	AATg	DRHE	-	1344	GGTt	DRNT	-
415	ACAg	DRHE	-	1345	GAAG	GTNR	-
416	ACCg	DRHE	-	1346	TGAG	GTNR	-
417	ACTg	DRHE	-	1347	GGGG	GTSR	-
418	AGAg	DRHE	-	1348	GGTT	GTSR	-
419	AGCg	DRHE	-	1349	GAAA	GQHS	-
420	AGTg	DRHE	-	1350	GGGG	GTHR	-
421	ATAg	DRHE	-	1351	GAAG	GQHS	-
422	ATCg	DRHE	-	1352	GTTG	GTHR	-
423	ATGg	DRHE	-	1353	GCGG	GQHS	-
424	ATTg	DRHE	-	1354	GGGG	GQHS	-
425	AAAg	NHAN	-	1355	GACc	SQNR	105
426	AACg	NHAN	-	1356	AGGG	DRKR	25533

427	AAGg	NHAN	-	1357	CAAA	DRNT	25533
428	AATg	NHAN	-	1358	GACa	SQNR	14175
429	ACAg	NHAN	-	1359	GATG	DRNT	14175
430	ACCg	NHAN	-	1360	GATa	SQNR	14243
431	ACGg	NHAN	-	1361	CATG	DRNT	14243
432	ACTg	NHAN	-	1362	ACAg	SQNR	5158
433	AGAg	NHAN	-	1363	GGGA	DRKR	5158
434	AGCg	NHAN	-	1364	GGGt	SQNR	297
435	AGGg	NHAN	-	1365	TTGG	DRNT	2435
436	AGTg	NHAN	-	1366	GACG	DRNT	12228
437	ATAg	NHAN	-	1367	GACt	SQNR	78
438	ATGg	NHAN	-	1368	AAGG	DRNT	78
439	AGCg	DRNN	-	1369	GAGa	SQNR	823
440	AAAg	SSDR	-	1370	AAAG	DRNT	823
441	AACg	SSDR	-	1371	GAGc	SQNR	5944
442	AAGg	SSDR	-	1372	CCTG	DRNT	5944
443	AATg	SSDR	-	1373	GAGt	SQNR	1741
444	AGAg	SSDR	-	1374	GCAG	DRNT	1741
445	AGCg	SSDR	-	1375	GATc	SQNR	6115
446	AGGg	SSDR	-	1376	GATg	SQNR	89
447	AGTg	SSDR	-	1377	TGAG	DRKR	-
448	ATAg	SSDR	-	1378	GAGG	DRNT	89
449	ATCg	SSDR	-	1379	TAGG	DRNT	95
450	ATGg	SSDR	-	1380	GAAA	SQNR	1009
451	ATTg	SSDR	-	1381	CAGG	DRNT	98
452	AAAg	GNER	-	1382	TTTG	DRNT	6115
453	AACg	GNER	-	1383	GGGG	DRNT	-
454	AAGg	GNER	-	1384	TGAG	DRNT	-

455	AATg	GNER	-	1385	GGTT	DRKR	-
456	ACCg	GNER	-	1386	TGAG	SQNR	-
457	ACGg	GNER	-	1387	TGAG	GTSR	-
458	ACTg	GNER	-	1388	TGAG	GTHR	-
459	AGAg	GNER	-	1389	GGGt	SQDR	15
460	AGCg	GNER	-	1390	GCGG	DRER	2
461	AGGg	GNER	-	1391	GCTG	SQDR	10
462	AGTg	GNER	-	1392	GCTt	SQDR	2
463	ATAg	GNER	-	1393	GCGt	SQDR	1000
464	ATCg	GNER	-	1394	GCTt	DRER	66
465	ATGg	GNER	-	1395	GGTG	DRHK	+
466	ATTg	GNER	-	1396	AAGG	DRHK	+
467	AACg	KSDR	-	1397	AGGG	DRHL	+
468	AAGg	KSDR	-	1398	GGGG	HSLH	+
469	AATg	KSDR	-	1399	GGGG	DRHK	+
470	AGGg	KSDR	-	1400	GGGG	DRER	+
471	AGTg	KSDR	-	1401	GTGT	DRER	+
472	ATAg	KSDR	-	1402	TGGG	DRHK	+
473	ATCg	KSDR	-	1403	TGGG	HSLH	+
474	AAAg	DRHN	-	1404	AGGG	DRHK	+
475	AACg	DRHN	-	1405	TCGG	DRER	+
476	AATg	DRHN	-	1406	CCGT	DRER	+
477	ACAg	DRHN	-	1407	CCGG	DRER	+
478	ACCg	DRHN	-	1408	tcgg	DRET	50
479	ACTg	DRHN	-	1409	gagt	DRNR	50
480	ATAg	DRHN	-	1410	gtag	GQAR	+
481	ATCg	DRHN	-	1411	gtag	AQSR	+
482	ATGg	DRHN	-	1412	gttg	SQAR	+

483	ATTg	DRHN	-	1413	gtgg	SQDR	2
484	AAAg	SQSA	-	1414	gggg	DRAR	31
485	AATg	SQSA	-	1415	gtgg	DRAR	2
486	ACCg	SQSA	-	1416	ggag	SQHR	0.5
487	ACTg	SQSA	-	1417	ggtg	SQHR	1
488	AGAg	SQSA	-	1418	gcgg	SQDR	2
489	AGCg	SQSA	-	1419	gtgg	DRET	12.5
490	AGTg	SQSA	-	1420	ggcg	SQHR	1
491	ATCg	SQSA	-	1421	gccg	SDDR	+
492	ATTg	SQSA	-	1422	gatg	ATNR	+
493	AACg	HTGT	-	1423	ggcg	SDHR	+
494	AAGg	HTGT	-	1424	gttg	GTAR	+
495	AATg	HTGT	-	1425	gtcg	SDAR	+
496	ACCg	HTGT	-	1426	gctg	ARER	+
497	ACGg	HTGT	-	1427	ataa	TQGQ	+
498	ACTg	HTGT	-	1428	gcgg	ARER	+
499	AGCg	HTGT	-	1429	gcgt	SDNR	+
500	AGGg	HTGT	-	1430	gcgt	ARDR	+
501	ATAg	HTGT	-	1431	GCCT	SDVR	4.9
502	AAAg	DRNQ	-	1432	GCGg	DRVR	4.9
503	AACg	DRNQ	-	1433	TGAG	AQHT	36.6
504	AATg	DRNQ	-	1434	GCAG	AQTR	4
505	ACAg	DRNQ	-	1435	GGAg	ADHR	1
506	ACCg	DRNQ	-	1436	GACT	ADNK	1
507	ACTg	DRNQ	-	1437	GCAt	AADR	13.7
508	AGAg	DRNQ	-	1438	GCTG	SNDR	13.7
509	AGCg	DRNQ	-	1439	GATG	STNK	13.7
510	AGTg	DRNQ	-	1440	GCTt	SHDR	4

511	ATAg	DRNQ	-	1441	GCTG	SQDK	4
512	ATCg	DRNQ	-	1442	GCCt	SDSK	36.6
513	ATTg	DRNQ	-	1443	ATCT	SDSK	36.6
514	AAAg	DRHQ	-	1444	CTCt	DADQ	13.3
515	AACg	DRHQ	-	1445	GCTC	SRDR	13.3
516	AATg	DRHQ	-	1446	GGAG	AQHK	13.3
517	ACAg	DRHQ	-	1447	TAGg	DRAQ	40.3
518	ACCg	DRHQ	-	1448	AACT	ADNT	40.3
519	ACGg	DRHQ	-	1449	GCTA	SATK	40.3
520	ACTg	DRHQ	-	1450	CGGA	SKHA	+
521	AGAg	DRHQ	-	1451	GGCT	SKHA	+
522	AGCg	DRHQ	-	1452	TATA	SKHA	+
523	AGTg	DRHQ	-	1453	GTGG	DRHK	+
524	ATAg	DRHQ	-	1454	GGGA	SKHA	+
525	ATCg	DRHQ	-	1455	GGCC	SKHA	+
526	ATGg	DRHQ	-	1456	GGAT	SKHA	+
527	ATTg	DRHQ	-	1457	CGGG	DRHK	+
528	AAAg	DREV	-	1458	GGGC	SKHA	+
529	AACg	DREV	-	1459	TCGG	DRHK	+
530	AATg	DREV	-	1460	GGGT	DRER	+
531	ACAg	DREV	-	1461	GCAG	DRER	+
532	ACCg	DREV	-	1462	GTGT	SKHA	+
533	ACTg	DREV	-	1463	GTAG	DRER	+
534	AGAg	DREV	-	1464	TCAG	TNDK	14.1
535	AGCg	DREV	-	1465	TCAG	GQDK	21.6
536	AGGg	DREV	-	1466	TGTT	QTHE	2.1
537	AGTg	DREV	-	1467	gggg	drer	+
538	ATAg	DREV	-	1468	cccg	dcht	+

539	ATCg	DREV	-	1469	gacc	sqhr	+
540	ATTg	DREV	-	1470	gacc	QSNR	+
541	TACg	DRNT	-	1471	gacc	ADNR	+
542	TATg	DRNT	-	1472	gacc	TSNR	+
543	TCAg	DRNT	-	1473	gacc	ATNR	+
544	TCCg	DRNT	-	1474	gacc	PTNR	+
545	TCGg	DRNT	-	1475	gcac	SRDR	+
546	TCTg	DRNT	-	1476	gcac	PRDR	+
547	TTAg	DRNT	-	1477	gcac	GRDR	+
548	GGGg	DRHT	0.5	1478	gcac	SHDR	+
549	ATGg	DRAV	+	1479	gcac	VRDR	+
550	AAGg	DRNN	+	1480	gcac	AADR	+
551	AAAg	GTNV	-	1481	gcac	SKDR	+
552	AACg	GTNV	-	1482	gcac	ARER	+
553	AAGg	GTNV	-	1483	gcac	GNSR	+
554	ACAg	GTNV	-	1484	gcac	GSSR	+
555	ACCg	GTNV	-	1485	gcac	GTTR	+
556	ACGg	GTNV	-	1486	gacc	SNNR	+
557	ACTg	GTNV	-	1487	gcgc	SKER	+
558	AGAg	GTNV	-	1488	gcgc	EKDR	+
559	AGCg	GTNV	-	1489	gcgc	YSDR	+
560	AGGg	GTNV	-	1490	gcgc	TTGR	+
561	ATAg	GTNV	-	1491	gcgc	GKDR	+
562	ATCg	GTNV	-	1492	gcgc	WAER	+
563	ATGg	GTNV	-	1493	gcgc	TEGR	+
564	ATTg	GTNV	-	1494	gcgc	KGDR	+
565	AAAg	THTN	-	1495	gcgc	DKDR	+
566	AACg	THTN	-	1496	gcgc	RNDH	+

567	AAGg	THTN	-	1497	gcgc	ERGR	+
568	ACAg	THTN	-	1498	gcgc	WTER	+
569	ACCg	THTN	-	1499	gcgc	SNDR	+
570	ACGg	THTN	-	1500	gcgc	ANDR	+
571	AGAg	THTN	-	1501	gcgc	ERDR	+
572	AGCg	THTN	-	1502	gcgc	RNDR	+
573	AGGg	THTN	-	1503	gcgc	SSDR	+
574	ATAg	THTN	-	1504	gcgc	YDGR	+
575	ATCg	THTN	-	1505	gcgc	KDDR	+
576	ATGg	THTN	-	1506	gcgc	RDDR	+
577	AACg	SQSA	-	1507	gtgc	GTAR	+
578	AAGg	SQSA	-	1508	tcgc	EQDR	+
579	ACGg	SQSA	-	1509	tcgc	SRDK	+
580	AGGg	SQSA	-	1510	tcgc	NRDK	+
581	ATGg	SQSA	-	1511	acgc	RDDR	+
582	AAAg	DRNN	-	1512	acgc	TGEK	+
583	AACg	DRNN	-	1513	acgc	RERT	+
584	AATg	DRNN	-	1514	acgc	GRQE	+
585	ACAg	DRNN	-	1515	acgc	EYER	+
586	ACCg	DRNN	-	1516	acgc	GESR	+
587	ACTg	DRNN	-	1517	TTAg	aqss	+
588	AGAg	DRNN	-	1518	GTTg	dntr	+
589	AGTg	DRNN	-	1519	GGAg	gqhe	+
590	ATAg	DRNN	-	1520	GAGg	drgt	+
591	ATCg	DRNN	-	1521	gcgg	drgt	+
592	ATTg	DRNN	-	1522	TAGg	drgt	+
593	AATg	GTNV	+	1523	ttgg	drgt	+
594	ATAg	SQSA	+	1524	TCGg	drgt	+

595	CAAg	HQHE	+	1525	GGGg	drhd	+
596	CAAg	GSHE	+	1526	TGGg	drhd	+
597	CAAg	GQHQ	+	1527	gcgg	ardr	+
598	CAAg	KNQN	+	1528	GAGg	gsnr	+
599	CAAg	TNHH	+	1529	gatg	slnr	+
600	CCAg	HQHE	+	1530	gacg	slnr	+
601	CCAg	SSHE	+	1531	gcgg	drvs	+
602	CCAg	RTDQ	+	1532	gtgg	drvs	+
603	CGAg	AQHE	+	1533	TCGg	drvs	+
604	CGAg	DRHN	+	1534	ttgg	drvs	+
605	CTAg	HQSE	+	1535	gatg	sinr	+
606	CACg	HQHE	+	1536	gacg	sinr	+
607	CACg	GSHE	+	1537	gcgc	CKDR	+
608	CACg	HQHD	+	1538	gcgc	YKCR	+
609	CACg	NNHE	+	1539	gcgc	NKSP	+
610	CACg	AGGR	+	1540	gcgc	CKQS	+
611	CCCg	GSHE	+	1541	gcgc	QQVT	+
612	CCCg	RSNE	+	1542	gcgc	QTSP	+
613	CCCg	KSHE	+	1543	gcgc	HVIN	+
614	CGCg	HQHE	+	1544	tgtc	EPRP	+
615	CGCg	GSHE	+	1545	tgtc	ESQP	+
616	CGCg	KGWV	+	1546	tgtc	QHQP	+
617	CGCg	RDWV	+	1547	tgtc	GRQA	+
618	CGCg	GHHE	+	1548	tgtc	ARRG	+
619	CGCg	RNTT	+	1549	tgtc	NESD	+
620	CTCg	HQHE	+	1550	tgtc	VNMD	+
621	CAGg	HQHE	+	1551	tgtc	RNGK	+
622	CAGg	NRNV	+	1552	tgtc	RSPW	+

623	CAGg	KDVT	+	1553	tggg	NYTt	+
624	CAGg	IHDH	+	1554	tggg	AYAt	+
625	CAGg	DRNI	+	1555	tggg	YYHt	+
626	CAGg	TNHH	+	1556	tggg	VTNt	+
627	CAGg	NNPP	+	1557	tggg	HRQt	+
628	CCGg	HQHE	+	1558	tggg	PFYt	+
629	CCGg	GSHE	+	1559	ttgg	LQSt	+
630	CCGg	TNRS	+	1560	ttgg	GRLt	+
631	CCGg	DRTA	+	1561	ttgg	FRSt	+
632	CGGg	AIDQ	+	1562	ttgg	SKRt	+
633	CGGg	RRGK	+	1563	ttgg	RGKt	+
634	CGGg	RETA	+	1564	ttgg	NGSt	+
635	CGGg	DRHE	+	1565	ttgg	RQPt	+
636	CTGg	IHDH	+	1566	gcgt	DRLS	+
637	CTGg	KTES	+	1567	gcgt	LSLA	+
638	CATg	HQHE	+	1568	gcgt	SVVL	+
639	CATg	GSHE	+	1569	ctgt	VNGP	+
640	CATg	GNHE	+	1570	ctgt	WSII	+
641	CATg	LGIG	+	1571	ctgt	AIWL	+
642	CATg	GAHE	+	1572	ctgt	MIMF	+
643	CATg	HQHD	+	1573	ctgt	ERCL	+
644	CATg	VSHE	+	1574	ctgt	AILT	+
645	CATg	GTHE	+	1575	ctgt	VNQR	+
646	CATg	YSKE	+	1576	GGGA	drer	+
647	CATg	GSGA	+	1577	gaac	sqnk	+
648	CCTg	HQHE	+	1578	gcag	sqdk	+
649	CCTg	GQHA	+	1579	atgg	drtg	+
650	CCTg	SGKE	+	1580	CTGg	drtg	+

651	CCTg	ACHE	+	1581	acgg	drtg	+
652	CGTg	QDDI	+	1582	gcgT	drea	+
653	CGTg	RDSS	+	1583	gcct	drer	+
654	CGTg	RQHS	+	1584	gcat	dger	+
655	CGTg	RLQH	+	1585	gcct	dger	+
656	CGTg	RRNK	+	1586	gctt	dger	+
657	CGTg	RSTA	+	1587	gcgg	eryr	+
658	CGTg	RQHE	+	1588	gcgg	trhr	+
659	CTTg	HQHE	+	1589	gcgg	srer	+
660	CTTg	GSHE	+	1590	gcgg	srar	+
661	CTTg	QNHE	+	1591	gcag	tqtr	+
662	CAAg	GQNE	+	1592	gcag	aqsr	+
663	CACg	KSAE	+	1593	gcTG	aqsr	+
664	CAGg	DRNE	+	1594	gcag	tqsr	+
665	CATg	GTNE	+	1595	gcAT	tqsr	+
666	CCAg	HTSE	+	1596	gcTG	gsdr	+
667	CCGg	DRTE	+	1597	gcTG	ssar	+
668	CCTg	NTSE	+	1598	gcTG	qlvr	+
669	CGAg	GQHE	+	1599	gcTG	atsr	+
670	CGGg	DRKE	+	1600	gcTG	stgr	+
671	CTAg	SQTE	+	1601	gcTG	ttar	+
672	CTTg	GTAE	+	1602	gcTG	atar	+
673	CAAg	DRNI	-	1603	gcgg	atar	+
674	CACg	DRNI	-	1604	gcCG	sdvr	+
675	CATg	DRNI	-	1605	gcCG	sdtr	+
676	CCAg	DRNI	-	1606	gcCG	sdar	+
677	CCCg	DRNI	-	1607	gcTG	sdsr	+
678	CCGg	DRNI	-	1608	gcCC	sdsr	+

679	CCTg	DRNI	-	1609	gcTC	sdsr	+
680	CGAg	DRNI	-	1610	gcgc	hrdr	+
681	CGCg	DRNI	-	1611	gcAC	hrdr	+
682	CGTg	DRNI	-	1612	gcgc	hssr	+
683	CTAg	DRNI	-	1613	gcgc	ytsr	+
684	CTCg	DRNI	-	1614	gcTC	ytsr	+
685	CTGg	DRNI	-	1615	gcgc	qrnr	+
686	CTTg	DRNI	-	1616	gcga	qrnr	+
687	CCAg	KSHE	-	1617	gcTC	qrnr	+
688	CGAg	KSHE	-	1618	gcTA	qrnr	+
689	CTAg	KSHE	-	1619	gcAC	aqtr	+
690	CGAg	RSNE	-	1620	gcAC	tqnr	+
691	CAAg	DRTA	-	1621	gcAC	gqar	+
692	CACg	DRTA	-	1622	gcAC	tqtr	+
693	CATg	DRTA	-	1623	gcgc	stsr	+
694	CCAg	DRTA	-	1624	gcag	stsr	+
695	CCCg	DRTA	-	1625	gcTG	stsr	+
696	CCTg	DRTA	-	1626	gcgg	stsr	+
697	CGAg	DRTA	-	1627	gcAC	sstr	+
698	CGCg	DRTA	-	1628	gcAT	sstr	+
699	CGTg	DRTA	-	1629	gcTC	sstr	+
700	CTAg	DRTA	-	1630	gcTT	sstr	+
701	CTCg	DRTA	-	1631	gcTC	stnr	+
702	CTTg	DRTA	-	1632	gcTC	ttar	+
703	CAAg	DRHE	-	1633	gcTC	stlr	+
704	CACg	DRHE	-	1634	gcTC	stir	+
705	CATg	DRHE	-	1635	gcAC	stir	+
706	CCAg	DRHE	-	1636	gcTT	stir	+

707	CCCg	DRHE	-	1637	gcAT	stir	+
708	CCTg	DRHE	-	1638	gcTC	ntsr	+
709	CGAg	DRHE	-	1639	gcTT	ntsr	+
710	CGCg	DRHE	-	1640	gcAT	stsr	+
711	CGTg	DRHE	-	1641	gcTC	gtsr	+
712	CTAg	DRHE	-	1642	gcTT	gtsr	+
713	CTCg	DRHE	-	1643	gcTC	tltr	+
714	CTTg	DRHE	-	1644	gcAC	tltr	+
715	CCAg	RSTA	-	1645	gcCC	tltr	+
716	CGAg	RSTA	-	1646	gcTC	sltr	+
717	CGGg	RSTA	-	1647	gcTT	sltr	+
718	CTAg	RSTA	-	1648	gcCC	sltr	+
719	CTGg	RSTA	-	1649	gcCT	sltr	+
720	CCGg	HQSE	-	1650	gcCC	hdnr	+
721	CGGg	HQSE	-	1651	gcCC	khsr	+
722	CTGg	HQSE	-	1652	gcCC	qhnr	+
723	CAAg	DRAE	-	1653	gcTC	qhnr	+
724	CACg	DRAE	-	1654	gcCA	adnr	+
725	CATg	DRAE	-	1655	gcga	qrdr	+
726	CCAg	DRAE	-	1656	gcCA	qrdr	+
727	CCCg	DRAE	-	1657	gcaa	gqhr	+
728	CCTg	DRAE	-	1658	gcaa	aqtr	+
729	CGAg	DRAE	-	1659	gcAT	aqtr	+
730	CGCg	DRAE	-	1660	gcTA	qtar	+
731	CGTg	DRAE	-	1661	gcTA	atsr	+
732	CTAg	DRAE	-	1662	gcTC	atsr	+
733	CTCg	DRAE	-	1663	Acgg	drhk	+
734	CTTg	DRAE	-	1664	gcTA	qltr	+

735	CTAg	KSAE	-	1665	gcga	rrdr	+
736	CTGg	KSAE	-	1666	gcTC	rrdr	+
737	CAAg	DRNE	-	1667	gcTA	ahtr	+
738	CACg	DRNE	-	1668	gcTC	ahtr	+
739	CATg	DRNE	-	1669	gcTA	thtr	+
740	CCAg	DRNE	-	1670	gcCA	thtr	+
741	CCCg	DRNE	-	1671	gcCC	thtr	+
742	CCTg	DRNE	-	1672	gcTA	hvhr	+
743	CGAg	DRNE	-	1673	Gcgg	drhl	+
744	CGCg	DRNE	-	1674	gcTG	hvhr	+
745	CGTg	DRNE	-	1675	gcCA	ahtr	+
746	CTAg	DRNE	-	1676	gcCA	ahnr	+
747	CTCg	DRNE	-	1677	gcCA	shnr	+
748	CTGg	DRNE	-	1678	gcTA	shnr	+
749	CTTg	DRNE	-	1679	gcCC	shnr	+
750	CAAg	GTNE	-	1680	gcTC	shnr	+
751	CACg	GTNE	-	1681	gcCA	rdar	+
752	CAGg	GTNE	-	1682	gcgg	hrdr	+
753	CCCg	GTNE	-	1683	gcTT	hrdr	+
754	CCGg	GTNE	-	1684	gcTG	hrdr	+
755	CGAg	GTNE	-	1685	gcgT	sryr	+
756	CGCg	GTNE	-	1686	gcAT	sryr	+
757	CGGg	GTNE	-	1687	gcTT	sryr	+
758	CTAg	GTNE	-	1688	gcgT	srsr	+
759	CTCg	GTNE	-	1689	gcgC	srsr	+
760	CTGg	GTNE	-	1690	gcAT	srsr	+
761	CTTg	GTNE	-	1691	gcAC	srsr	+
762	CACg	HTSE	-	1692	gcTT	srsr	+

763	CAGg	HTSE	-	1693	gcTC	srsr	+
764	CGCg	HTSE	-	1694	gcAT	aqsr	+
765	CGGg	HTSE	-	1695	gcAC	aqsr	+
766	CAAg	DRTE	-	1696	gcTT	aqsr	+
767	CACg	DRTE	-	1697	gcTC	aqsr	+
768	CATg	DRTE	-	1698	gcAT	qqhr	+
769	CCAg	DRTE	-	1699	gcAC	qqhr	+
770	CCCg	DRTE	-	1700	gcAT	sqhr	+
771	CCTg	DRTE	-	1701	gcag	sqtr	+
772	CGAg	DRTE	-	1702	gcTT	sttr	+
773	CGCg	DRTE	-	1703	gcTC	sttr	+
774	CGTg	DRTE	-	1704	gcTT	sthr	+
775	CTAg	DRTE	-	1705	gcTC	sthr	+
776	CTCg	DRTE	-	1706	gcTT	stvr	+
777	CTTg	DRTE	-	1707	gcTC	stvr	+
778	CAAg	NTSE	-	1708	gcAT	stvr	+
779	CACg	NTSE	-	1709	gcAC	stvr	+
780	CAGg	NTSE	-	1710	gcTT	thsr	+
781	CTCg	NTSE	-	1711	gcTC	thsr	+
782	CTAg	GQHE	-	1712	gcTT	qhtr	+
783	CAAg	DRKE	-	1713	gcTC	qhtr	+
784	CACg	DRKE	-	1714	gcAT	qhtr	+
785	CAGg	DRKE	-	1715	gcAC	qhtr	+
786	CATg	DRKE	-	1716	gcgT	qhtr	+
787	CCAg	DRKE	-	1717	gcgC	qhtr	+
788	CCCg	DRKE	-	1718	gcTT	thtr	+
789	CCGg	DRKE	-	1719	gcCT	sdsr	+
790	CCTg	DRKE	-	1720	gcCT	shtr	+

791	CGAg	DRKE	-	1721	gcTT	shtr	+
792	CGCg	DRKE	-	1722	gcCT	sdrr	+
793	CGTg	DRKE	-	1723	gcCC	sdrr	+
794	CTAg	DRKE	-	1724	aaag	GAAN	+
795	CTCg	DRKE	-	1725	aaag	HQNL	+
796	CTGg	DRKE	-	1726	aaag	GNNT	+
797	CTTg	DRKE	-	1727	aaag	NQNL	+
798	CAAg	SQTE	-	1728	aaag	TQNN	+
799	CAGg	SQTE	-	1729	aaag	GQNA	+
800	CCGg	SQTE	-	1730	aaag	HQNV	+
801	CGGg	SQTE	-	1731	aaag	GQNT	+
802	CAGg	GTAE	-	1732	aaag	TQNH	+
803	CTAg	GTAE	-	1733	ataa	AQSL	+
804	CCAg	GSHE	+	1734	ataa	GQAA	+
805	CGCg	FNHE	+	1735	ataa	GQST	+
806	CCGg	DRHD	+	1736	ataa	NQGQ	+
807	CTGg	DRAE	+	1737	ataa	GQSS	+
808	CCTg	GSHE	+	1738	ataa	IQST	+
809	GAGg	DRAR	+	1739	tcag	NTLR	+
810	GAGg	ARSR	+	1740	tcag	HQDR	+
811	GTGg	SRSR	+	1741	tcag	TTDK	+
812	GAGg	DKAR	+	1742	ggtt	SEHR	+
813	GCCg	DKDR	+	1743	tgtt	HMHH	+
814	GAGg	DKSR	+	1744	tgtt	HHHV	+
815	GAGg	DKTR	+	1745	tgtt	HHHQ	+
816	GGCg	DKVR	+	1746	tgtt	HHHA	+
817	GCCg	AKDR	+	1747	tgtt	HHHN	+
818	GGGg	DRDK	+	1748	tgtt	HLHQ	+

819	GAGg	DREK	+	1749	gcga	drer	+
820	GTAg	GQER	+	1750	tggt	drht	+
821	GTAg	GQTR	+	1751	gcgg	drhr	+
822	GGCg	SDKR	+	1752	ccgg	drht	+
823	GGCg	GDKR	+	1753	ccgg	drhS	+
824	GCCg	GDDR	+	1754	TCGg	drhs	+
825	GCTg	GDER	+	1755	tcgc	NKDK	+
826	GGTg	GTAR	+	1756	gtgg	drht	+
827	GCTg	GTDR	+	1757	gagt	drer	+
828	GCTg	GTTR	+	1758	ccgg	drha	+
829	GCTg	QTSR	+	1759	ccgg	drhV	+
830	GCTg	QTTR	+	1760	acgg	drhE	+
831	GAGg	ARKR	+	1761	acgt	drer	+
832	GGCg	AQKR	+	1762	ataa	TQAQ	+
833	GGCg	SQKR	+	1763	ggtt	SDHR	+
834	GGCg	GTKR	+	1764	tcag	TTNS	+
835	GGTg	DTHR	+	1765	tgga	drht	+
836	GTCg	DRHR	-	1766	ggga	HRHV	+
837	GAAg	DRNR	-	1767	gcga	arer	+
838	GATg	DRNR	-	1768	acgg	drht	+
839	GACg	DRNR	-	1769	ccgg	drhr	+
840	GTAg	DRNR	-	1770	aaag	NQNA	+
841	GGAg	DRAR	-	1771	aaag	NQNN	+
842	GGTg	DRAR	-	1772	gcgt	arer	+
843	GGCg	DRAR	-	1773	gcgg	drhk	+
844	GAAg	DRAR	-	1774	gcgc	TRdr	+
845	GATg	DRAR	-	1775	acgg	drhY	+
846	GACg	DRAR	-	1776	aaag	GNAN	+

847	GTTg	DRAR	-	1777	ccgg	drhN	+
848	GTCg	DRAR	-	1778	tcgg	drhv	+
849	GCAg	DRAR	-	1779	acgg	drhv	+
850	GCTg	DRAR	-	1780	tgtt	QHHT	+
851	GCCg	DRAR	-	1781	gaga	drer	+
852	GTAg	DRDR	-	1782	gtgg	drtg	+
853	GTTg	DRDR	-	1783	gcgg	drtg	+
854	GTCg	DRDR	-	1784	ttgg	drtg	+
855	GCAg	DRDR	-	1785	TCGg	drtg	+
856	GGAg	DRSR	-	1786	gcgg	drvn	+
857	GGTg	DRSR	-	1787	gcta	QTTR	+
858	GGCg	DRSR	-	1788	gcta	ATTR	+
859	GAAg	DRSR	-	1789	tcta	GTTR	+
860	GATg	DRSR	-	1790	gcta	GTDR	+
861	GACg	DRSR	-	1791	ggga	HRHA	+
862	GTTg	DRSR	-	1792	aagg	DRGA	+
863	GTCg	DRSR	-	1793	aagg	DRNV	+
864	GGTg	DRTR	-	1794	aagg	DRND	+
865	GGCg	DRTR	-	1795	aagg	DRQA	+
866	GTCg	DRTR	-	1796	aagg	DRGI	+
867	GGAg	DKDR	-	1797	aagg	DRQT	+
868	GGAg	DKTR	-	1798	gcgc	arer	+
869	GAAg	DKTR	-	1799	gcgt	dger	+
870	GATg	DKTR	-	1800	gcTC	stsr	+
871	GACg	DKTR	-	1801	gcgg	drha	+
872	GTTg	DKTR	-	1802	gcTA	RRdr	+
873	GCAg	DKTR	-	1803	Acgg	drhn	+
874	GGTg	AKDR	-	1804	gcAT	sqtr	+

875	GAAg	AKDR	-	1805	aaag	NQNT	+
876	GGGg	AKER	-	1806	aaag	VQNT	+
877	GTGg	AKER	-	1807	GGCG	drer	+
878	GCAg	AKER	-	1808	gcgg	dret	+
879	GGAg	DRDK	-	1809	GCAA	drer	+
880	GAAg	DRDK	-	1810	gacc	RDNR	+
881	GTAg	DRDK	-	1811	gacc	SSNR	+
882	GGAg	DREK	-	1812	gcgc	GKER	+
883	GAAg	DREK	-	1813	gcgc	SRER	+
884	GATg	DREK	-	1814	gcgc	YDTR	+
885	GTAg	DREK	-	1815	gcgc	ATDR	+
886	GTTg	DREK	-	1816	gtgc	GTGR	+
887	GTCg	DREK	-	1817	gtgc	KESR	+
888	GCAg	DREK	-	1818	tgtc	STEH	+
889	GTAg	SQNR	-	1819	ttgg	WHMt	+
890	GCAg	SQNR	-	1820	gcat	drer	+
891	GCTg	SQNR	-	1821	gaag	sqnk	+
892	GGGg	SQSR	-	1822	GTGA	drer	+
893	GGAg	SQSR	-	1823	gcgg	ERdr	+
894	GGTg	SQSR	-	1824	gcTG	GNdr	+
895	GGCg	SQSR	-	1825	gcgT	trdr	+
896	GAGg	SQSR	-	1826	gcgc	srtr	+
897	GAAg	SQSR	-	1827	gcgT	srtr	+
898	GATg	SQSR	-	1828	gcAC	RTdr	+
899	GTTg	SQSR	-	1829	gcAT	gqar	+
900	GTCg	SQSR	-	1830	gcAT	tqtr	+
901	GTAg	AQHR	-	1831	gcAC	stsr	+
902	GATg	GQDR	-	1832	gcTC	NGdr	+

903	GGGg	SQDR	-	1833	gcTT	stsr	+
904	GGAg	SQDR	-	1834	gcTC	star	+
905	GGTg	SQDR	-	1835	gcTT	star	+
906	GGCg	SQDR	-	1836	Ccgg	drhg	+
907	GAAg	SQDR	-	1837	Tcgg	drhg	+
908	GATg	SQDR	-	1838	gcCC	adnr	+
909	GACg	SQDR	-	1839	gcCT	adnr	+
910	GTTg	SQDR	-	1840	gcaa	QRdr	+
911	GTCg	SQDR	-	1841	gcAT	gqhr	+
912	GGGg	GQER	-	1842	gcTC	thtr	+
913	GGAg	GQER	-	1843	Acgg	drhl	+
914	GGTg	GQER	-	1844	gcgT	SKdr	+
915	GGCg	GQER	-	1845	gcgT	SSdr	+
916	GAGg	GQER	-	1846	gcgT	HRdr	+
917	GAAg	GQER	-	1847	ggtt	SVHR	+
918	GATg	GQER	-	1848	tgtt	HHHS	+
919	GACg	GQER	-	1849	tgtt	HMHA	+
920	GTGg	GQER	-	1850	tgtt	HMHD	+
921	GTTg	GQER	-	1851	tgtt	HMHQ	+
922	GCGg	GQER	-	1852	AAAT	TQNT	0.12
923	GGGg	GQTR	-	1853	TCAG	HQDK	0.038
924	GGTg	GQTR	-	1854	ACAT	GQTR	0.11
925	GGCg	GQTR	-	1855	TGTT	HMHE	0.11
926	GAGg	GQTR	-	1856	GGTT	SDKR	0.038
927	GATg	GQTR	-	1857	GGGA	HRHL	0.11
928	GTGg	GQTR	-	1858	GCTA	HTTR	0.12
929	GTTg	GQTR	-	1859	ATAA	AQSA	0.12
930	GCGg	GQTR	-	1860	AAGG	DRNA	0.038

Table A.10: Information of training dataset in DB1 database. In the table, the bases of the DNA sequence are in the 5'-3' order. The amino acids at the contacting positions 2, -1, 3, 6 are retained. The binding affinities are described by either the quantitative information or binding status (binding /non-binding).

B Visualization Colouring

B.1 Non-linear dimensionality reduction methods

B.1.1 Relative information from PCA

Table B.1 provides the detailed information about the relationship between the number of the principal components (PCs) and the data samples. As mentioned in Subsection 4.2.2, it is impossible to use the first two PCs to explain the specific 320-D binary data set.

Number of Eigenvec-tor	Explained Variances	Number of Eigenvec-tor	Explained Variances	Number of Eigenvec-tor	Explained Variances	Number of Eigenvec-tor	Explained Variances
1	9.72%	59	79.72%	117	94.50%	175	98.90%
2	16.10%	60	80.11%	118	94.63%	176	98.93%
3	20.08%	61	80.49%	119	94.76%	177	98.97%
4	23.82%	62	80.87%	120	94.88%	178	99.00%
5	26.60%	63	81.24%	121	95.00%	179	99.04%

6	29.11%	64	81.61%	122	95.13%	180	99.07%
7	31.56%	65	81.98%	123	95.25%	181	99.10%
8	33.97%	66	82.34%	124	95.36%	182	99.13%
9	36.13%	67	82.69%	125	95.48%	183	99.16%
10	38.21%	68	83.04%	126	95.59%	184	99.19%
11	40.20%	69	83.39%	127	95.71%	185	99.22%
12	41.91%	70	83.72%	128	95.82%	186	99.25%
13	43.50%	71	84.05%	129	95.92%	187	99.27%
14	45.06%	72	84.38%	130	96.03%	188	99.30%
15	46.58%	73	84.70%	131	96.13%	189	99.33%
16	48.00%	74	85.02%	132	96.23%	190	99.35%
17	49.34%	75	85.33%	133	96.33%	191	99.38%
18	50.59%	76	85.63%	134	96.42%	192	99.41%
19	51.82%	77	85.93%	135	96.51%	193	99.43%
20	53.01%	78	86.23%	136	96.60%	194	99.46%
21	54.17%	79	86.53%	137	96.69%	195	99.48%
22	55.28%	80	86.81%	138	96.78%	196	99.50%
23	56.34%	81	87.08%	139	96.86%	197	99.53%
24	57.37%	82	87.35%	140	96.94%	198	99.55%
25	58.36%	83	87.62%	141	97.02%	199	99.57%
26	59.31%	84	87.89%	142	97.09%	200	99.59%
27	60.25%	85	88.15%	143	97.17%	201	99.61%
28	61.16%	86	88.42%	144	97.24%	202	99.63%
29	62.04%	87	88.67%	145	97.31%	203	99.64%
30	62.87%	88	88.93%	146	97.38%	204	99.66%
31	63.69%	89	89.18%	147	97.44%	205	99.67%
32	64.50%	90	89.43%	148	97.51%	206	99.69%
33	65.28%	91	89.68%	149	97.58%	207	99.70%
34	66.05%	92	89.92%	150	97.64%	208	99.72%
35	66.81%	93	90.16%	151	97.71%	209	99.73%
36	67.55%	94	90.39%	152	97.77%	210	99.75%
37	68.26%	95	90.62%	153	97.83%	211	99.76%
38	68.96%	96	90.84%	154	97.89%	212	99.78%
39	69.64%	97	91.06%	155	97.95%	213	99.79%
40	70.28%	98	91.27%	156	98.01%	214	99.81%
41	70.91%	99	91.48%	157	98.07%	215	99.82%
42	71.53%	100	91.69%	158	98.13%	216	99.84%
43	72.13%	101	91.89%	159	98.18%	217	99.85%

44	72.71%	102	92.09%	160	98.24%	218	99.86%
45	73.28%	103	92.28%	161	98.29%	219	99.88%
46	73.82%	104	92.47%	162	98.34%	220	99.89%
47	74.34%	105	92.65%	163	98.39%	221	99.90%
48	74.84%	106	92.83%	164	98.44%	222	99.91%
49	75.33%	107	93.00%	165	98.49%	223	99.93%
50	75.81%	108	93.17%	166	98.53%	224	99.94%
51	76.28%	109	93.34%	167	98.58%	225	99.95%
52	76.75%	110	93.50%	168	98.62%	226	99.96%
53	77.20%	111	93.66%	169	98.67%	227	99.97%
54	77.64%	112	93.81%	170	98.71%	228	99.98%
55	78.07%	113	93.95%	171	98.75%	229	99.98%
56	78.50%	114	94.09%	172	98.79%	230	99.99%
57	78.91%	115	94.23%	173	98.82%	231	99.99%
58	79.32%	116	94.37%	174	98.86%	232	100%

Table B.1: Statistical information of eigenvectors. In this table, the proportion of data information which can be explained by different number of PCs are summarised.

B.1.2 Locally Linear Embedding (LLE)

According to the description in Subsection 4.2.2, the performance of the Locally Linear Embedding (LLE) model depends on the selection of the K nearest neighbours. To identify K nearest neighbours for the model, a cost function which is based on Euclidean distance is defined as follow:

$$\varepsilon(W) = \sum_{i}^{N} |\mathbf{x}_i - \sum_{j=1}^{K} W_{ij}\mathbf{x}_j|^2 \tag{B.1}$$

where W_{ij} is a weight between a point i and its neighbours j. The appropriate weights are obtained by optimising the cost function which is subjected to two constraints: 1) each data point $\mathbf{x}_i$ is reconstructed only from its neighbours; 2) the rows of the weight matrix sum to one: $\sum_{j=1}^{K} W_{ij} = 1$. With the two constraints, the cost function can be minimised

by minimised by using a Lagrange multiplier to enforce the constraint that $\sum_{j=1}^{K} W_{ij} = 1$. Through comparing the value of ε, the number of nearest neighbours (K) can be identified.

B.2 Amino acids classification check list

Table B.2 provides the definition of the colour codes based on the hydrophobicity and hydrophilicity of amino acids. As listed in the table, the amino acids with the hydrophobicity properties are defined to be coloured in red or orange, and the relevant values are marked as '1' or '2'. With the changes of the physicochemical properties, the colour code is altered from red to blue, and the relevant values are increased from '1' to '6'. Since there are four amino acids in each zinc finger protein to participate the interaction with DNA sequence, the colour codes which are applied in the visualisation results are calculated based on the properties of the four amino acids. For example, when the four amino acids in the protein is 'VLIF', the relative value of the colour code ought to be '4'. Table B.3 lists the statistical information of the training dataset based on the hydrophobicity and hydrophilicity of the zinc finger in each data sample.

Characteristic	Value	Colour	Amino acid	Abbreviation
Hydrophobic (4)	1	red	Valine	V
			Leucine	L
			Isoleucine	I
			Phenylalanine	F
	2	Orange	Glycine	G
			Alanine	A
			Methionine	M
			Cysteine	C
			Proline	P
↓	3	Yellow	Tryptophan	W
			Tyreonine	Y
	4	Green	Serine	S
			Threonine	T
	5	Cyan	Asparagine	N
			Glutamine	Q
	6	Blue	Aspartic	D
			Glutamic	E
			Lysine	K
			Arginine	R
Hydrophilic (24)			Histidine	H

Table B.2: Amino acid colour map

Hydrophobicity	Number of data	Proportion
4	0	0
5	0	0
6	2	0.10%
7	2	0.10%
8	2	0.10%
9	1	0.05%
10	3	0.16%
11	4	0.22%
12	22	1.18%
13	6	0.32%
14	33	1.77%
15	76	4.09%
16	138	7.42%
17	115	6.18%
18	177	9.52%
19	331	17.80%
20	277	14.89%
21	129	6.94%
22	238	12.80%
23	101	5.43%
24	203	10.91%
Hydrophilicity	Total: 1860	100%

Table B.3: Statistics of training dataset based on physicochemical characteristic of amino acid. This table summaries the hydrophobicity and hydrophilicity of the zinc fingers in the training dataset.

C Prediction Models

C.1 D.1 320-D original database creation

As summarised in Table 3.3, there are in total 1860 data samples in the training dataset. In order to reconstitute the 320-D original database for the prediction models investigation, Table C.1 is generated as a reference for the database creation. In the table, 25 published papers are listed using the defined index in Appendix B.1 Table A.1. The second column provides the number of adopted data samples from each published paper. And the related proportion is listed in the third column. Since the proportion of three categories is defined as 5:4:1, the adopted data samples from each individual data source are separated into three groups according to the proportion and the number of the data samples in each group is presented in the last three columns. When creating the 320-D original database, for example, from the first data source, 132 data samples are randomly selected to be the training data, 107 data samples for the test data set, and 26 data samples for the validation

data set. Repeating the random selection process one hundred times, the database is reconstituted. This method can ensure that all of the 26 data sources are covered in each category, and the number of data samples in each group remains unchanged.

Data source	No. of adopted data samples	Percentage of data samples	No. of training data	No.of test data	No. of validation data
1	264	14.19%	132	107	25
2	214	11.505%	107	86	21
3	43	2.312%	21	17	5
4	326	17.53%	163	131	32
5	52	2.80%	26	21	5
6	385	20.70%	193	153	39
7	288	15.48%	144	116	28
8	31	1.67%	15	12	4
9	0	0	0	0	0
10	0	0	0	0	0
11	11	0.59%	6	4	1
13	6	0.323%	3	2	1
14	9	0.484%	5	3	1
15	1	0.054%	1	0	0
16	0	0	0	0	0
17	18	0.968%	9	7	2
18	13	0.699%	7	5	1
19	18	0.968%	9	7	2
20	110	5.914%	55	44	11
21	5	0.269%	3	1	1
22	10	0.538%	5	4	1
23	19	1.02%	10	5	4
24	14	0.753%	7	5	2
25	3	0.161%	2	1	0
26	20	1.075%	10	6	4
Total	1860	100%	933	737	190

Table C.1: Statistical information of the adopted 25 data sources in the original database.

C.2 PCA based reconstruction data visualisation

This section includes nine Figures which show the NeuroScale visualisation results of the reconstruction datasets based on different number of eigenvectors. In these Figures, the number of eigenvectors increase from 50 to 234. The sub-Figures for 233 eigenvectors and 234 eigenvectors, respectively, show identical visualisation results, which proves that the data reconstructed by the first 233 eigenvectors from PCA can describe the characteristics of the interaction as effectively as the 320-D original data.

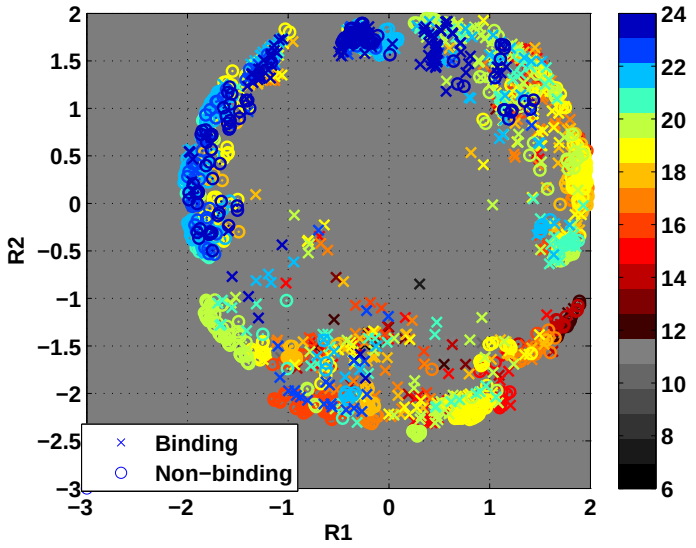


Figure C.1: Visualisation result of the reconstructed database using 50 eigenvectors.

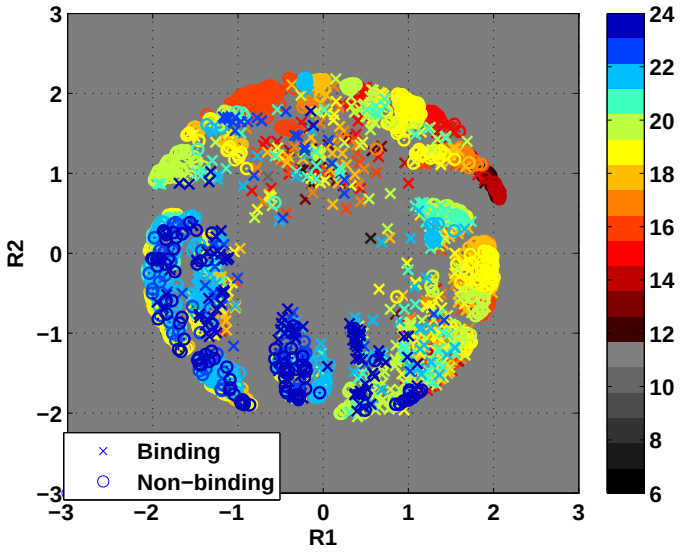


Figure C.2: Visualisation result of the reconstructed database using 100 eigenvectors.

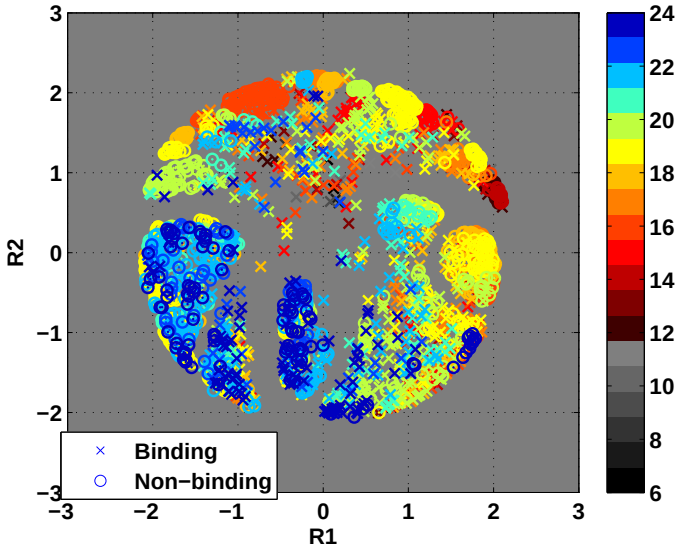


Figure C.3: Visualisation result of the reconstructed database using 150 eigenvectors.

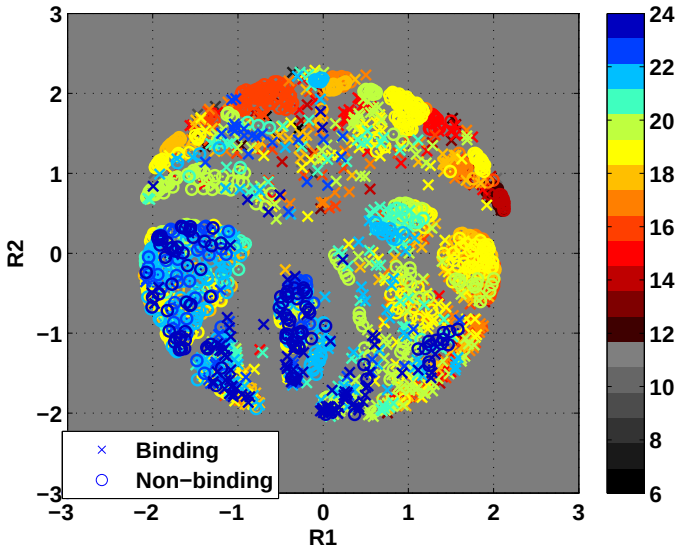


Figure C.4: Visualisation result of the reconstructed database using 200 eigenvectors.

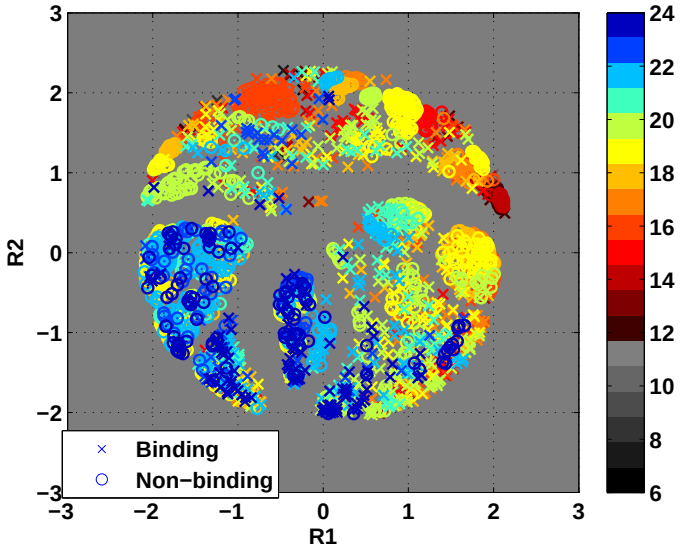


Figure C.5: Visualisation result of the reconstructed database using 210 eigenvectors.

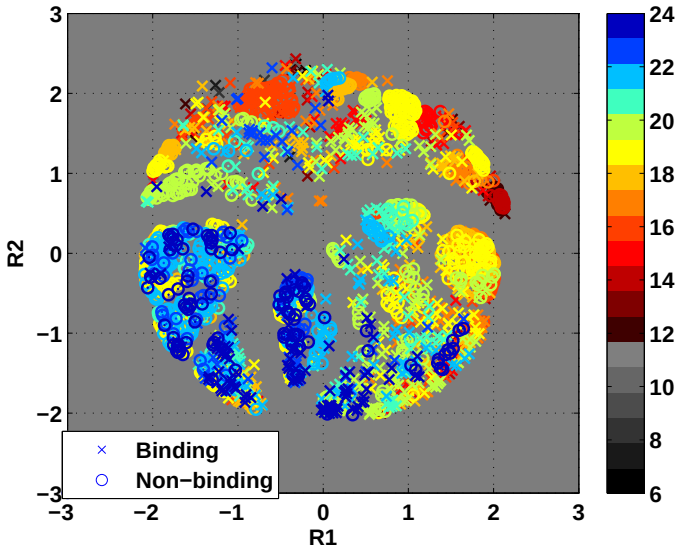


Figure C.6: Visualisation result of the reconstructed database using 220 eigenvectors.

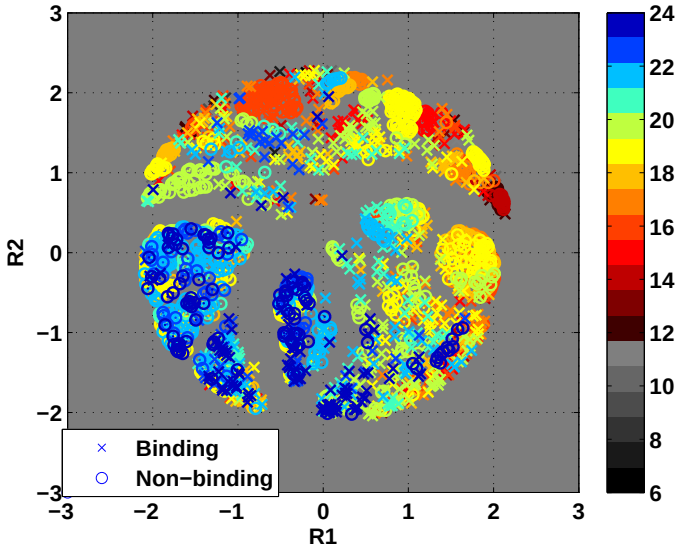


Figure C.7: Visualisation result of the reconstructed database using 230 eigenvectors.

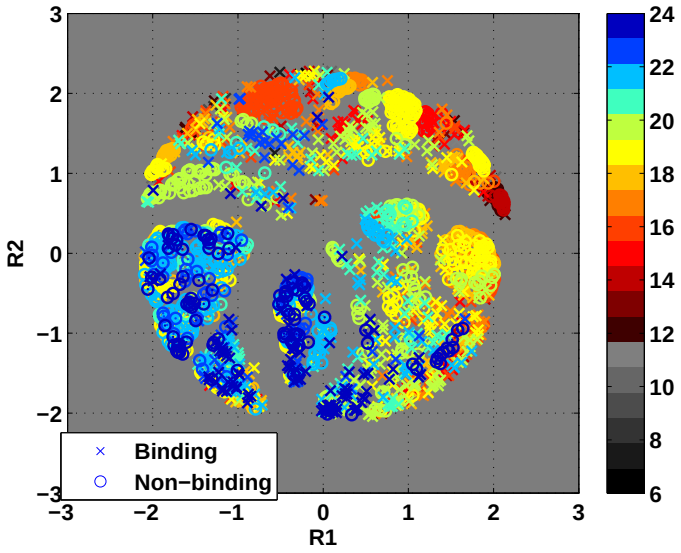


Figure C.8: Visualisation result of the reconstructed database using 233 eigenvectors.

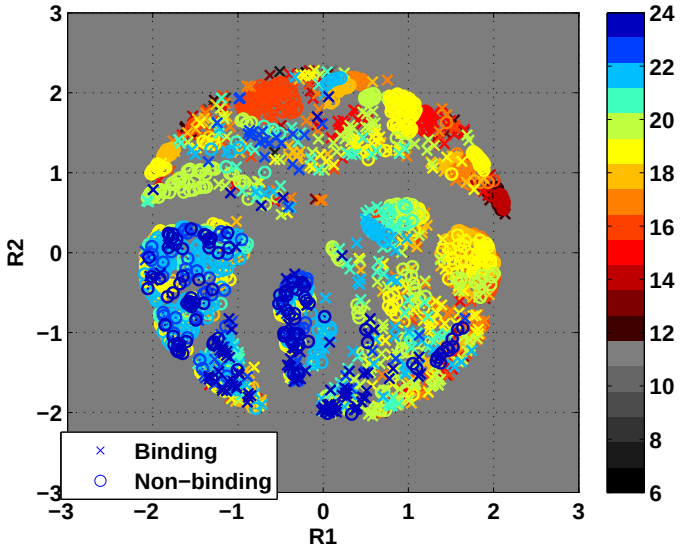


Figure C.9: Visualisation result of the reconstructed database using 234 eigenvectors.

C.3 Quality criteria - Receiver operator characteristic (ROC)

As defined in Subsection 5.1.4, the ROC curve as a two-dimensional graph is employed to depict relative tradeoffs between benefits (true positive) and costs (false positive) Fawcett (2006). Figure C.10 shows a confusion matrix Fawcett (2006) in which there are four possible outcomes compared with target outputs, given a classifier and a group of data

examples which contains only either one of two statuses, binding and non-binding. If P is the total number of the binding examples and N represents the total number of the non-binding examples, then the **true positive rate (sensitivity)** of a classifier is estimated as:

$$tp\ rate \approx \frac{\text{Positivescorrectlyclassified}}{\text{Totalpositives}} = \frac{TP}{TP+FN}$$

The **false positive rate** of the classifier is:

$$fp\ rate \approx \frac{\text{Positivesfalseclassified}}{\text{Totalpositives}} = \frac{FP}{FP+TN}$$

And the **accuracy** of the classifier is:

$$accuracy = \frac{TP+TN}{P+N}$$

The **specificity** of the classifier is:

$$specificity = \frac{TN}{FP+TN}$$

In this thesis, the binding status of a given data example is represented as [0 1] (binding) or [1 0] (non-binding). To verify the binding status of the prediction outcome for the example, the Euclidean distance between the prediction result and the target output (either [0 1] or [1 0]) is calculated. If the predicted outcome is [1 0], but the target output is [0 1], the maximum Euclidean distance approximately equal to 1.414; if the predicted outcome is [1 0], and the target output is [1 0], the minimum Euclidean distance between them is 0. In Figure C.11, the intermediate value of the Euclidean distance, 0.707, is marked. The red solid line is the adjustable threshold. Through altering the value of the threshold , the prediction binding status changes. The ROC curve can be plotted according to different *tp* rate and *fp* rate which are calculated by adjusting the threshold.

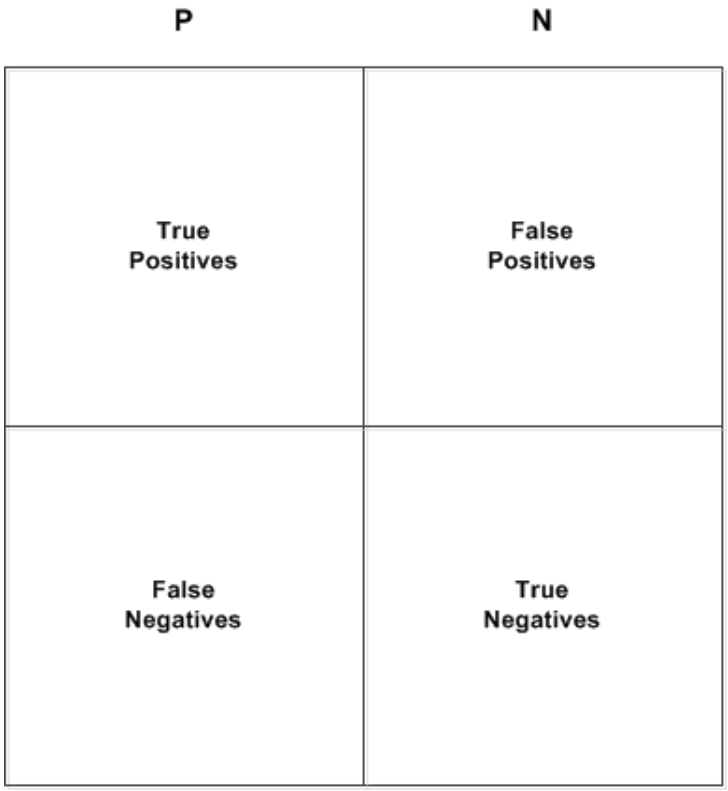


Figure C.10: Confusion matrix. In this matrix, the four possible outcomes are arranged into four blocks. P denotes the total number of the positive samples; N is the total number of the negative samples.

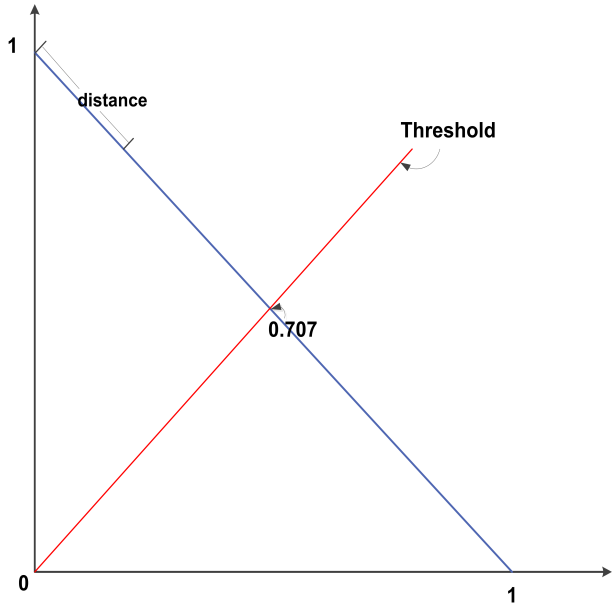


Figure C.11: The ROC curve calculation standard. The distance is calculated based on the Euclidean metric. The two points marked on the X, Y axis represent the binding and non-binding statuses: [0 1] and [1 0]. The range of the threshold adjustment is from 0 to 1.414.

C.4 Parameters and results of relevant prediction algorithms

C.4.1 Prediction results based on Minkowski

Figure C.12, C.13 and C.14 are the plots of the normalised classification error of the *k*-NN, MLP and RBF models. For the 2-D Minkowski metric based reconstruction database, the number of nearest neighbours for the *k*-NN is adjusted from 1 to 21 with the interval of 2. According to Figure C.12, when 9 neighbours are used to define the target test data samples, the normalised error of the *k*-NN reaches the smallest value at 0.4351 and the error for the validation data is 0.4384. Figure C.13 shows the normalised error of the MLP classifier where the range of the hidden centres is same as that in the model when used for the 2-D Euclidean distance based reconstruction database. With the similar reducing trend as of Figure 5.3, 57 hidden centres are selected this time to implement the prediction of validation data samples. The relevant normalised error of the validation dataset is 0.4288. Figure C.14 presents the results of the normalised error for the RBF model. The number of centres is changed from 2 to 80 with the interval of 2. Compared with the MLP, the RBF classifier is more erroneous than the MLP. With respect to the errors of both training and test data, 46 hidden centres are selected for this model; the normalised error of validation dataset with 46 hidden centres is 0.7133.

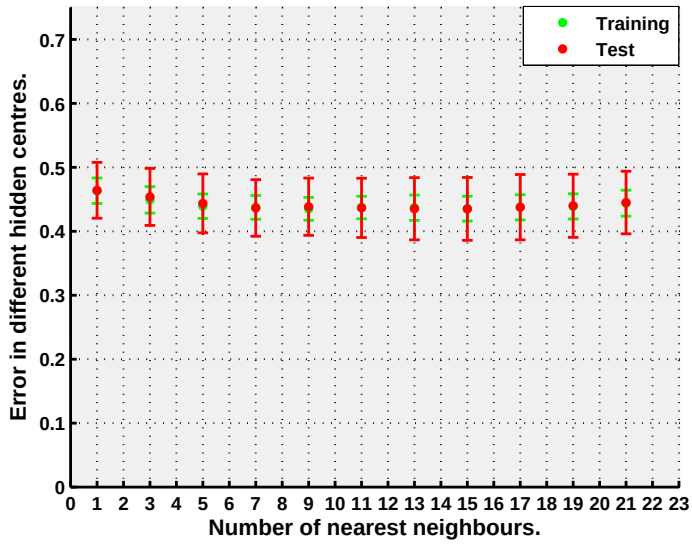


Figure C.12: The *k*-NN normalised classification error of the 2-D reconstruction data based on the Minkowski distance. When the number of neighbours is 9, both test and validation datasets have the smallest normalised classification error: 0.4351 and 0.4384.

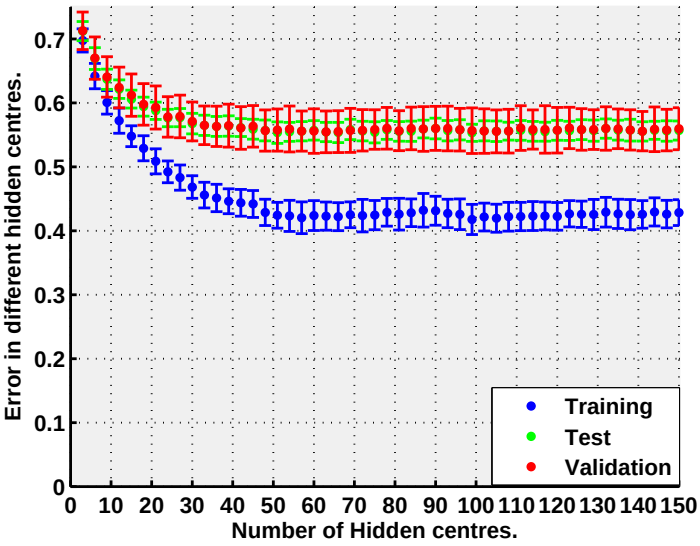


Figure C.13: The MLP normalised classification error for the the 2-D reconstruction data based on the Minkowski distance. The normalised errors for the training dataset are generally better than those for the test and validation datasets. When the hidden centres set to be 57, the error of training dataset is 0.4203, the error of test dataset is 0.5557, and the error of validation dataset is 0.5559.

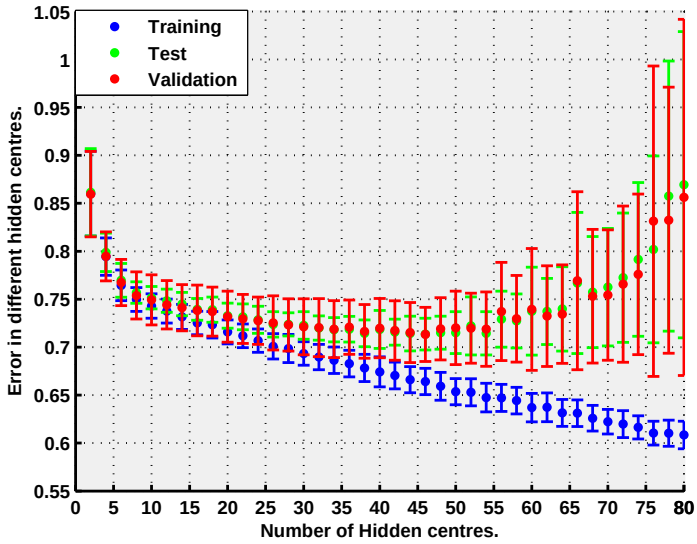


Figure C.14: The RBF normalised classification error for the 2-D reconstruction data based on the Minkowski distance. When the number of hidden centres is lower than 55, the differences of the normalised errors between three datasets are very small. Hereafter, due to over-training, they become larger and larger as well as the error bar of the normalised error for the test and validation datasets. When the hidden centres is 46, the error of test dataset has the lowest value: 0.7136, while the error of training dataset is 0.6642, and the error of validation dataset is 0.7133.

Figure C.15 depicts the ROC curves with respect to the classifiers for the 2-D Minkowski metric based reconstruction database.

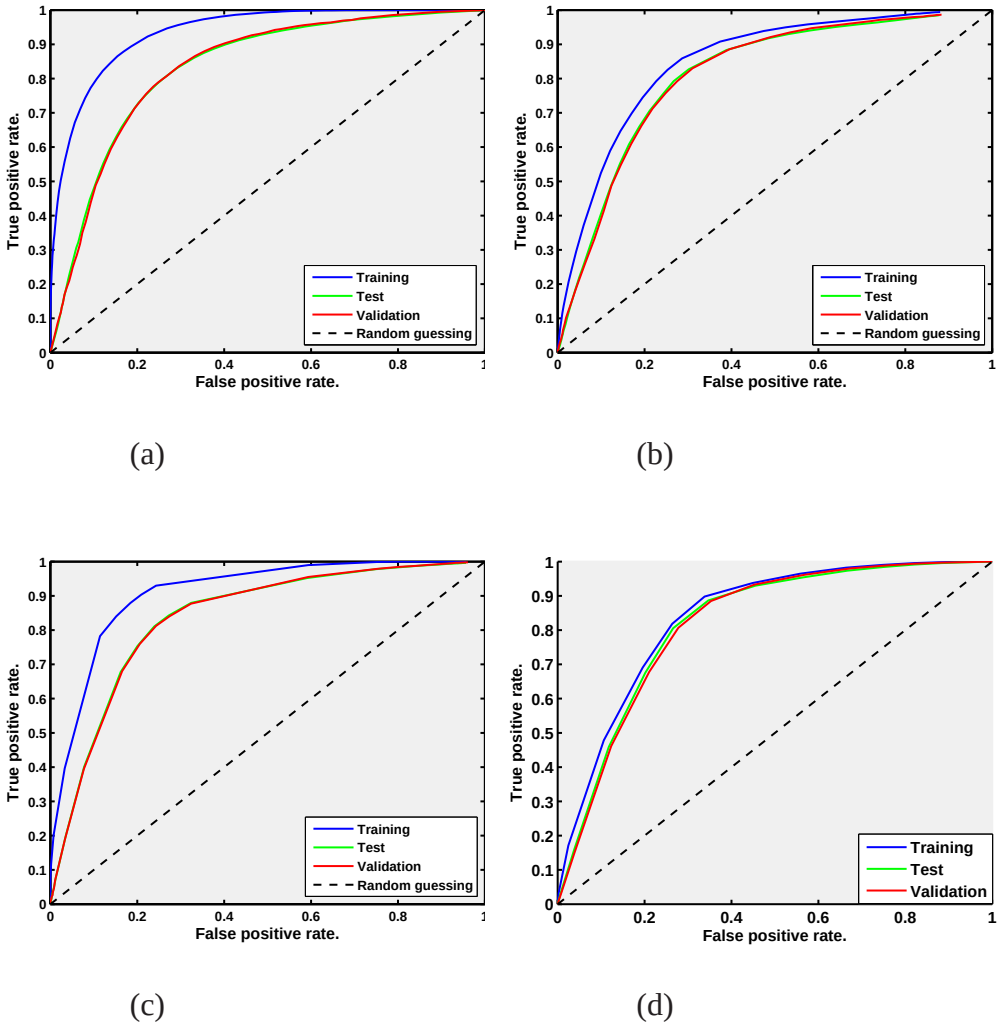


Figure C.15: The ROC curves of different classifiers using the 2-D reconstruction datasets based on the Minkowski distance. (a) MLP classifier (AUC values: 0.9346, 0.8310 and 0.8315); (b) RBF classifier (AUC values: 0.7301, 0.6921 and 0.6965); (c) SVM regression classifier (AUC values: 0.8680, 0.7934 and 0.7970) and (d) RVM classifier (AUC values: 0.8390, 0.8212 and 0.8184). Generally, the classifiers performs much better than random guessing (AUC: 0.5).

C.4.2 Prediction results based on 320-D original data

Figure C.16, C.17 and C.18 show the normalised classification error of the *k*-NN, MLP and RBF models respectively. For the 320-D original database, the number of nearest neighbours for the *k*-NN is adjusted from 1 to 11 with the interval of 2. According to Figure C.16, when one neighbour is selected to define the target test data samples, the normalised error is at its smallest of 0.342; and for the validation data it is 0.3710. Figure C.17 shows the normalised error of the MLP classifier. In this Figure, the normalised

error for the training dataset is zero when the number of the hidden centres is changed from 3 to 150 with the interval of 3. This is because of the over-training of the high dimensional input. When 21 hidden centres are selected, the normalised error for the validation dataset is 0.2144. Figure C.19 presents the results of the normalised error for the RBF model. The number of hidden centres is changed from 2 to 180 with the interval of 2. Comparing with the error in the MLP algorithm, the error of the RBF classifier is higher. Regarding the errors for both training and test data sets, 142 hidden centres are selected for the validation data set, where the normalised error is 0.5616. Figure C.19 plots the ROC curves of the MLP, RBF and SVM regression models.

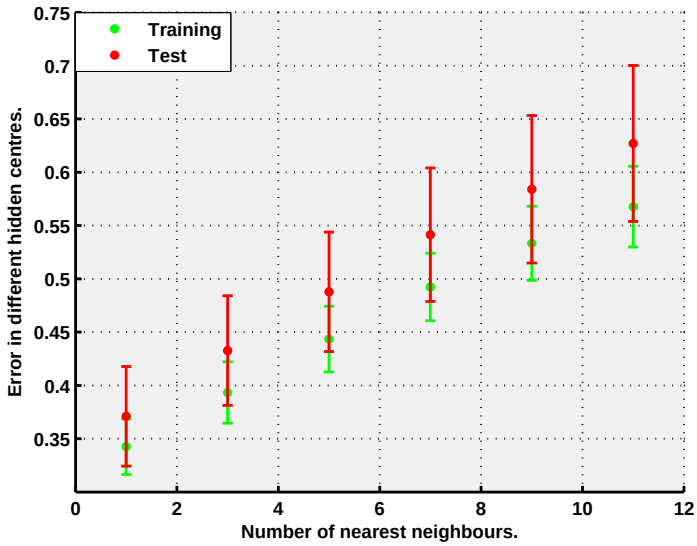


Figure C.16: The *k*-NN normalised classification error for the 320-D original data. When there is only one neighbour. the normalised classification errors for the test and validation datasets are smallest at 0.3275 and 0.3257, respectively.

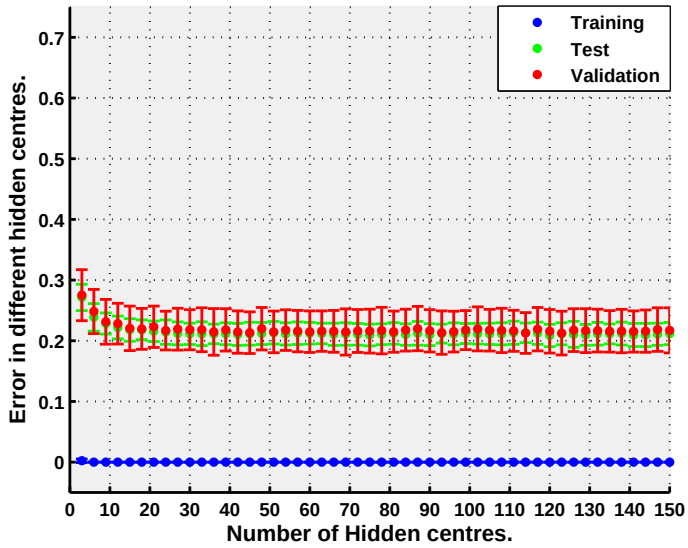


Figure C.17: the MLP normalised classification error for the 320-D original data. The normalised errors for the training dataset are zero due to the over-training of the high dimensional input. When the hidden centres set to be 21, the error of training dataset is 7.7504×10^{-13}, the error of test dataset is 0.2104, and the error of validation dataset is 0.2144.

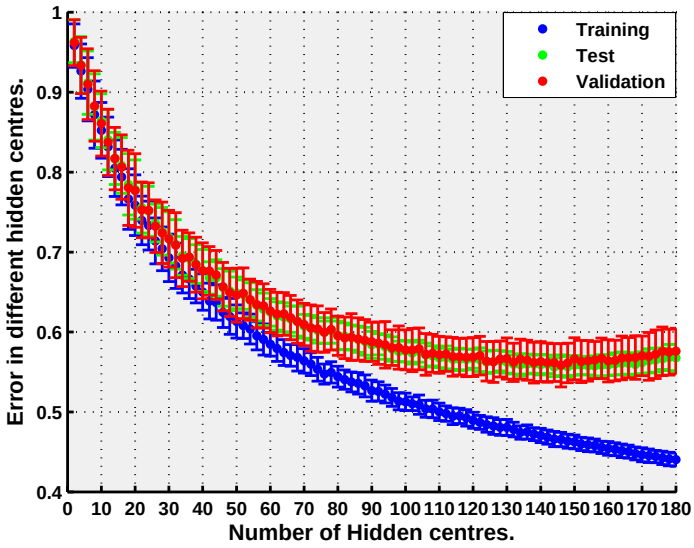


Figure C.18: The RBF normalised classification error for the 320-D original data. When there are fewer than 80 hidden centres, the normalised error is similar between three datasets. Hereafter, the error difference becomes significant. When the hidden centres is 142, the error of test dataset has the lowest value: 0.5597, while the error of training dataset is 0.4685, and the error of validation dataset is 0.5616.

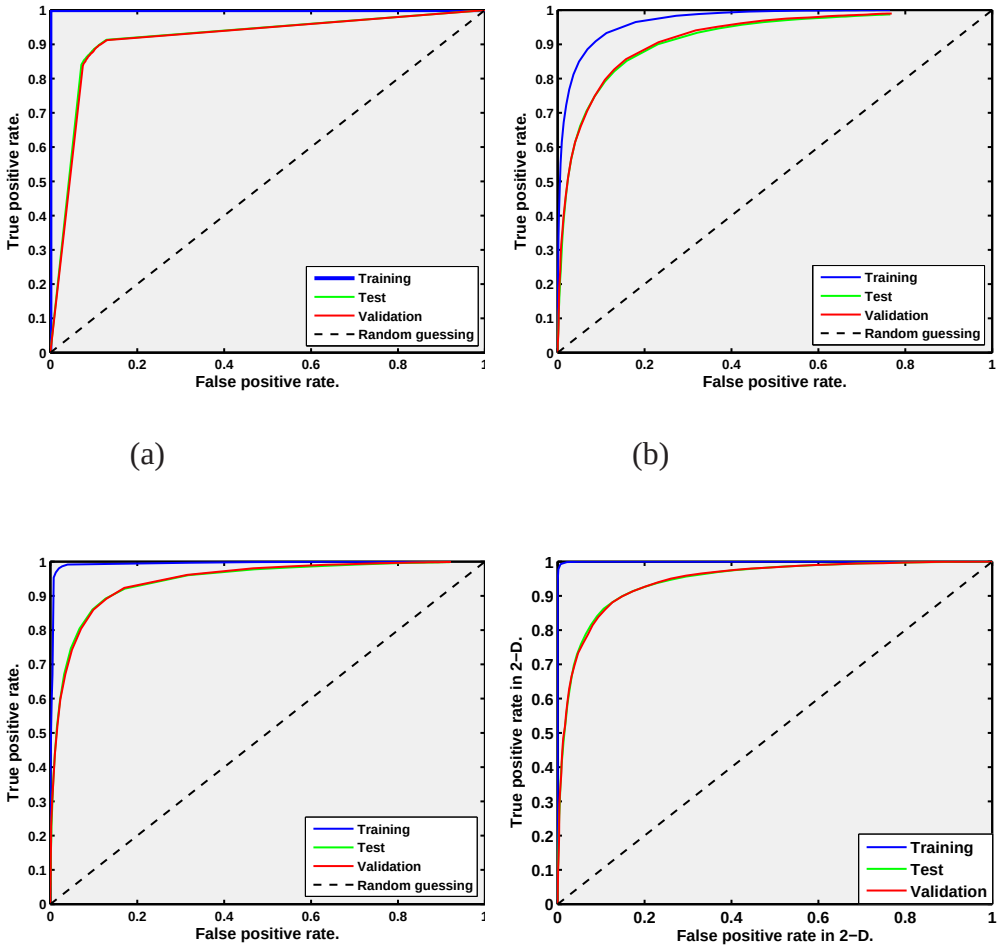


Figure C.19: The ROC curves for the 320-D original data. (a) MLP classifier (AUC values: 1, 0.9140 and 0.9118); (b) RBF classifier (AUC values: 0.7357, 0.6837 and 0.6904); (c) SVM regression classifier (AUC values: 0.9142, 0.8633 and 0.8665) and (d) RVM classifier (AUC values: 0.9997, 0.9440 and 0.9439). Generally, the classifiers performs much better than random guessing (AUC: 0.5).

C.4.3 Prediction results on 320-D reconstruction data

In Figure C.20, C.21 and C.22, the normalised classification errors of the *k*-NN, MLP and RBF models are plotted. For the 320-D reconstruction database, the number of nearest neighbours for the *k*-NN is adjusted from 1 to 11 with the interval of 2. According to Figure C.20, when 5 neighbours are selected to define the target test data samples, the normalised error has the smallest value at 0.2693 and the error for the validation data is 0.2851. Figure C.21 shows the normalised error of the MLP classifier. In this Figure, the

same as the results of the 320-D original database, there is no normalised errors for the training dataset as the number of hidden centres changes from 3 to 150 with the interval of 3. When 45 hidden centres are selected, the normalised error for the validation dataset is 0.2144. Figure C.22 presents the results of the normalised error for the RBF model. The number of hidden centres is changed from 2 to 180 with the interval of 2. Taking the errors of both training and test data as a reference, 142 hidden centres are selected for the validation data set, resulting the normalised error of validation dataset being 0.5607. Figure C.23 shows the ROC curves of the selected regression models which constructed using the 320-D reconstruction database.

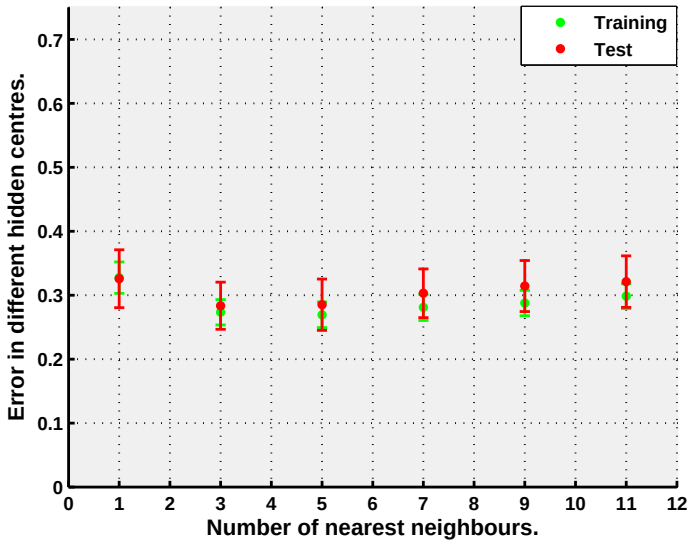


Figure C.20: The *k*-NN normalised classification error for the 320-D reconstruction data. When there is only one neighbour, the smallest normalised classification errors are achieved for the test and validation datasets at 0.2104 and 0.2144, respectively.

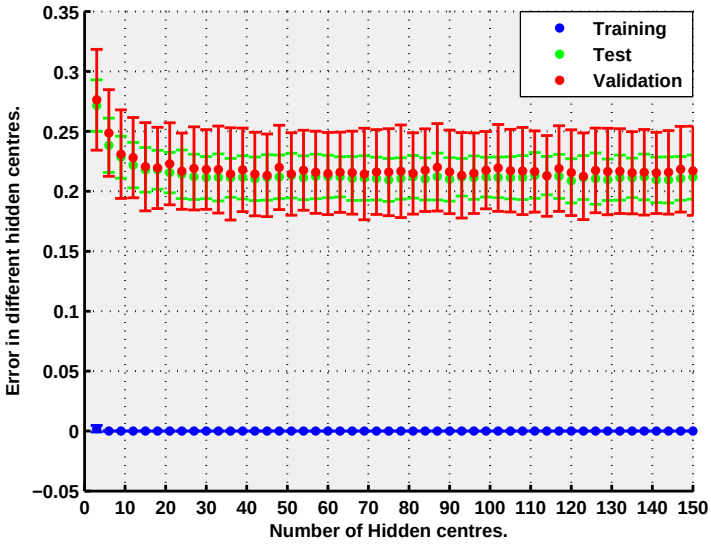


Figure C.21: The MLP normalised classification error for the 320-D reconstruction data. When the hidden centres is 45, the error of training dataset is 7.5916×10^{-13}, the error of test dataset is 0.2116, and the error of validation dataset is 0.2133.

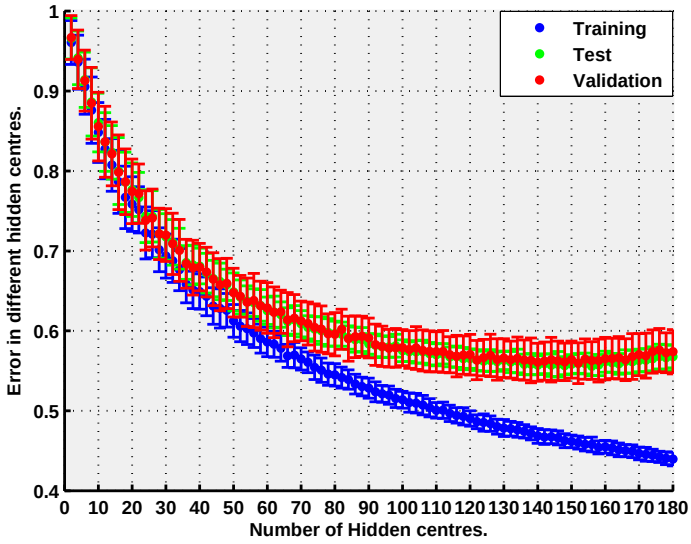


Figure C.22: The RBF normalised classification error for the 320-D reconstruction data. The differences of the normalised errors between the three datasets becomes significant for more than 90 hidden centres. When the hidden centres is 140, the error of test dataset has the lowest value: 0.5571, while the error of training dataset is 0.4686, and the error of validation dataset is 0.5607.

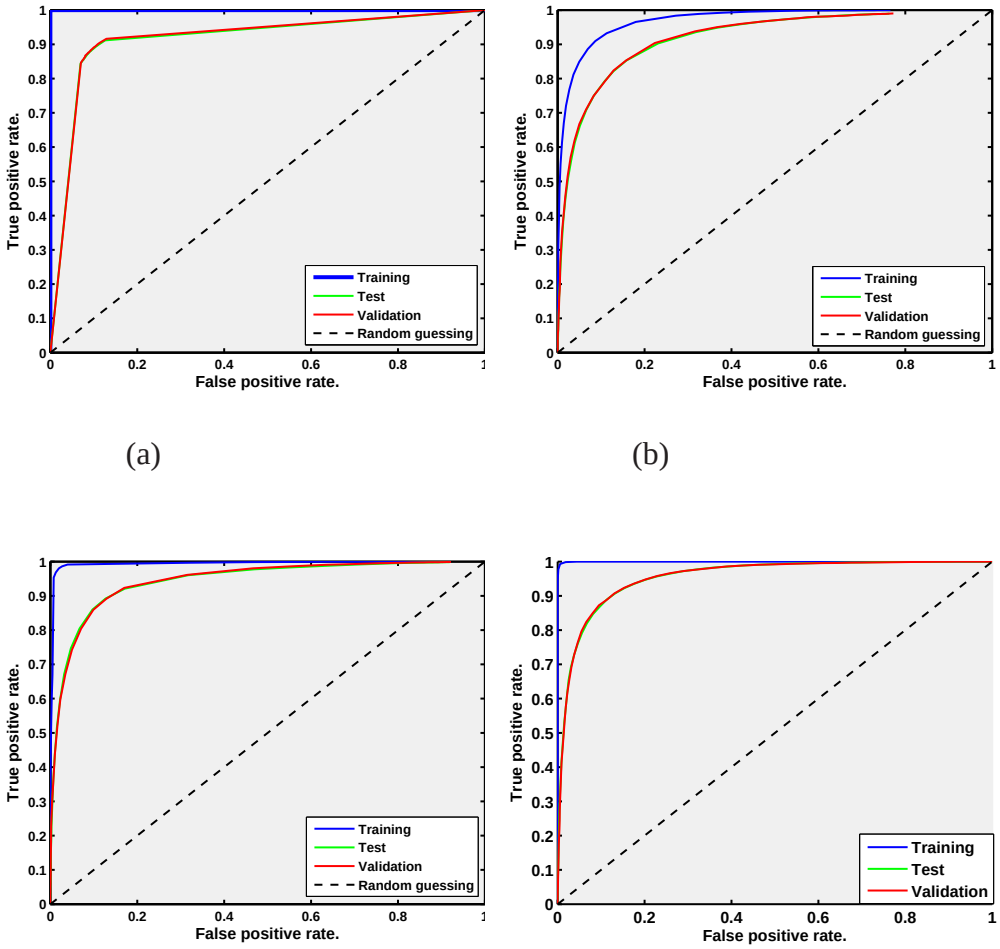


Figure C.23: The ROC curves for the 320-D reconstruction data. (a) MLP classifier (AUC values: 1, 0.9145 and 0.9162); (b) RBF classifier (AUC values: 0.7367, 0.6838 and 0.6944); (c) SVM regression classifier (AUC values: 0.9142, 0.8633 and 0.8665) and (d) RVM classifier (AUC values: 0.9983, 0.9543 and 0.9542). Generally, the classifiers performs much better than random guessing (AUC: 0.5).

	Data set	Linear	KNN	MLP	RBF	SVM(Cla.)	SVM(Reg.)	RVM
TP	Training (438)	256.98	——	399.92	365.24	——	399.16	366.70
	Test (345)	197.01	**276.93**	**282.25**	268.86	259.28	**277.14**	**276.23**
	Validation(99)	54.93	**77.28**	**78.75**	75.16	72.49	**77.69**	**76.85**
FP	Training (0)	168.33	——	51.64	72.89	——	40.98	76.40
	Test (0)	134.2	56.36	66.97	67.19	87.61	58.03	67.32
	Validation (0)	33.34	13.79	16.67	16.54	21.20	14.13	18.50
TN	Training (495)	321.57	——	438.26	417.01	——	448.92	410.61
	Test (392)	259.25	**337.09**	326.48	326.26	305.84	**335.42**	324.59
	Validation (91)	61.31	**80.86**	77.98	78.11	73.45	**80.52**	77.87
FN	Training (0)	186.12	——	43.18	77.86	——	43.94	79.29
	Test (0)	146.54	66.62	61.30	74.69	84.27	66.41	68.86
	Validation (0)	40.42	18.07	16.60	20.19	22.86	17.66	16.78

Table C.2: The ROC parameters of prediction models using the 2-D Euclidean distance.

	Data set	Linear	KNN	MLP	RBF	SVM(Cla.)	SVM(Reg.)	RVM
TP	Training (438)	286.34	——	383.33	365.86	——	389.51	362.65
	Test (345)	223.64	249.64	264.24	272.22	**283.62**	**278.55**	276.70
	Validation(99)	61.56	70.80	74.06	75.52	**79.14**	**77.46**	76.92
FP	Training (0)	175.24	——	75.62	123.96	——	90.28	80.45
	Test (0)	140.91	114.04	91.66	105.10	106.70	94.17	66.85
	Validation (0)	34.15	26.99	22.44	26.03	26.80	23.03	18.43
TN	Training (495)	314.66	——	414.28	365.94	——	399.62	360.69
	Test (392)	252.54	279.41	301.79	288.35	**286.75**	**299.28**	288.89
	Validation (91)	60.50	67.66	72.21	68.62	**67.85**	**71.62**	68.34
FN	Training (0)	156.76	——	59.77	77.24	——	53.59	129.21
	Test (0)	119.91	93.91	79.31	71.33	59.93	65	104.56
	Validation (0)	33.79	24.55	21.29	19.83	16.21	17.89	26.31

Table C.3: The ROC parameters of prediction models using the 2-D Minkowski distance.

	Data set	Linear	KNN	MLP	RBF	SVM(Cla.)	SVM(Reg.)	RVM
TP	Training (438)	403.86	——	**443.10**	404.78	——	399.16	**440.01**
	Test (345)	283.86	236.45	**300.33**	282.01	310.30	277.14	**305.57**
	Validation(99)	78.84	66.11	**83.34**	78.97	87.05	77.69	**84.79**
FP	Training (0)	26.58	——	0	42.13	——	40.98	3.09
	Test (0)	37.69	24.04	36.18	50.34	30.70	58.03	37.98
	Validation (0)	9.16	5.77	8.96	11.95	7.50	14.13	10.56
TN	Training (495)	463.32	——	**489.90**	447.77	——	448.92	**487.02**
	Test (392)	356.10	369.41	**357.27**	343.11	362.75	335.42	**348.78**
	Validation (91)	85.61	88.88	**85.69**	82.70	87.15	80.52	**83.86**
FN	Training (0)	39.24	——	0	38.32	——	43.94	2.88
	Test (0)	59.35	107.10	43.22	61.54	33.25	66.41	44.67
	Validation (0)	16.39	29.24	12.01	16.38	8.30	17.66	10.79

Table C.4: The ROC parameters of prediction models using the 320-D original data.

	Data set	Linear	KNN	MLP	RBF	SVM(Cla.)	SVM(Reg.)	RVM
TP	Training (438)	403.86	——	**443.10**	403.09	——	434.35	**440.01**
	Test (345)	283.86	273.67	**300.47**	282.15	310.30	295.41	**305.57**
	Validation(99)	78.84	76.07	**83.72**	78.52	87.05	81.95	**84.79**
FP	Training (0)	26.58	——	0	42.56	——	9.76	3.09
	Test (0)	37.69	31.38	34.78	50.55	30.70	38.21	37.98
	Validation (0)	9.16	7.68	8.49	12.18	7.50	9.38	10.56
TN	Training (495)	463.32	——	**489.90**	447.34	——	480.14	**487.02**
	Test (392)	356.10	362.07	**358.67**	342.90	362.75	355.24	**348.78**
	Validation (91)	85.61	86.97	**86.16**	82.47	87.15	85.27	**83.86**
FN	Training (0)	39.24	——	0	40.01	——	8.75	2.88
	Test (0)	59.35	69.88	43.08	61.40	33.25	48.14	44.67
	Validation (0)	16.39	19.28	11.63	16.83	8.30	13.40	10.79

Table C.5: The ROC parameters of prediction models using the 320-D reconstruction data.

D New data study

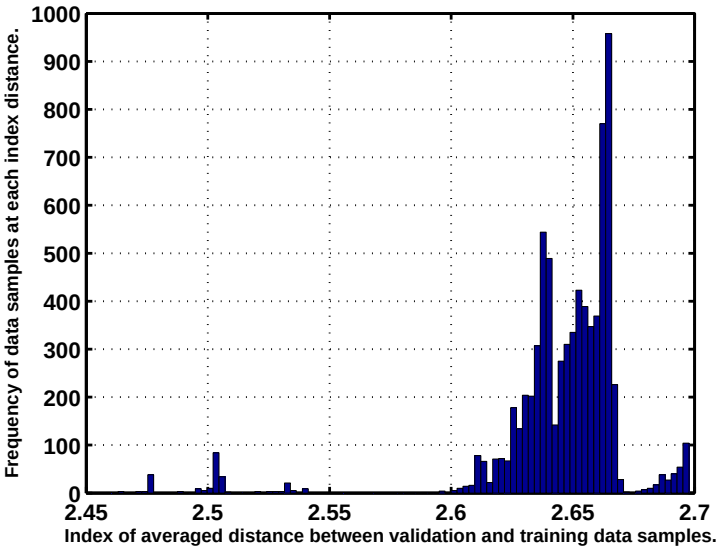
D.1 Visualisation results

According to the statistical histograms shown in Figure 6.1 for studying the most extreme situation where data samples are completely different from the training dataset, the nine data samples are generated and listed in Table D.1.

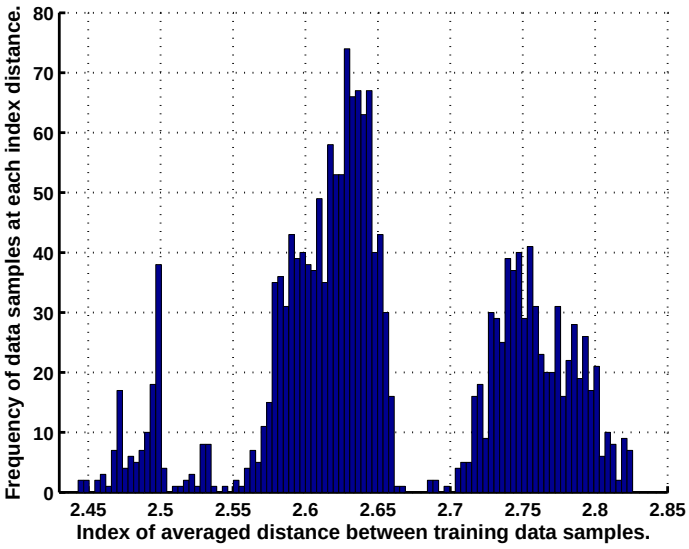
Index	**DNA (5'-3')**	a_2	a_{-1}	a_3	a_6
1	AAAt	A	C	C	C
2	AAAt	I	F	F	F
3	CAAt	Y	Y	Y	C
4	GCCa	L	Y	W	Y
5	CCCc	L	M	M	Y
6	ATAc	I	M	F	W
7	TAAa	E	P	R	M
8	CGGa	V	W	I	M
9	GATt	P	F	L	M

Table D.1: Structure informations of the nine generated data samples. According to the histograms shown in Figure 6.2, the nine data samples which are completely different from the training dataset (database DB1) are generated. The DNA sequence are presented in the 5'-3' order, and the listed amino acids labels at position 2, -1, 3 and 6.

In Subsection 6.2.1, the visualisation results of the database DB2 based on the Euclidean metric has been discussed. Figure D.1 plots the histogram of the dissimilarities between the validation and training datasets in data space. In Figures D.2 and D.4(a), the visualisation results of the test dataset which using the Minkowski metric as the dissimilarity measure, and relevant histogram of the dissimilarities in the input space are are presented, respectively. Moreover, the representation results of the validation dataset (DB3) based on the Minkowski metric and relevant histogram are also provided in Figures D.3 and D.4(b).

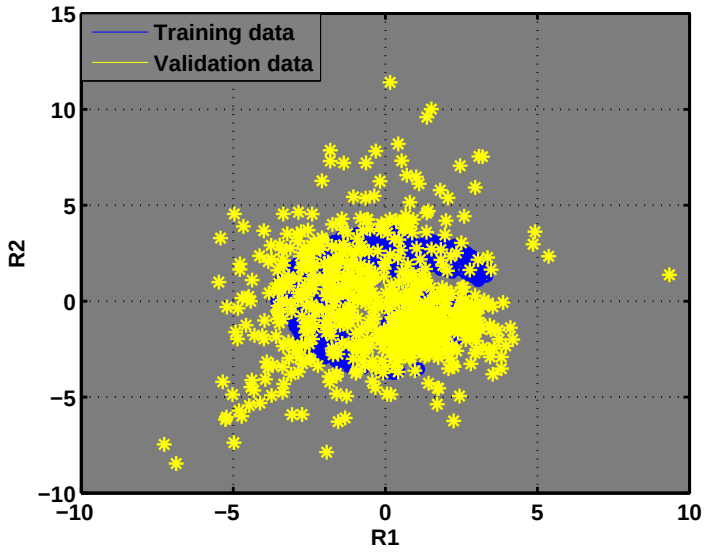


(a) Histogram of the averaged Euclidean distance between each validation data samples and the training dataset in the data space.

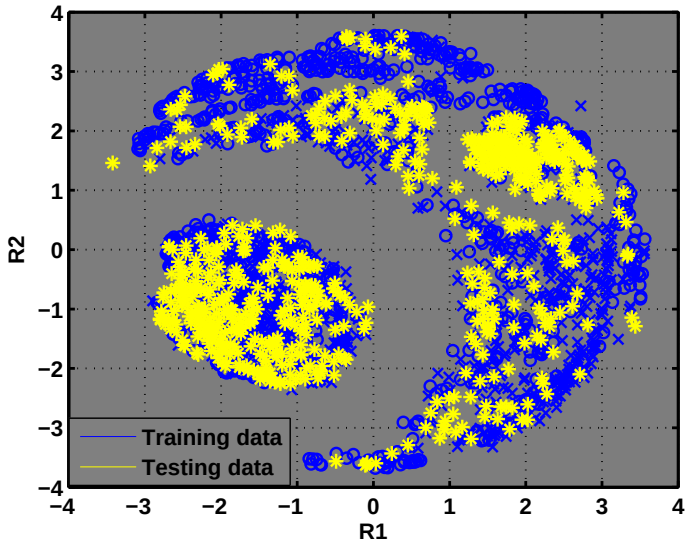


(b) Histogram of the averaged Euclidean distance between the training dataset in the data space.

Figure D.1: Histogram of the averaged Euclidean distance between the validation and training datasets. The distance changes from 2.45 to 2.7, and data samples are mainly in the ranges: 2.6 to 2.7. (a) is the histogram of the averaged distance between each test data samples and the training dataset. (b) as a reference plots the histogram of the averaged distance between each test data samples and the training dataset.

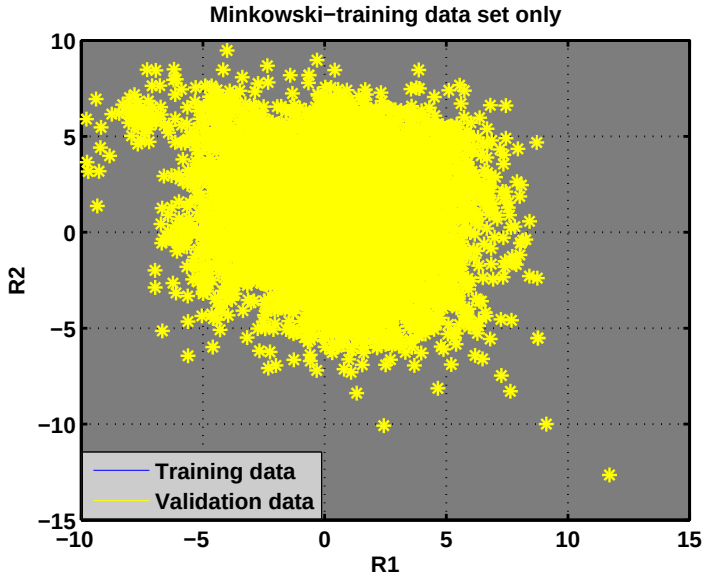


(a) Visualisation result of the test dataset which only the training dataset has been trained. Similar as Sub-Figure 6.7(a), some test data samples are projected external to the main visualisation area.

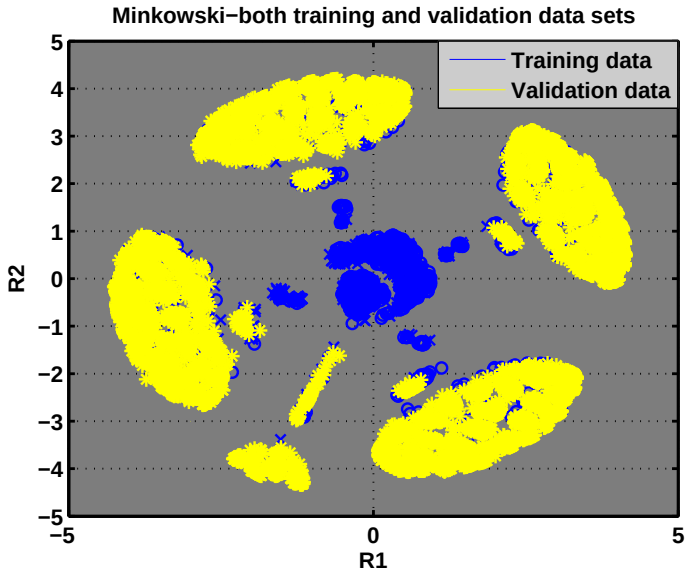


(b) Visualisation result from re-training on the whole dataset. In the figure, all data samples can be projected into the main visualisation area. The test data samples are projected to the clusters which have same amino acid colour coding and similar structure information.

Figure D.2: The visualisation results of the test dataset based on Minkowski metric. Data samples in bule colours are from the training dataset, yellow colours represent the test data samples. (a) is the result only use the training dataset to train the visualisation model. (b) is the visualisation result which both the training dataset and the validation dataset have been trained.

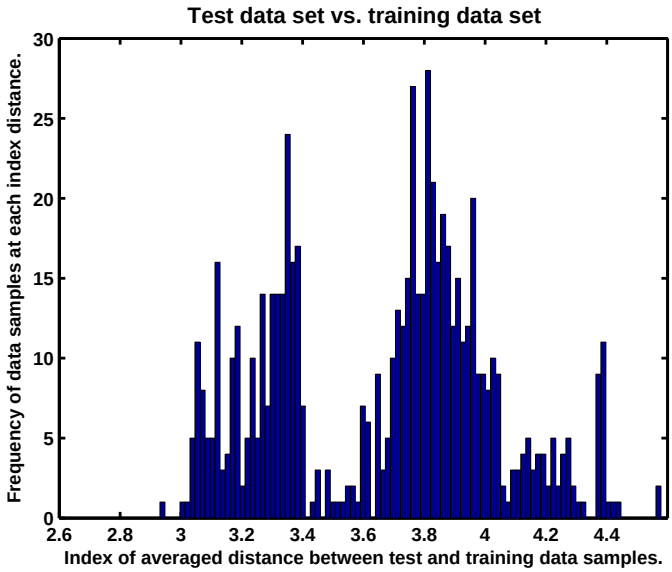


(a) In the figure, as the validation dataset is much larger than the training dataset, the training dataset are completely overlapped by the projected validation data samples.

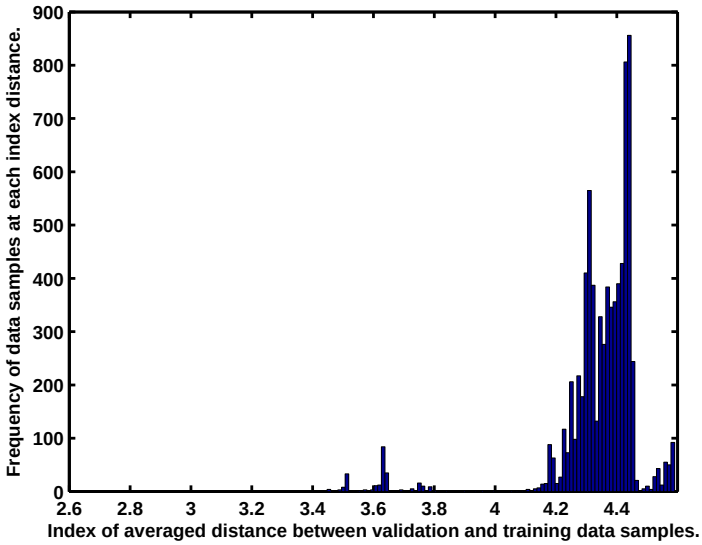


(b) Similar as Sub-Figure 6.10b, while the model is re-trained on the whole dataset, the projected validation dataset can be clustered into the relevant groups. The training dataset represented in the centre of the main area have different structure features from the validation data samples.

Figure D.3: Visualisation results of validation dataset (database DB3) by using the Minkowski metric in the data space. (a) is the result only use the training dataset to train the visualisation model; (b) is the result use both the training and validation datasets to train the NeuroScale model.



(a) Histogram of the averaged Minkowski distance between the test and training datasets.



(b) Histogram of the averaged Minkowski distance between the validation and training datasets.

Figure D.4: Histogram of the averaged Minkowski distance between different datasets. (a) is the histogram of the averaged distance between each test data samples and the training dataset. The distance changes from 2.8 to 4.4, and data samples are mainly in two distance ranges: 2.8 to 3.4 and 3.4 to 4.4. (b) the histogram of the averaged distance between each validation data samples and the training dataset. The distance changes from 3.4 to 4.6, and data samples are mainly in the distance range: 4 to 4.6.