N-gram methods of analyzing DNA sequence

by

© Charles Chilaka

A thesis submitted to the

School of Graduate Studies

in partial fulfillment of the

requirements for the degree of

Master of Science

Interdisciplinary Program in Scientific Computing

Memorial University of Newfoundland

June 2015

St John's Newfoundland & Labrador

Abstract

An DNA sequencing microarray experiment produces a 4 x N data matrix, comprising the signal strength of experimental DNA / reference oligonucleotide binding for each of four possible bases (A,C,G,T) at each of N positions. The highest of the four signals at each position is expected to result from the perfect match, and hence be the correct base call. Variation in absolute and relative signal strength may interfere with reliability of base calling. Variable base composition of the reference oligonucleotides influences this in ways that are not fully understood.

I used an n-gram representation of oligonucleotides in a neural network analysis to predict normalized signal intensities from an Affymetrix DNA sequencing experiment. For a DNA oligonucleotide, an n-gram can take on 4^n values, e.g., a 1-gram uses the frequencies of each base, a 2-gram the di-nucleotide frequencies, and so on. Neural networks use a variable number of neurons in the hidden layer to create a correspondence between an input data set (divided into Training, Validation, and Test sets)

and an output target set.

I used a data set reduced from a sequence of 15,392 bases to 594 lines. I used 1- and 2-grams and their composite, with 20 to 40 neurons in the hidden layers of the neural network. For all models, the base with normalized value of 1.0 (that with the highest absolute value) was predicted for 100% of times. For 1-grams compared with the composite, regression values increased from 0.9898 to 0.9918, and measures of performance plots improved (decreased) from 3.195×10^{-3} to 2.525×10^{-3}. For a 1- and 2-gram composition with 30 neurons in the hidden layer of the neural network, the diagonal of the confusion matrix had a high percentage value of 99.8%. Receiver Operating Characteristic (ROC) curves showed points in the upper-left corner, a sign of good test.

The analysis suggests that neural network analysis of oligonucleotides as higher-order n-grams could be used to predict intensities in an Affymetrix or any other DNA sequencing microarray experiment.

Acknowledgement

This thesis would not have been possible without advice from my supervisors, Professor Wolfgang Banzhaf (Computer Science), Professor Steve Carr (Biology) and Professor Nabil Shalaby (Mathematics). You stood by me in my trying moments and made this possible through your painstaking indept analysis, reviews and accepting my shortcomings. May you be blessed by God.

I also thank my wife, Alice and children, Uriel, Ecclesius and Naomi for providing the needed comfort and distractions. I appreciate the effort of my great friends, Feng Wu and Brad Sheppard who provided some insights on Python programming. I also thank Peter Gill for critically reading the manuscript.

To High Chief and Chief (Mrs) Emmanuel Iroh, Dikeoha 1 of Umunama who provided much needed support in times of great need, may the good Lord enlarge your coast.

To my mum, Onyenyinyedi and Daa Cy, my amiable Mother-in-Law I love you both.

To God be all the glory.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction: Biology of DNA

Many classical works have been done to better understand the basic building blocks of human life via what is called Deoxyribonucleic acid, DNA. DNA is a double-helical molecule (dsDNA) where each helix comprises an aperiodic series of nucleotides (bases) known as Adenine, Cytosine, Guanine and Thymine designated by the sequence of letters A, C, G, and T respectively. Adenine and Guanine are two-ringed molecules and are collectively known as Purines while Cytosine and Thymine are single-ringed molecules known as Pyrimidines. These bases occur on the inside of the molecule. The two helices (strands) each have a directionality designated $5' - 3'$ and run in opposite (antiparallel) directions.

When biologists analyze the interaction of species and the function of cells, researchers

depend greatly on the collection and analysis of large DNA datasets to understand the interactions between species. DNA sequences are not numeric in nature. Their conversion to numerical values enables the application of powerful digital signal processing techniques to them. An area of contemporary research is Bioinformatics where we use software tools to understand the arrangement of sequences and its' effects on the condition of any living organism. Bioinformatics combines mathematics, computer science, biology, chemistry and statistics and tries to solve biological problems with computers.

The central structural feature of DNA is the base pairing and the sequence itself is the central information content. These pairings are that Adenine (A) in one strand is always paired with a Thymine (T) in the other strand, and are held together by two hydogen (H) bonds. Also, Cytosine (C) in one strand is always paired with a Guanine (G) in the other strand, and are held together by three hydrogen (H) bonds. The A/T and G/C pairs are the same size and shape, which allows their aperiodic arrangement in the DNA molecule [4], [106], [117].

The bioinformatic content of a single-stranded DNA (ssDNA) can be specified by writing out the sequence of base letters and identifying the 5' and 3' ends. Because of the pairing relationships, the sequence and direction of either strand is implicit in the other. The biological process of replication of DNA allows the two strands to

separate and each makes a base complementary copy of itself, so as to produce two molecules from the original one by semiconservative process as shown in Figure 1.1. The knowledge of the sequence of one strand allows that strand to be characterized mathematically and thus helps in the bioinformatic analysis of a DNA molecule.

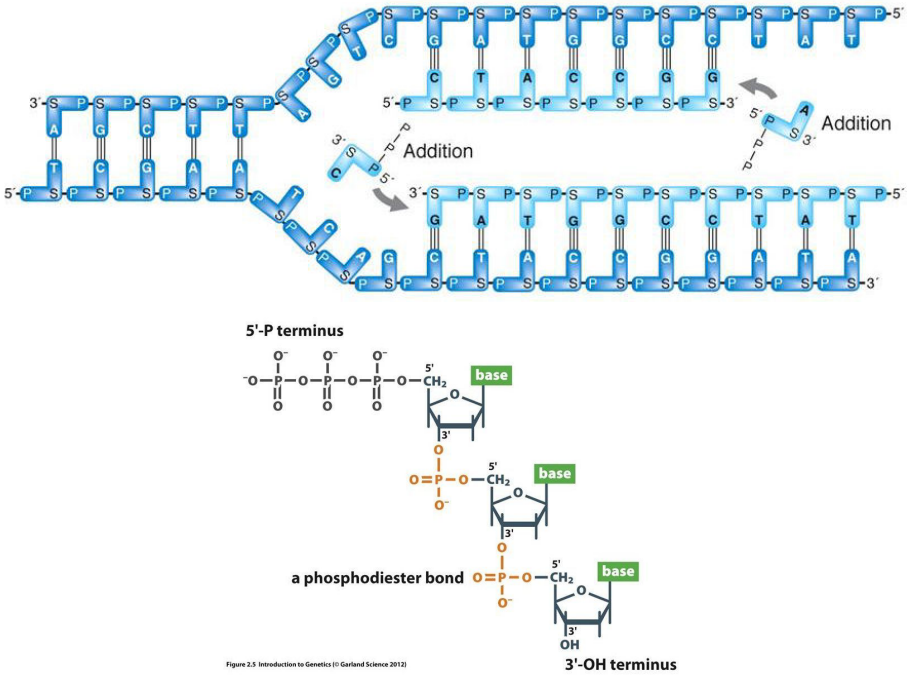


Figure 1.1: Replication and directionality in a DNA sequence adopted from [106]

In 1975, Frederick Sanger showed a novel method of DNA sequencing known as dideoxy sequencing [4]. His method exploited the base-pair complements together with an understanding of the basic enzymology and biochemistry of DNA replication. He used radioactively-labelled dideoxynucleic acids, ddNTP's which acted as

terminators in chain extension. By doing separate reactions with ddA, ddC, ddG and ddT, for the four nucleotides respectively, a complete set of fragments is produced that differ by ±1 base-pair. The sequence is then read as a 'ladder" as shown in Figure 1.2.

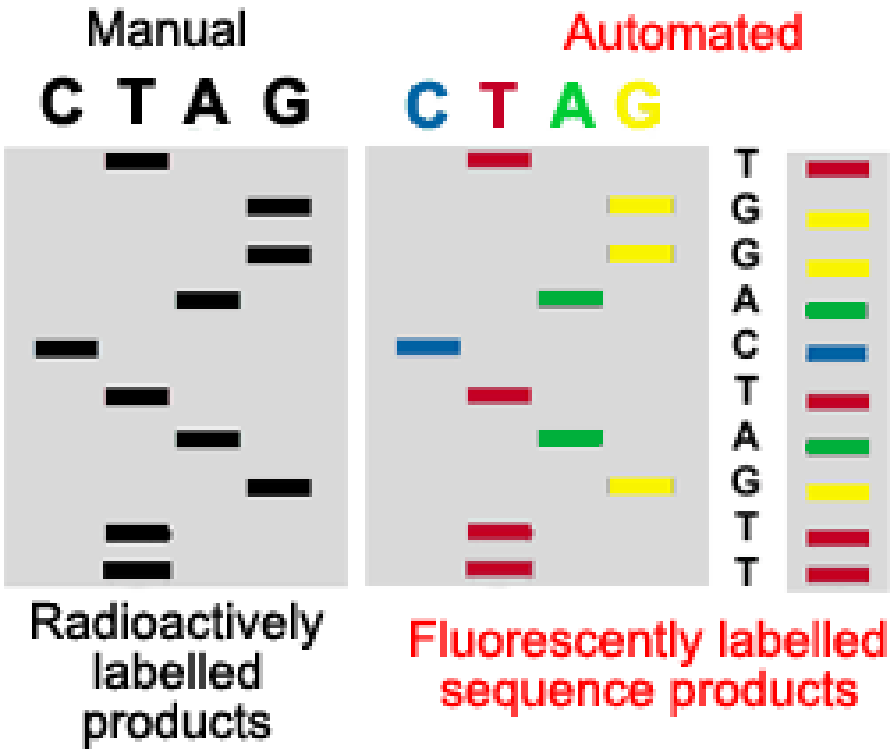


Figure 1.2: Radioactively and Fluorescently labeled sequences

Sanger Sequencing monitors the replication of an existing ssDNA as its single stranded complement, in the presence of modified bases so that the order of addition can be tracked.

This method was automated by the late 1980s with the use of fluorescently-labeled ddNTP's and gave rise to the "molecular revolution" of the 1990's. The development of polymerase chain reaction also contributed to this revolution. The break-through whole genome sequencing projects such as HUG0 in 2004 enabled researchers to sequence the human genome with 3 billion base pairs. In automated sequencing, added bases are modified so as to fluoresce at different wavelengths in the green spectrum, which are then represented as different 'pseudocolours", [43]: "green" for Adenine, "blue" for Cytosine, "yellow/black" for Guanine and "red" for Thymine as also shown in Figure 1.2. The sequence of the replicated strand may then be "read" directly as a succession of differently colored terminal bases, interpreted as sequences of ACGTs. This sequence can then be presented as a reverse complement corresponding to the order of bases in the original ssDNA template. Sanger sequencing ideas remain in use as the standard sequencing technology for many projects.

1.1 DNA sequencing techniques

We give a discussion of the use of microarrays in sequencing.

1.1.1 From Sanger sequencing to next generation sequencing

Sanger sequencing is gradually been supplanted by a variety of so called "Next-Generation" sequencing technologies of which those based on microarrays (DNA chip) technology are among the earliest [101]. One of those based on this technology is Affymetrix [42], a pioneer in microarray technology and a leader in genomics analysis. Its mission statement is *"to enable the translation of biological knowledge into routine practice"*. The Affymetrix chip is based on molecular "Sequencing by Hybridization" which is distinguishable from mathematical "Sequencing by Hybridization". Mathematical sequencing by hybridization is explained later using graph theory concepts.

A DNA chip is a small piece of silicon glass ($1cm^2$) to which a large number of synthetic, single-stranded DNA oligonucleotides have been chemically bonded. These oligonucleotides function as DNA probes and stick (anneal) selectively only to those DNA molecules whose nucleotide sequences are exactly complementary. A DNA chip can be used as Variant Detector Arrays to identify DNA sequences that differ by Single Nucleotide Polymorphisms (SNP's) [101].

A SNP is a variant base-pair that occurs at a homologous position in different genomes of the same or different species and constitute a major class of the naturally-occuring genetic variation among organisms. They are sometimes responsible for genetic diseases that result from a SNP that causes amino acid change which in turn results in a

modified protein with abnormal physiological function. An example is the Sickle Cell Anaemia, an inherited blood disorder, which is caused by $A \rightarrow T$ SNP in the second position of the sixth codon of the coding strand of the gene for Beta-hemoglobin which results in the substitution of Valine for Glutamine in the protein. This results in a hemoglobin molecule that can form crystals.

Figure 1.3, [101] shows a set of DNA oligonucleotides that differ only at the last position, corresponding to a known SNP site in the genome.

Fluorescently-tagged genomic DNA fragments anneal preferentially to those oligos with which they are perfectly complementary: In Figure 1.3, an allele [1] with a T SNP binds to the A oligo and an allele with a C SNP binds to the G oligo. A computer reads the position of the two fluorescent tags and identifies the individual as a C/T heterozygote [2] [84]. Similarly, the single spots in the other three columns of the 4×4 VDA indicate that the individual is homozygous [3] at the three corresponding SNP positions. The 4×4 array fits into one corner of a 256-oligo VDA chip for 64 SNPs (lower right). Current generation of microarrays can accommodate hundreds of thousands of oligos.

[1] Different forms of the same gene,located in the same position along the chromosome in each individual

[2] Individuals that have two different alleles of a particular gene

[3] Individuals that have two copies of the same specific allele of a particular gene

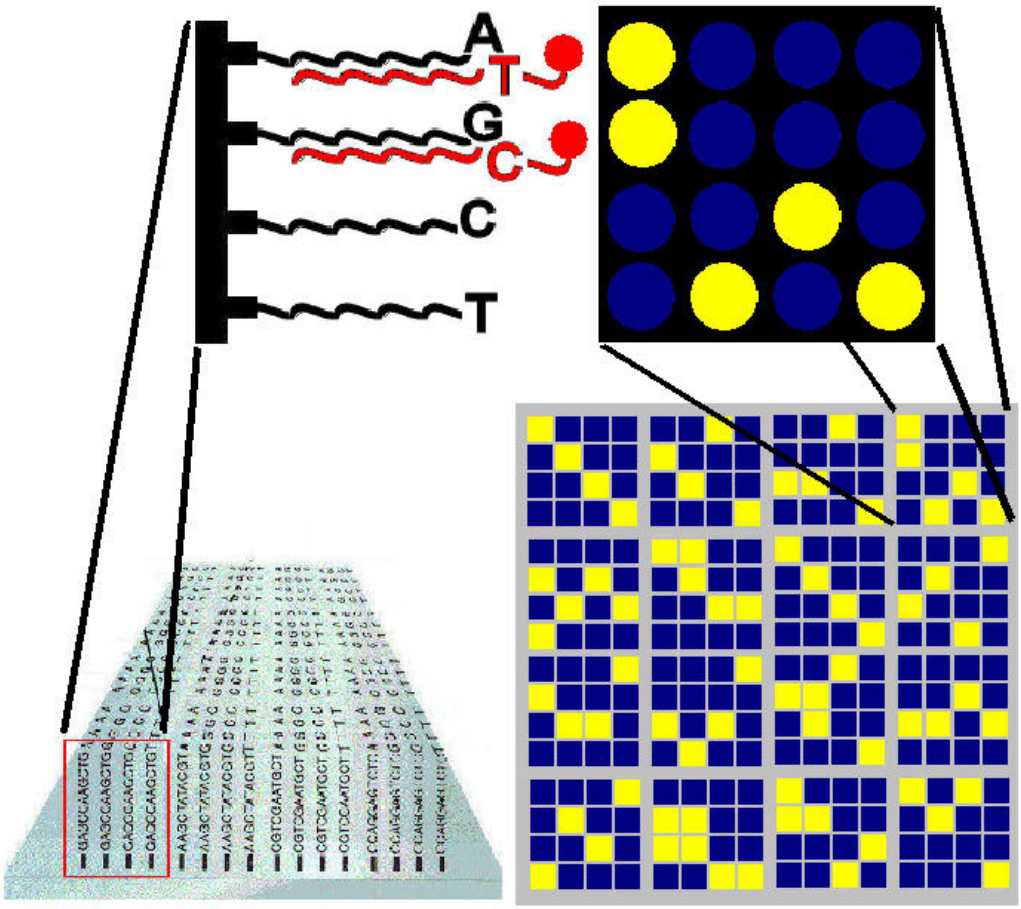


Figure 1.3: DNA microarrays as Variant Detector Arrays adopted from [101]

Variant Detector Arrays (VDAs) do not measure gene expression but rather variation in SNPs among samples of interest. VDAs rely on the ability of a ssDNA in the experimental sample to recognize and bind to its perfect oligonucleotide complement. A refinement of VDA microarrays is to evaluate, not just known SNPs, but all potential SNPs within a particular gene region.

1.1.2 Detection of SNPs using the DNA chip

The array design of a Genechip normally utilizes two types of probes, [22]: Perfect match probes which are probes that have complete complementarity to their target sequence and mismatch probes which are probes with a single mismatch to the target, centered in the middle of the oligonucleotide which is usually of length 25. For every perfect match feature, a mismatch feature is included which is identical to the perfect match sequence, except for a nucleotide variant on the 13th nucleotide which is the central nucleotide. Common SNP detection experiment using DNA chip works as follows [101]:

- A set of 4 artificial 25-base long ssDNA oligonucleotides are tethered at one end to the four adjacent spaces (quartet) on the chip.
- The length 25 is chosen such that a random match is extremely unlikely.
- In the full experiment, there will be many thousands of such quartets.
- In each quartet, one of the oligonucleotides is an exact match to the DNA sequence of the reference individual.

- The other three differ from the reference at the 13th position, by the three possible SNP variants at that position e.g.: C, G and T versus A.

The schematic representation of a DNA re-sequencing microarray is shown in Figure 1.4, [101] for four successive base positions. It shows molecular sequencing by hybridization for four successive positions as overlapping Variant Detector Arrays.

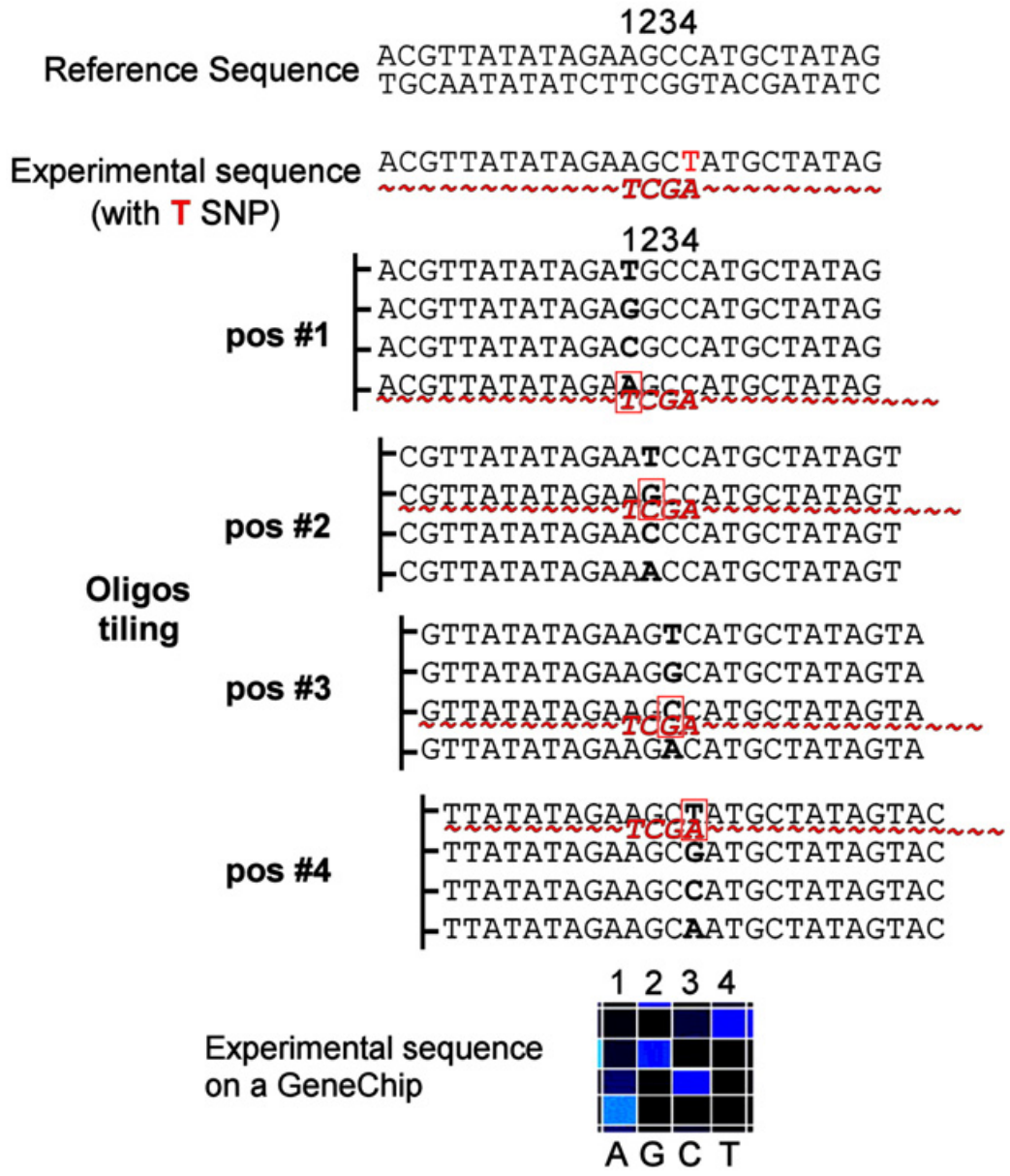


Figure 1.4: Schematics of DNA re-sequencing microarray experiment adopted from [101]

In Figure 1.4, a reference DNA sequence is represented in a series of overlapping ("tiled") oligonucleotide probes, each of length 25bp. For each oligo, three variants are included that vary in the middle (13th) base, one for each of the three alternate code letters. In the example shown in Figure 1.4, four successive bases in the reference DNA sequence are **AGCC**: the four alternate oligos tiling the first position are (top to bottom) **TGCC**, **GGCC**, **CGCC**, and **AGCC**. The same arrangement occurs for oligos tiling the next three positions; the order of the variant bases in each set of oligos is constant (**T,G,C,A** = 1st, 2nd, 3rd, 4th rows).

Consider an experimental DNA sequence with SNP at the last position **AGCT**. The sequence of complementary strand (...**TCGA**...) is an exact match for only one of the four variant oligos at each tiling position. Mismatch at this position most strongly effects binding: the absolute degree of binding is measured at each oligo, and computer imaging of the microarray shows this as a more or less intense pseudocolour. In this case preferential annealing to the 4th, 3rd, 2nd and 1st oligos at four successive positions indicates that the original (complementary) experimental sequence is **AGCT**.

Generally, DNA from the individual of interest is prepared as small pieces of ssDNA, each fluorescently tagged at one end. This experimental DNA is allowed to hybridize to the DNA target quartet. The data of each experimental DNA/target quartet

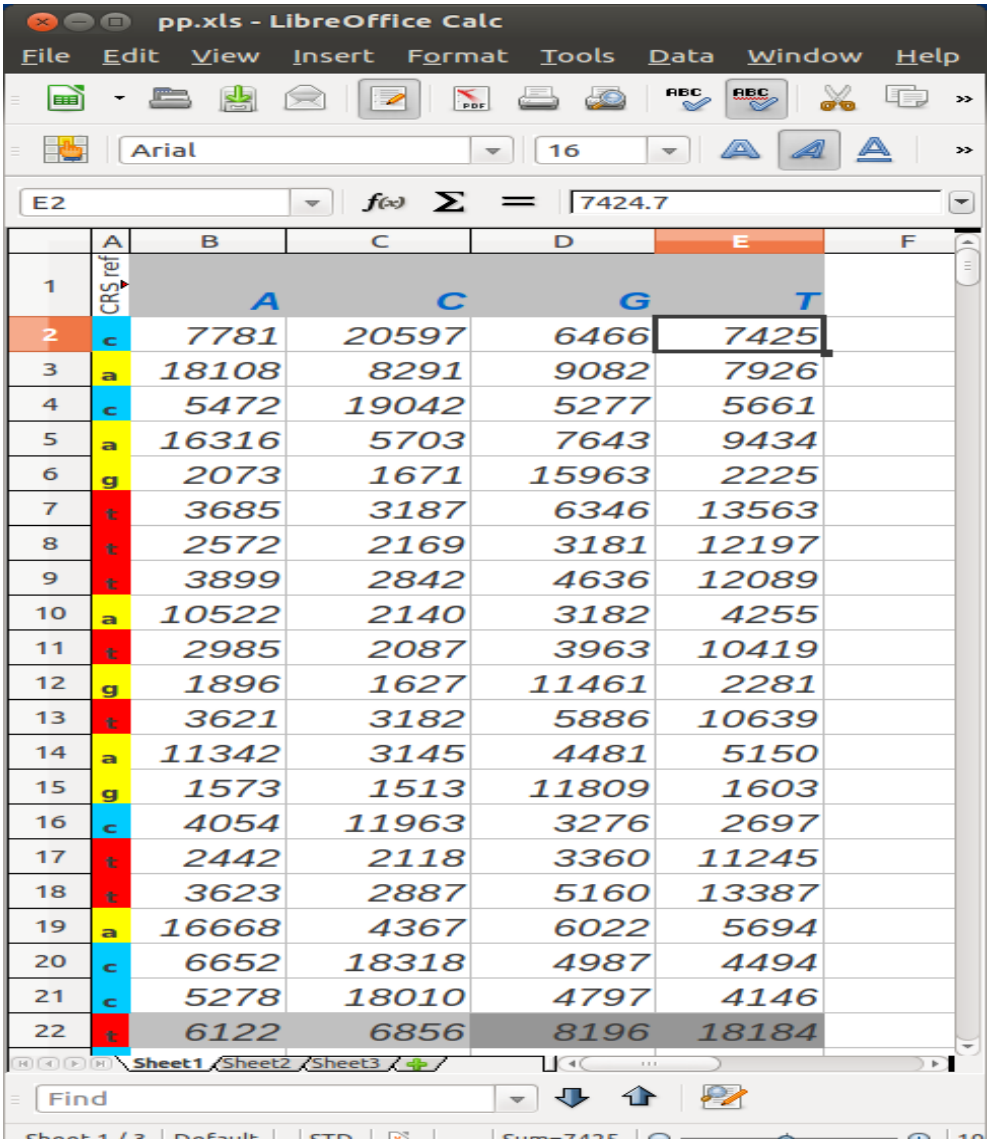
combination then consists of four numbers, indicating a fluorescence signal over the range of 10^0 to 10^5.

1.1.3 The Affymetrix Excel DNA spreadsheet

After an Affymetrix experiment for any form of hybridization, which involves, among other things, staining of a DNA array and washing, each array is scanned using the Affymetrix Genechip Scanner 3000 and analysed using Genechip DNA analysis software. The output for each array scanned consists of probe intensity values for each cell, corresponding to Adenine (A), Cytosine (C), Guanine (G) and Thymine (T) for both the sense and anti-sense strands. Note that a typical DNA strand is read backwards from the 5' end to the 3' end since DNA synthesis proceeds by the elongation of the primer chains always in the 5'-to-3' direction. Using an excel spreadsheet algorithm, calls are generated that summed the sense and anti-sense values for each position. My thesis concentrates on one of the strands but not both. A one strand screen shot page of an Affymetrix data set is shown in Figure 1.5 giving the call [4] and the respective signal intensity outputs (profiles) of all the four

[4]base with highest intensity along each row according to the Cambridge Reference Sequence which is a reference sequence for human mitochondrial DNA. It is deposited in the GenBank NCBI database under accession number NC_012920

nucleotides.



	A	B	C	D	E	F
1	CRS ref	A	C	G	T	
2	c	7781	20597	6466	7425	
3	a	18108	8291	9082	7926	
4	c	5472	19042	5277	5661	
5	a	16316	5703	7643	9434	
6	g	2073	1671	15963	2225	
7	t	3685	3187	6346	13563	
8	t	2572	2169	3181	12197	
9	t	3899	2842	4636	12089	
10	a	10522	2140	3182	4255	
11	t	2985	2087	3963	10419	
12	g	1896	1627	11461	2281	
13	t	3621	3182	5886	10639	
14	a	11342	3145	4481	5150	
15	g	1573	1513	11809	1603	
16	c	4054	11963	3276	2697	
17	t	2442	2118	3360	11245	
18	t	3623	2887	5160	13387	
19	a	16668	4367	6022	5694	
20	c	6652	18318	4987	4494	
21	c	5278	18010	4797	4146	
22	t	6122	6856	8196	18184	

Figure 1.5: A screen shot of Affymetrix data

The Probe set is the quartet of 25-mer oligonucleotides corresponding to a particular base in the numbered Reference (Ref) DNA sequence. There are four signals in each quartet, corresponding to variants at the 13th position: base order on the first (forward) strand is T G C A. At each position, the presumptive base call is the one with highest hybridization to the experimental DNA and therefore the highest signal. Signal to noise ratio dS/N is a measure of confidence in the call and is the difference between the highest (max) and second-highest (2nd) signals, divided by the sum of signals (sum), [103]. Thus $dS/N = (max–2nd)/Sum$. The strongest signal needs to be checked and compared with other signals.

In practice, only the experimental ssDNA that is exactly complementary in one member in each quartet will stick and show a fluorescence signal. The experimental ssDNA will stick to the exact match most strongly and to the other three less strongly, at various degrees of intensity. The identification of the base present in the experimental DNA is then a matter of identifying the strongest signal intensity, which is >> 99 percent of cases. Empirical rules for mismatch between strands, and (or) SNP detection are also given in [103].

These signal intensities are what I want to develop an improved mathematical method for their prediction via their normalized values. I seek to establish the relationship between the arrangement of the bases that make up one strand of DNA and the nor-

malized signal intensities recorded by the Affymetrix Genechip.

The rest of the thesis is organized as follows: Further parts of Chapter One deal with more definitions and examples of relevant concepts and techniques for DNA sequencing. In Chapter Two, I present some of the many numerical representations and visualizations of DNA sequence and introduce the notion of n-gram and some of its distance measurements. Some correlation measurements amongst bases that make up the DNA sequence are also discussed. Chapter Three presents ideas from data mining and Artificial Neural Networks. Chapter Four deals with sequence encoding schema, evaluation functions and diagrams from our various simulations using the Matlab neural network and Chapter Five deals with results, conclusion and some ideas for future research.

1.2 Definitions and examples

I formally present the definitions of terms associated with DNA and graph theory which are of importance in this thesis. All the definitions given can be found in biology and graph theory texts, [4, 5], [11], [23], [26], [68], [76], [84], [106], [117, 118].

Deoxyribonucleic acid, DNA : This is the hereditary material in almost all biological organisms. It is made of four (4) chemical bases known respectively

as Adenine (A), Cytosine (C), Guanine (G) and Thymine (T). These bases are arranged so that they are complementary i.e. A is complementary to T, and G is complementary to C.

Genome : The entire DNA of a living organism is called its genome. It is the complete genetic material of an organism. The size of the genome is the total number of base-pairs of one copy (e.g. half of the total for a diploid [5] organism).

Base-pair : A base-pair consists of two nitrogenous bases (Adenine and Thymine or Guanine and Cytosine) held together by weak bonds. Two strands of DNA are held together in the shape of a double helix by the bonds between base-pairs.

Base sequence : Base sequence is the order of nucleotides in a DNA molecule.

Sequencing : Sequencing refers to determining the order in which the bases that make up a DNA sequence occur in any given chain.

Nucleotide : A nucleotide is a subunit of DNA consisting of nitrogenous bases (Adenine, Guanine, Thymine or Cytosine), a sugar-phosphate backbone (deoxyribose in DNA). Thousands of nucleotides are linked to form a DNA molecule.

Hybridization : This refers to the binding between complementary single-stranded

[5] any cell with two chromosome sets

DNA molecules.

Example 1. *A probe ATTACCGT will hybridize with a target CTAATGGCAAT since it is complementary to TAATGGCA.*

Alphabet : An *alphabet* L is defined as a finite, non-empty set of symbols.

String : A *string* is a finite sequence of symbols from a given alphabet. A string is primitive if no character appears more than once. A typical DNA strand made of only the four letters representing the nucleotides is primitive since it can have only A, C, G or T as the basic strand.

Empty string : An *empty string* is a string that does not contain any symbols.

Length of string : The *length* $|v|$ of a string v is the number of positions for symbols in the string.

Substring : A string $u = a_1a_2...a_m$ is a *substring* of string $v = b_1b_2...b_n, m \leq n$ if $a_1a_2...a_m = b_1b_{i+1}...b_{i+m-1}$ for some i, $1 \leq i \leq n$

Superstring : A superstring of a set of strings $S = s_1, s_2, ..., s_n$ is a string s containing each s_i, $1 \leq i \leq n$.

Example 2. *The string S= ATCCCACAGTCCAGT with 3-spectrum*

S= ATC, CCA, CAG, TCC, AGT has the string s= ATCCAGT as a super-string.

Spectrum (S,n) : The unordered multiset of all possible (L-n+1) n-mer substrings in a string S of length L.

Cardinality of a Spectrum : The cardinality of a spectrum is given by $|L-n+1|$ where L is the length of the sequence and n is the length of the substring been investigated.

Example 3. *For the sequence ATAGGCAAA, its' length is 9 and the 4-mers are ATAG, TAGG, AGGC, GGCA, GCAA, CAAA. Hence the cardinality of this given spectrum is* $9-4+1=6$.

Overlap (s_i, s_j) : This is the length of the maximal prefix of a string s_i that matches the suffix of a string s_j.

Example 4. *The overlap of the two DNA strings*

S_1=*TTTGGCATCAAATCTAAAGGCATCAAA and*

S_2 =*AAAGGCATCAAAGATGCCTTTGGTACA, is the string AAAGGCAT-CAAA of length 12.*

Edit distance : An edit distance also known as Levenshtein distance is the minimum number of elementary edit operations needed to transform one string

into another where the edit operations are insertion of a symbol, deletion of a symbol, substitution of one symbol for another.

Example 5. *The edit distance between the sequence $S_1 = TGCATAT$ and the sequence $S_2 = ATCCGAT$ is at least 4. These are as follows:*

Insert A at the front of S_1 ....ATGCATAT

Delete T in the sixth position of S_1....ATGCAAT

Substitute G for A in the 5th position of S_1....ATGCGAT

Substitute C for G in the 3rd position of S_1....ATCCGAT

This fourth operation resulted in the string ATCCGAT which is equal to S_2.

An advantage of edit distance computation is that it allows us to compare strings of different lengths [76].

Hamming Distance : Given two $l-$mers v and w, we compute the Hamming distance between them $d_H(v, w)$ as the number of positions that differ in the 2 strings.

Example 6. *The Hamming distance between the two strings $S_1 = ATTGTC$ and $S_2 = ACTCTC$ is $d_H(S_1, S_2) = 2$.*

One major shortfall of Hamming distance computation is that it rigidly assumes

that the i-th symbol of one sequence is already aligned against the ith symbol of the other sequence [23], [76].

Graph : A graph is a pair (V, E) of sets, V nonempty and each element of E an unordered pair of distinct elements of V. The elements of V are called vertices and the elements of E are called edges. A pseudograph is like a graph but it may contain loops and/or multiple edges. In DNA context, vertices represent strings of length say n (n-mers from the spectrum) and edges represent overlapping n-mers.

Weighted graph : A weighted graph is a graph G(V,E) together with the function $\theta : E \to [0, \infty]$. If e is an edge, the nonnegative real number $\theta(e)$ is called the weight of e. In other words, a weighted graph is a graph in which there is a nonnegative number associated with each edge.

Walk : A walk in a pseudograph is an alternating sequence of vertices and edges, beginning and ending with a vertex, in which each edge is incident with the vertex immediately preceding it and the vertex immediately following it.

Path : A path in a pseudograph is a walk in which all vertices are distinct.

Digraph : A directed graph (digraph) is a pair (V, E) of sets, V nonempty and each

element of E an ordered pair of distinct elements of V. The elements of V are called vertices and the elements of E are called arcs.

Cyclic digraph : A directed graph with at least one directed circuit is said to be cyclic. A directed graph is acyclic otherwise. In an acyclic digraph, there exists at least one source (a vertex whose in-degree is zero) and at least one sink (a vertex whose out-degree is zero).

de Bruijin graph : Given a I-mer spectrum S, let $\hat{S}$ denote the set of all n-1 mers in S. The de Bruijin graph for S is a directed graph $G = (V, E)$ with node set $V = \hat{S}$ and in which two nodes v and w are connected by a directed edge (v,w) if S contains an n-mer, whose $n-1$ prefix is v and whose $n-1$ suffix is w.

Hamiltonian path : This is a path in a graph that visits every vertex exactly once.

Example 7. *The sequence $S = ATGCAGGTCC$ with 3-spectrum given by ATG, AGG, TGC, TCC, GTC, GGT, GCA, CAG has a Hamiltonian path $H = ATGCAGGTCC$.*

Eulerian path : This is a path in a graph that visits every edge exactly once.

1.3 Hybridization and factors affecting it

Earlier, many methods existed for the sequencing of DNA but none has been proven to be capable of sequencing the full length of DNA sequences of most organisms at once until the advent of better sequencing technologies. Instead, to detect the existence of certain DNA segments in a sample, short DNA segments are sequenced and their positions mapped in the entire genome. One of the methods that have been used to do this complex task is called Sequencing by Hybridization, SBH, [6, 7], [23], [68], [76], [96]. Sequencing by Hybridization, SBH was proposed simultaneously and independently in the 1980's by E. Southern [28], Drmanac et al [87], Bains and Smith [113], Lysov et al [125]. It relies on the hybridization of an unknown Deoxyribonucleic acid, DNA fragment from a large array of short probes [76]. It is a non-enzymatic method of determining the order in which nucleotides occur on a DNA strand. In this method, unknown DNA is labelled and compared with known sequences. Hybridization occurs if there is a substring of the largest sequence that is complementary to the probe.

Hybridization has been known to be influenced by the following factors: temperature, ionic composition of the solution, the DNA sequence and its complexity, the length, and number of GC pairs present in the helix [1], [103], [121]. Homopolymers (uninterupted stretch of a single nucleotide) also affect hybridization via the signal

intensity of the nucleotide sequence. Moreover, there is a strong correlation between the overall probe intensity strength as measured by an Affymetrix Genechip for each position in the nucleotide sequence and the signal to noise ratio [102], [103]. Sensitivity and specificity are also two of the measurements of quality in any microarray experiment [6].

1.4 Sequencing by Hybridization as a mathematical construct

This form of Sequencing by Hybridization, SBH, requires one to obtain a string where all the n-mers that are substrings of that string are exactly the set of n-mers obtained. The ideas are mainly from graph theory where usage is made of short DNA sequences to represent vertices and overlapping DNA sequences as edges of a graph. In DNA context, vertices represent strings of length say l (l-tuples from the spectrum) and edges represent overlapping l-tuples (weights). There are Hamiltonian and Eulerian path approaches to this form of hybridization.

1.4.1 The Hamiltonian/Eulerian path approach

Recall that the Hamiltonian path of a graph is a sequence of vertices connected by edges that visits every vertex exactly once [26]. To determine such a path is difficult computationally, since we have to check all possible paths. The problem is known to be NP complete [68], [76]. Unlike the Hamiltonian path approach, however, the Eulerian path approach is relatively easy to handle. Recall, too, that the Eulerian path in a graph is a path that visits each edge exactly once. It is not too hard to prove that a connected graph has an Euler path if and only if the absolute difference of the in-degree and the out-degree for every vertex is 1 for at most 2 vertices and 0 for the rest [23], [26].

SBH can be reduced to a Shortest Superstring Problem (SSP): Given a set of strings $s_1, ..., s_n$, find the shortest string s such that each s_i appears as a substring of s. SBH provides information about the I-tuple (ordered set of I elements) but does not provide information about their positions. It gives the I-tuple composition of fragment and can be seen as a special case of SSP [68], [76].

Example 8. *Given the set of strings*

$S = AAA, AAB, ABA, ABB, BAA, BAB, BBA, BBB$. *The concatenation superstring is given by* $AAAAABABAABBBAABABBBABBB$ *and the shortest superstring*

is AAABBBABAA.

We can reduce the Shortest Superstring Problem, SSP, to a Travelling Salesman Problem (TSP) [26], which is to find a Hamiltonian cycle of least weight in a weighted connected graph by establishing overlaps between sequences [76].

Example 9. *Consider the two DNA strings:*

S=TTTGGCATCAAATCTAAAGGCATCAAA and

P=AAAGGCATCAAAGATGCCTTTGGTACA.

The overlap is the string AAAGGCATCAAA of length 12 and we construct a graph with n vertices representing the n strings $s_1, s_2, ..., s_n$. Insert edges of length overlap s_i, s_j between vertices s_i and s_j. We then find the shortest path which visits every vertex exactly once. This is the Travelling Salesman Problem as it reduces to almost a complete directed graph [26], [74]. In Sequencing by Hybridization, all the s_i have the same length, and all the I-tuples p and q overlap in last n-1 letters of p and first n-1 letters of q.

Example 10. *Consider the string: $S = ATC, CCA, CAG, TCC, AGT$.*

The concatenation gives $S = ATCCCACAGTCCAGT$. As a Shortest Superstring Problem, this gives ATCCAGT. As a Travelling Salesman Problem, it is just enough to show a path that touches the 3-mers ATC, TCC, CCA, CAG AGT as shown in Figure 1.6.

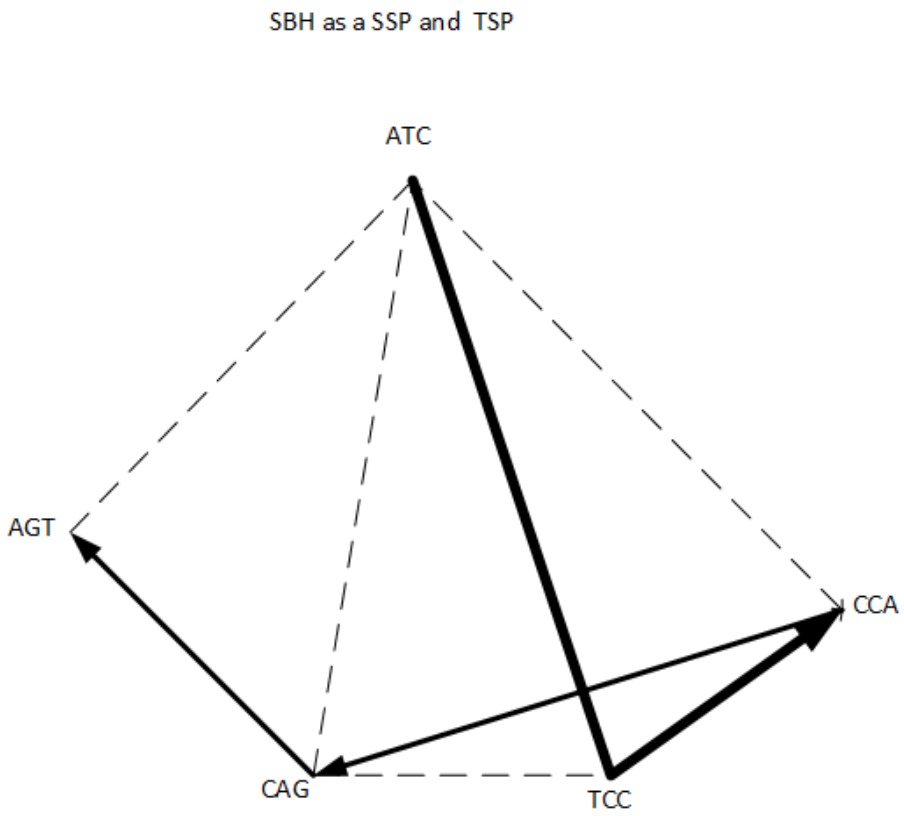


Figure 1.6: Reduction of SBH as a Shortest Superstring Problem and Travelling Salesman Problem

1.5 SBH as a Hamiltonian path problem

We can recast the SBH as a Hamiltonian path problem. This is a problem of finding a path in a graph that visits every vertex exactly once where the vertices are the I-tuples and edges are pairs of overlapping I-tuples.

Example 11. *Given the sequence:* $S = ATGCAGGTCC$ *with 3-spectrum given by*

ATG,AGG,TGC,TCC,GTC,GGT,GCA,CAG. We can have a Hamiltonian path given by $H = ATGCAGGTCC$. *For another spectrum which is given by* $S = ATG, TGG, TGC, GTG, GGC, GCA, GCG, CGT$, *there are two possible sequence reconstructions given by ATGCGTGGCA and ATGGCGTGCA as shown in Figure 1.7.*

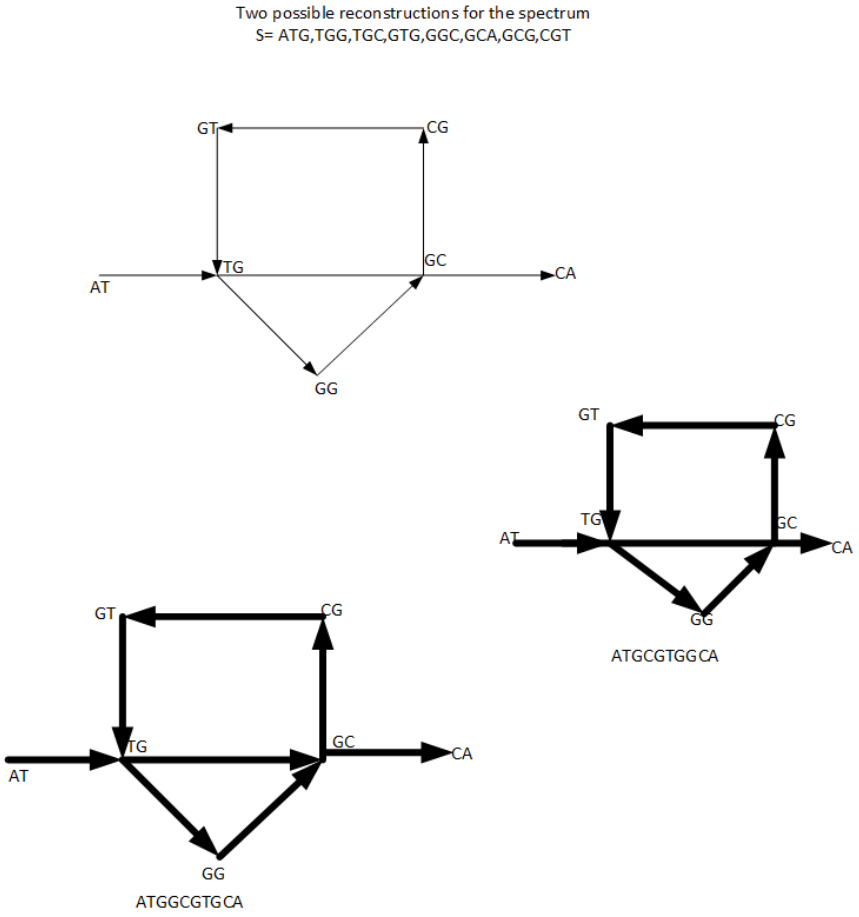


Figure 1.7: Two possible sequence reconstructions given the de Bruijin graph

This actually shows that for large DNA fragments, the overlap graphs become rather complicated and hard to analyze. The SBH problem can also be recast as an Eulerian Path Problem, [76] where the I-mers represent the edges between vertices. There is an edge from x_1 to x_2 if there is an I-mer whose first I-1 character is x_1 and whose last I-1 character is x_2. Other proposed methods of DNA sequence assembly could be found in [93], [99].

In general, molecular SBH is different from mathematical SBH. The molecular SBH relies on the hybridization of an (unknown) DNA fragment with a large array of short probes. Given a short (8-to 30-nucleotide) synthetic fragment of DNA called a Probe, and ssDNA as the Target, this target will bind (hybridize) to the probe if there is a substring of the target that is a Watson-Crick complement of the probe.

Example 12. *A probe sequence $S_p = ACCGTGGA$ will hybridize with a target sequence $S_t = CCCTGGCACCTA$ since it is complementary to the substring TGGCACCT of the target sequence. Hence probes can be used to test the unknown target DNA and determine its composition.*

SBH as a mathematical construct relies on the ideas from graph theory. We make use of the Hamiltonian and Eulerian path approaches. The Hamiltonian path approach shows a one-to-one correspondence between paths that visit each vertex of a graph at least once and DNA fragments with the spectrum S.

Example 13. *Hamiltonian: The 3-spectrum*

S= ATG, AGG,TGC,TCC,GTC,GGT,GCA,CAG has a path visiting all vertices leading to the sequence reconstruction S_r *= ATGCAGGTCC. The vertices are the 3-tuples and the edges are the overlapping 3-tuples. Another spectrum gives more sequence reconstructions.*

Example 14. *The Eulerian path for the 3-spectrum*

S= ATG,TGG,TGC,GTG,GGC,GCA,GCG,CGT has the vertices corresponding to the 2-tuples and the edges corresponding to the 3-tuples from the spectrum. This leads to two possible sequence reconstructions S_a*=ATGGCGTGCA and* S_b*= ATGCGTGGCA.*

Hence these two forms of SBH (molecular and mathematical) are distinguishable. It is worthy to note that the number of sequence reconstructions in mathematical SBH is bounded by the number of Eulerian cycles (paths) [76].

Chapter 2

Numerical DNA representations

DNA sequence is a set of alphabets made of four letters (strings): A, C, G and T representing the nitrogenous bases: Adenine, Cytosine, Guanine and Thymine respectively. They are not numeric in their original nature. The conversion of DNA sequences to numerical signals is important since it opens the possibility to employing powerful digital signal processing techniques for analyzing genomic data. Some of the desirable properties of a DNA numerical representation include [56]:

- Each nucleotide has equal weight (e.g; magnitude) since there is no biological evidence to suggest that one is more important than another.
- Distances between all pairs of nucleotides within the sequence itself should be equal since there is again no biological evidence to suggest that any pair is closer

than another.

- Representations should be compact to minimize redundancy and should allow access to a range of mathematical analysis tools.

Example 15. *Once the DNA is converted into digital signal, while retaining the biological meaning of the represented information, we can utilize the spectral analysis technique to find the particular nucleotide or exons* [1]*, since they exhibit the 3-base periodicity property [56].*

Other uses of the conversion of nucleotide sequences to numerical values are in analysis of DNA sequences using methods of statistical physics [18], computation of alignment free distances from DNA sequences [25] and prediction of nucleotide sequences by use of genomic signals [80].

This thesis is based on one of the forms of DNA representations called n-gram which is introduced later in this chapter. The only difference between it and other forms of DNA numerical representations is that n-gram method maps any DNA sequence to 4, 16, 64, 256 and so on dimensional feature vectors.

We give some examples of such DNA numerical representations [9], [12], [20], [25], [34], [37–40], [55], [58, 59], [71], [82, 83], [88], [91], [98], [110], [129].

[1] amino acid coding sequences

2.1 Voss (binary) method

This is the earliest and most popular form of mapping of DNA. The DNA is represented with four binary indicator sequences $x_A[n]$, $x_C[n]$, $x_G[n]$ and $x_T[n]$, where the presence of a nucleotide at a particular base pair position is represented by 1, and the absence of it is represented by 0.

Example 16. *The Voss representation of the DNA sequence* $S = GCTATCTATC$ *is given by*

$$x_A[n] = [0, 0, 0, 1, 0, 0, 0, 1, 0, 0]$$

$$x_C[n] = [0, 1, 0, 0, 0, 1, 0, 0, 0, 1]$$

$$x_G[n] = [1, 0, 0, 0, 0, 0, 0, 0, 0, 0]$$

$$x_T[n] = [0, 0, 1, 0, 1, 0, 1, 0, 1, 0]$$

2.2 Tetrahedron representation

This form of DNA sequence representation [56] reduces the number of indicator sequences from four to three but in a way symmetric to all the four sequences, where the four indicator sequences $x_A[n]$, $x_C[n]$, $x_G[n]$, $x_T[n]$ are mapped to the four 3-dimensional vectors pointing from the center to the vertices of a regular tetrahe-

dron.The resolving of the four 3-dimensional vectors results in the following colour outputs,

$$(a_r, a_g, a_b) = (0, 0, 1), \tag{2.1}$$

$$(t_r, t_g, t_b) = \Big(\frac{2\sqrt{2}}{3}, 0, -\frac{1}{3}\Big), \tag{2.2}$$

$$(g_r, g_g, g_b) = \Big(-\frac{\sqrt{2}}{3}, -\frac{\sqrt{6}}{3}, -\frac{1}{3}\Big), \tag{2.3}$$

$$(c_r, c_g, c_b) = \Big(-\frac{\sqrt{2}}{3}, \frac{\sqrt{6}}{3}, -\frac{1}{3}\Big), \tag{2.4}$$

which give rise to the following three numerical sequences

$$x_r[n] = \frac{\sqrt{2}}{3}\Big(x_T(n) - x_C(n) - x_G(n)\Big), \tag{2.5}$$

$$x_g[n] = \frac{\sqrt{6}}{3}\Big(x_C(n) - x_G(n)\Big), \tag{2.6}$$

$$x_b[n] = \frac{1}{3}\Big(3x_A(n) - x_T(n) - x_C(n) - x_G(n)\Big), \tag{2.7}$$

where r, g, b are red, green and blue indicators respectively. The main application of tetrahedron representation is in obtaining DNA spectograms of biomolecular sequences and this also provide local frequency information for all four bases by displaying the resulting three magnitudes by superposition of the corresponding three primary colours, red for $|X[k]|_r$, blue for $|X[k]|_b$, green for $|X[k]|_g$ and $X[k]$ is the discrete Fourier transform of the sequence given by

$$X[k] = \sum_{n=0}^{L-1} x_\alpha(n) e^{-j\frac{2\pi}{L}kn}, \quad k = 0, 1, 2, ..., L-1 \tag{2.8}$$

where $x_\alpha(n)$ are the respective indicator sequences, $\alpha \in A, C, G, T$. The sequence $X[k]$ gives a measure of the frequency content at frequency k, which corresponds to the underlying period of L/k samples. Hence the $X_\alpha[k]$, $\alpha \in A, C, G, T$ are the discrete Fourier transforms of the binary indicator sequences $x_\alpha(n)$, $\alpha \in A, C, G, T$.

2.3 Z-curves

The Z-curve method [40], [59], [69], [95], is another way of representation of DNA sequences into an equivalent three dimensional vectors based on the symmetry of the rectangular tetrahedrons.

Consider a DNA sequence of N bases and let the accumulated numbers of the bases A, C, G, T be given respectively by A_n, C_n, G_n, T_n respectively. The Z-curve consists of a series of nodes P_n (n=0,1,2,..,N) whose coordinates are denoted by x_n, y_n, z_n. These three coordinates are determined by the use of the four integers, A_n, C_n, G_n, T_n. The relationship is called the Z-transform given by

$$\begin{cases} x_n = (A_n + G_n) - (C_n + T_n) \\ y_n = (A_n + C_n) - (C_n + T_n) \\ z_n = (A_n + T_n) - (C_n + G_n) \end{cases}$$

for x_n, y_n, $z_n \in [-N, N]$, $\quad n = 0, 1, 2, ..., N$ where $A_0 = C_0 = G_0 = T_0 = 0$ and thus x_n=y_n=$z_n = 0$.

Given the coordinates of a Z-curve, the corresponding DNA sequence can be reconstructed by the use of the inverse Z-transform which is expressed as [40]

$$\begin{pmatrix} A_n \\ C_n \\ G_n \\ T_n \end{pmatrix} = \frac{n}{4}\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + \frac{1}{4}\begin{pmatrix} 1 & 1 & 1 \\ -1 & 1 & -1 \\ 1 & -1 & -1 \\ -1 & -1 & -1 \end{pmatrix}\begin{pmatrix} x_n \\ y_n \\ z_n \end{pmatrix} \tag{2.9}$$

for x_n, y_n, $z_n \in [-N, N]$, $\quad n = 0, 1, 2, ..., N$ and the relationship

$A_n + C_n + G_n + T_n = n$.

2.4 H-curves

In this method [27], [40], a vector function g(z) is defined as

$$g(z) = \begin{cases} i + j - k & \text{if z= A} \\ -i - j - k & \text{if z=C} \\ -i + j - k & \text{if z=G} \\ i - j - k & \text{if z=T} \end{cases} \tag{2.10}$$

where i, j and k are the unit vectors pointing in the direction of the Cartesisn x, y, z axes respectively. The 3-D curve (H-curve) consisting of a series on n-joined base

vectors is defined as

$$H_{ln} = h(z) = \sum_{1}^{n} g(z). \tag{2.11}$$

2.5 DNA walks

The first type of this form of DNA representation [40] is the 1-dimensional DNA walk. Let a DNA sequence of length L, be denoted as [x(i), i=1, 2, 3,...,L]. The presence of purines (A,G) and pyrimidines (C,T) in DNA sequence correspond to x(k) values of +1 and -1 respectively, where k denotes the position in the sequence. A DNA walk s(k) is defined as the cummulative sum of x(i) given by

$$s[k] = \sum_{i=1}^{k} x(i). \tag{2.12}$$

Other two letter alphabets such as Strong (S) $\in (C, G)$ or Weak (W) $\in (A, T)$, Keto (K) $\in (T, G)$ or aMino (M) $\in (A, C)$ can be employed to obtain the 1-dimensional DNA walks. The dimensionality of the numerical DNA sequence can be more than one for a better representation. The complex representation for DNA walk can be

defined as follows

$$x(k) = \begin{cases} 1 & \text{for k= A} \\ j & \text{for for k= C} \\ -j & \text{for for k = G} \\ -1 & \text{for for k = T} \end{cases} \tag{2.13}$$

and can be plotted in a 3-d Cartesian coordinate system by treating the accumulated values of the real part, imaginary part and the variable k as the values for x, y and z axes. This method of DNA representation can be used to extract useful information such as the long range correlation information and sequence periodicities from DNA sequences [40].

2.6 Integer number representation

This representation [38], is obtained by mapping numerals (1, 3, 2, 0) respectively to the four nucleotides as C=1, G=3, A=2, and T=0. This form of mapping is closely related to the Galois Indicator representation, which maps the CGAT nucleotides to a Galois field of 4, $GF(4)$ which is formed by assigning numerical values to the nucleotides as $C = 1$, $G = 3$, $A = 0$ and $T = 2$ in a nucleotide sequence. This representation suggests that $C < G$ and $A < T$ which is not always the case.

2.7 Paired nucleotide/Atomic number representation

Here the paired nucleotides are assigned with atomic numbers as A, G = 62 and C, T=42 respectively [59], [83]. Closely related to this form of representation is the atomic number representation, where atomic numbers are assigned to each nucleotide as C=58, G=78, A=70, T=66. There is also the molecular mass representation of a nucleotide sequence which is formed by mapping the four nucleotides to their molecular masses as C=110, G=150, A=134, and T = 125.

2.8 EIIP method

The EIIP (Electron-ion interaction pseudopotential) method represents the distribution of the free electrons energies along a nucleotide sequence [9]. The values for the individual nucleotides are given by $A = 0.1260$, $C = 0.1340$, $G = 0.0806$ and $T = 0.1335$.

Example 17. *The sequence S= GCTATCTATC can be converted to one numerical*

sequence given by

$$x[n] = [0.0806, 0.1340, 0.1335, 0.1260, 0.1335,$$

$$0.1340, 0.1335, 0.1260, 0.1335, 0.1340]$$

The above represents the distribution of the free electrons energies along the DNA sequence.

2.9 Double curve representation, DCR

Unlike the binary representation, the double curve representation [90], [115] takes two bases at a time. Any DNA sequence can be converted to six numerical sequence, based on the cumulative occurence of a chosen nucleotide pair AT, AC, AG, TC, TG or CG. The double curve representation of the base pair AT is defined as

$$x_{AT}[n] = \sum_{i=1}^{n} u(i), \quad n = 1, ..., L$$

where L is the length of the sequence and u(n) is defined as

$$u(n) = \begin{cases} +1 & \text{for base A} \\ -1 & \text{for base T} \\ 0 & \text{other bases} \end{cases}$$

Example 18. *For the sequence, $S = CATTGCCAGT$, the respective DCR are given by*

$$\begin{aligned} x_{AT}[n] &= [0,1,0,-1,-1,-1,-1,0,0,-1] \\ x_{AC}[n] &= [-1,0,0,0,0,-1,-2,-1,-1,-1] \\ x_{AG}[n] &= [0,1,1,1,0,0,0,1,0,0] \\ x_{TC}[n] &= [-1,-1,0,1,1,0,-1,-1,-1,0] \\ x_{TG}[n] &= [0,0,1,2,1,1,1,1,0,1] \\ x_{CG}[n] &= [1,1,1,1,0,1,2,2,1,1] \end{aligned}$$

The results in [90, 91] showed that the double curve representation is more informative than single base binary representations.

2.10 Paired numeric representation

This method is based on statistical property that introns [2] are rich in nucleotides A and T while exons are rich in nucleotides C and G [82]. It assigns values +1 and -1 to show the presence of A-T and C-G nucleotide pairs respectively. The main application of the paired numeric mapping is the gene and exon prediction whereby

[2] non-amino acid coding sequence

it exploits one of the differential properties of exons and introns, since introns are rich in nucleotides A and T and exons are rich in nucleotides G and C.

Example 19. *For the sequence,* $S = GCTATCTATC$*, the PNR is given by the single sequence* $P(n) = [-1, -1, 1, 1, 1, -1, 1, 1, 1, -1]$

2.11 Real/Complex number representation

For the real number representation, the nucleotide mappings are C= 0.5, G=-0.5, A=-1.5 and T=1.5, which bears complementary property. The complex number representation reflects the complementary nature of C-G and A-T pairs by mapping nucleotides $C = -1 - j$, $G = -1 + j$, $A = 1 + j$, and $T = 1 - j$.

2.12 Structural profile method

Structural information of physical properties of DNA molecule are utilized to mapping a nucleotide sequence to numerical sequence [115]. These structural informations include DNA bending stiffness (measured in nanometer, nM), duplex free energy, duplex disrupt energy and propeller twist. These structural profiles are calculated from conversion tables with step size of 1 along the DNA sequence, which takes di- or tri-nucleotides at a time.

Example 20. *The DNA bending stiffness profile for $S = CATTGCCAGT$ is given by $x_b[n] = [60, 20, 35, 60, 85, 130, 60, 60, 60]$.*

2.13 The code13 method

This is a kind of 1-sequence complex value numerical representation. It assigns values 1, -1, -j and j to the presence of the four nucleotides A, C, G, T respectively.

Example 21. *The sequence $S = CATTGCCAGT$ is given by*

$x[n] = [-1, 1, j, j, -j, -1, -1, 1, -j, j]$

2.14 Quaternion technique

In this method [59], pure quaternions are assigned to each base with

$$Q(z) = \begin{cases} i + j + k & \text{if z= A} \\ i - j - k & \text{if z=C} \\ i - j + k & \text{if z=G} \\ -i + j - k & \text{if z=T} \end{cases} \tag{2.14}$$

2.15 Internucleotide distance technique

This method represents each DNA nucleotide with a number representing the distance between the current nucleotide and the next similar nucleotide [10]. During scanning of the sequence from left to right, if a similar nucleotide is not found, the sequence value of the current nucleotide is the length of the remaining sequence.

Example 22. *The DNA sequence x[n]=AGTTCTACCGAGC, has the ID as*

$ID[n] = (6, 8, 1, 2, 3, 7, 4, 1, 4, 2, 2, 1, 0)$

2.16 Dot plot

Another way of representation of DNA sequence is the dot plot. Dot plots are a simple technique for visually comparing two base sequences of similar size to highlight slight differences [7]. One sequence is arranged along the horizontal axis and the other along the vertical axis. The base at each sequence on one axis is compared to every one of the other. The spectrum plot contains a dot where there is an exact match. This can be seen as an overlap from the graph theory point of view as shown in Figure 2.1.

Table 2.1 gives a summary of some of the numerical representations of a DNA sequence which we mentioned. [56] gives a detailed look at the merits and demerits of some of the DNA numerical representation methods.

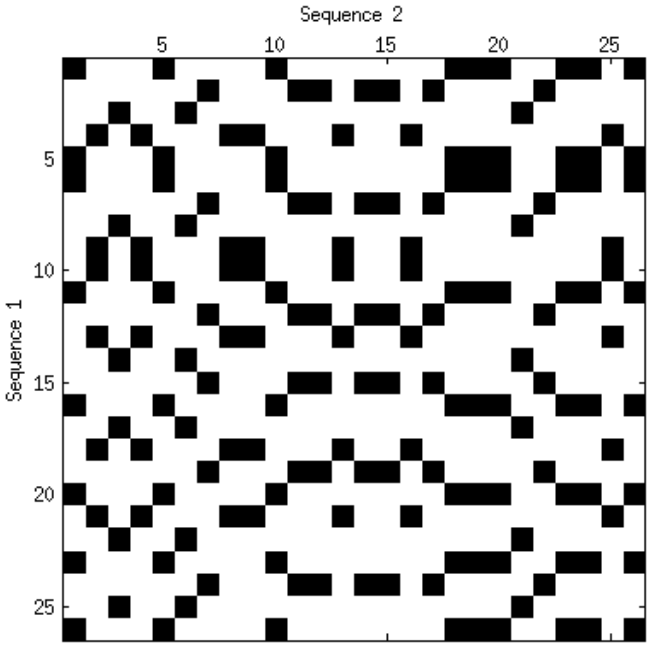


Figure 2.1: A sample dot plot of two sequences

Name	Code			
	C	G	A	T
1 Integer Number	1	3	2	0
2 Single Galois Indicator	1	3	0	2
3 Paired nucleotide Atomic number	42	62	62	42
4 Atomic Number	58	78	70	66
5 Molecular mass	110	150	134	125
6 EIIP	0.1340	0.0806	0.1260	0.1335
7 Paired Numeric	-1	-1	1	1
8 Real Number	0.5	-0.5	-1.5	1.5
9 Complex Number	-1-j	-1+j	1+j	1-j
10 K-Twin Pair Code	-1	-1	j	j
11 K -Bipolar-Pair Code I	-1	1	j	-j
12 K -Bipolar-Pair Code II	-1	1	-j	j
13 K- Quaternary Code I	-1	-j	1	j
14 K- Quaternary Code II	-1	-j	j	1
15 K- Quaternary Code III	-j	-1	1	j
16 K- Quaternary Code IV	-j	-1	j	1

Table 2.1: DNA sequence numeric representations

2.17 The N-gram

An n-gram as introduced by Shannon C. in 1948 [14] is a subsequence of length n of a sequence over a given alphabet. The sequence may be a message in a natural or artificial language, a discrete approximation of a continuous signal e.g. over a speech/signal or any sequence of symbols generated by a stochastic process. Any such "text" can be approximated by the set of n-gram statistical data (e.g; frequency distribution and the respective mean and standard deviation), and two such texts can be compared based on the distance of such approximations [73].

$N-$grams are sequences of n-words. Given a sequence S of $L-$words $L = l_1 l_2 ... l_N$, over a vocabulary L, and n, a positive integer, an n-gram of the sequence S is any subsequence $s_i ... s_{i+n-1}$ of n-consecutive words. There are L-n+1 such n-gram in S. For a vocabulary L with L distinct words, there are L^n possible unique n-gram. In a biological context, n-gram can be a sequence of nucleotides or n-amino acids. For instance, the sequence "AAANTSDSQKE" has two counts of 2-gram (digrams) AA, and one count each of 2-grams, AN, NT, TS, SD, DS, SQ, QK and KE [44].

Most of the analysis of plain text is based on Zipfs' law [33] which states that the most frequent word in any kind of text is expected to be twice as frequent as the second most frequent word. For a 25-mer of a nucleotide sequence, this may not always be

the case.

The earliest use of n-gram was in communication theory, [14]. There are other important aspects of scientific research and analysis where n-gram ideas have been used. Some are in information retrieval [5], textual information systems [8], G-protein coupling specificity prediction [13], classification of constrained DNA elements [24], string matching [29], relationships between n-gram patterns and protein secondary structure [49], statistical sequence analysis [60], prediction of HIV-1 coreceptor usage [65], visual framework for sequence analysis [105], whole genome promoter prediction [109], indexing DNA sequences [118], efficient and effective KNN sequence search [119], clustering of heartbeat signals [122], large-scale clustering of DNA texts [128]. The use of n-grams enables one to see the relationship between the individual members of the sequence in consideration. We give a formal definition of an n-gram following [29].

Definition 1. *N-gram: Given a sequence L and a positional integer n, a positional n-gram of L is a pair (i, g) where g is a subsequence of length n starting at the i-th element i.e, $g = L[i, i + n - 1]$.*

The set $G(L, n)$ consisting of all n-grams of L is obtained by sliding a window of length n over sequence L.

Example 23. *If $L = (A, C, G, T)$, its complete di-grams are AA, AC, AG, AT, CA,*

CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, TT.

By convention, the positional information of the n-gram are normally skipped. The tri-grams for the above example can be computed analogously. In general, for a 4-item set of unique strings, like the DNA alphabet, we have 4^n n-gram to consider. There are a total of $L-n+1$ n-gram in $G(L, n)$ where L is the length of the nucleotide and n is the length of the sliding window being considered.

The idea of n-gram has also been used to show that the edit distance between two strings and the longest common subsequence problem are n-gram distance and similarity, respectively [32]. In particular, when the edit operations are limited to insertions and deletions with no substitutions, the edit distance problem is equivalent to the longest common subsequence problem which is the problem of finding the longest common subsequence of two strings [76]. To measure similarity and vector space representation, the value similarity function [31] is used which quantifies the ratio of common edges between two graphs, taking into account the weight (associated numerical value) of common edges. Other forms of distance and similarity measures include normalized Euclidean distance, [13], Cosine similarity and Euclidean distance, [86], binary weighted cosine metric [94], and Jaccard similarity function, [97].

2.17.1 N-grams and their distance definitions

We define some n-gram proximity functions [5]. Denote the set of all strings over L as L^*, i.e $L^* = L^0 \cup \text{Ł}^1 \cup \text{Ł}^2 \cup \text{Ł}^3 \cup ...$

Definition 2. *Let $u = a_1a_2a_3..a_m \in \text{Ł}^*$ and let $x \in L^n$ be an n-gram. If $a_ia_{i+1}...a_{i+n-1} = x$ for some i, then u has an occurrence of x.*

Let $G(u)[x]$ denote the total number of occurrences of x in u.

Definition 3. *The n gram profile of x is the vector*

$$G_n(u) = G(u)[x] \quad x \in L^n. \tag{2.15}$$

Definition 4. *Let $u, v \in L^*$ and $n \in \mathcal{N}^+$. The n-gram distance between u and v is given by*

$$D_n(u, v) = \sum_{x \in L^n} |G(u)[x] - G(v)[x]| \tag{2.16}$$

Example 24. *Let $L = (a, b)$ and $u = abba$ and $v = babba \in L^*$. The di-gram profiles listed in lexicographical order are $(0, 1, 1, 1)$ and $(0, 1, 2, 1)$. The di-gram distance $D_2(u, v) = 1$.*

Observe that the distance definition given by Equation 2.16 is an L_1 norm (Manhattan distance) of the difference of their n-gram profiles [29]. The n-gram distance is a pseudo metric as shown if $u = agaa$ and $v = aaga$ since $u \neq v$ yet $D_2(u, v) = 0$

failing the identity condition of a metric space. Other properties of n-gram distance are also listed in [29]. One advantage of the n-gram distance formulation is that it allows one to compare strings of different lengths, a major difference from sequencing by hybridization.

2.17.2 Feature Vector Computation for nucleotides

Given a DNA of length say L, we count the number of appearances of (overlapping) strings of lengths n=2 (dinucleotide), n=3 (trinucleotide), n=4 (tetranucleotide) in the sequence. There are exactly $4^n, n = 2, 3, 4$ types of such strings, respectively. Consider a DNA sequence of length L. We denote the frequency of appearance of the n-string $\alpha_1\alpha_2...\alpha_n$ by $f(\alpha_1\alpha_2...\alpha_n)$ where each $\alpha_i \in A, C, G, T$ for DNA sequence. This frequency divided by the total number of n-grams given by $L-n+1$ of an n-string in the given sequence may be taken as the probability $p(\alpha_1\alpha_2...\alpha_n)$ of appearance of the string $\alpha_1\alpha_2...\alpha_n$ in the DNA sequence. This probabilities act as features and can be seen as normalized frequency counts given by [17], [109], [123],

$$v_i^n = \frac{f_i^n}{|L| - n + 1}, \quad 1 \leq i \leq 4^n, \quad \text{for n} = 2, 3, 4 \tag{2.17}$$

where L is the length of the sequence. The denominator in Equation 2.17 denotes the number of n-grams that are possible in a sequence of length L and hence v_i^n denotes the proportional frequency of occurrence of the i-th feature for a particular n-value.

Hence each DNA sequence of the data set is represented as a 16-dimensional feature vector $(v_1^2, v_2^2, ..., v_{16}^2)$ for n = 2, as a 64-dimensional feature vector $(v_1^3, v_2^3, ..., v_{64}^3)$ for n= 3, and as a 256-dimensional feature vector $(v_1^4, v_2^4, ..., v_{256}^4)$ for n=4.

2.17.3 Dinucleotide frequencies

Given a 25 base DNA sequence, S_1 = ACGTAACGTTACTGCAGTCATGACG.

By considering neighbouring bases, we obtain 16 dinucleotides (2-grams) given by

$S_1(XY)$ = AA,AC,AG,AT,CA,CC,CG,CT,GA,GC,GG,GT,TA,TC,TG,TT.

Denote by F_{XY} and F_X, the cumulative numbers of the dinucleotide XY and nucleotide X respectively.

Definition 5. *The absolute frequency $P_L(XY)$ is the ratio of the cumulative numbers of the dinucleotide XY to that of the first nucleotide X [126]*

$$P_L(XY) = \frac{F_{XY}}{F_X}. \tag{2.18}$$

For a DNA sequence L, the mononucleotide absolute frequency vector is defined by $V_L(1) = [P_L(A), P_L(C), P_L(G), P_L(T)]$. The dinucleotide absolute frequency vector is defined by $V_L(2) = [P_L(AA), P_L(AC), P_LAG, ..., P_L(TT)]$. The trinucleotide absolute frequency vector is defined by $V_L(3) = [P_L(AAA), P_L(AAC), P_L(AAG), ..., P_L(TTT)]$ and the tetra-nucleotide absolute frequency vector is defined by

$V_L(4) = [P_L(AAAA), P_L(AAAC), P_L(AAAG), ..., P_L(TTTT)]$. In this way we can obtain a correspondence between the DNA sequence and a 4, 16, 64, 256 component vector respectively. A DNA sequence can be analysed by studying the corresponding mono, di, tri and tetra nucleotide frequency vectors [116], [126].

Example 25. *Given the sequence* $S_1 = ACGTAACGTTACTGCAGTCATGACG$, *the absolute frequency matrix* $P_L(XY)$ *is given by*

$$\begin{pmatrix} 0.1428 & 0.571 & 0.1428 & 0.1428 \\ 0.333 & 0 & 0.5 & 0.1667 \\ 0.1667 & 0.1667 & 0 & 0.5 \\ 0.333 & 0.1667 & 0.333 & 0.1667 \end{pmatrix} \tag{2.19}$$

Let F_X denote the frequency of the nucleotide X in a sequence S_1 and F_{XY} the frequency of the dinucleotide XY. A standard assessment of the dinucleotide bias is through the odds ratio calculation [99]

$$\rho_{XY} = \frac{F_{XY}}{F_X F_Y}. \tag{2.20}$$

For ρ_{XY} values sufficiently larger (smaller) than 1, the XY dinucleotide is considered of high (low) relative abundance compared with the random association of its component mononucleotides [99].

Example 26. *Consider the sequence*

$S_1 = ACGTAACGTTACTGCAGTCATGACG$. The odds ratio frequency matrix ρ_{XY} is given by

$$\begin{pmatrix} 0.020 & 0.095 & 0.0238 & 0.0238 \\ 0.0476 & 0 & 0.0833 & 0.0278 \\ 0.0238 & 0.0278 & 0 & 0.0833 \\ 0.0476 & 0.0278 & 0.0556 & 0.0278 \end{pmatrix} \tag{2.21}$$

In general, given any nucleotide sequence S, the following relation is true: $\rho_{XY} < P_{XY}$ i.e. the entries (elements) of the odds ratio matrix are less or equal to the elements of the absolute frequency matrix.

Chapter 3

Data mining algorithms for DNA sequence

3.1 Introduction

The amount of data in biological sequences is all the time increasing and is becoming overwhelming today. Microarrays or bioarrays are a high throughput technology producing large amounts of data that earlier have only been noticed in other research areas and disciplines. Together with image processing and database methodologies, data analysis is one of the important tasks for those involved in biological sequence analysis.

Microarray data analysis is dependent upon many factors, principal amongst them are: volume of data, dimensionality, quality and normalization. It is noteworthy that the volume of many microarray data is always large vis-a-vis their dimension. The dimensionality here relates to presenting the data in the form of a matrix consisting of nucleotide sequences in row values and signal intensities (observed values) as measured and recorded by the measuring instrument across columns. The dimensionality problem for microarray data consists of thousands of nucleotides made of Adenine, Cytosine, Guanine and Thymine arranged in a given order (variables) and few observations (intensities).

The quality of data also presents a problem of data variation or noise. This variation can be due to both relevant (incomplete, inaccurate and inconsistencies in data) and non-relevant sources like the hybridization condition or the chip quality. It is somehow difficult to estimate the amount of unwanted variation with only a few observations. Ways have to be found to either eliminate the source of variation by having very precise and consistent experimental protocols or by applying statistical techniques to deal with such sources of variation.

Normalization is another key feature of data quality since some numerical elements in the data could be too large compared to other members of the data. Normalization is an essential step before performing any kind of analysis. Generally, in data mining,

normalization is carried out on the attribute values which could be the signal intensity profiles of the DNA sequence as recorded by the Affymetrix Genechip or the n-gram profiles as computed. This normalization is done since using the Euclidean formula, the effect of some attributes might be completely dwarfed by others. Normalization of the attribute values to lie between 0 and 1, we calculate [46], [48]

$$N_1(i) = \frac{y_i - miny_i}{maxy_i - miny_i} \tag{3.1}$$

where y_i is the actual value of the attribute i and the maximum and minimum are taken over all instances in the training set. The formula implicitly assumes numeric attributes. Hence, the difference between two values is just the numerical difference between them. There are variants of the above form of normalization which we call row-wise normalization given in Equations 3.2 and 3.3:

$$N_2(i) = \frac{y_i - miny_i}{max - min} \tag{3.2}$$

and

$$N_2(i) = \frac{y_i - min}{max - min} \tag{3.3}$$

where y_i is the actual value of the attribute i, $miny_i$ is the minimum over all instances in the training set, max/min are the maximum/minimum values along each row. For the task at hand, I used the row-wise normalization given in Equation 3.3. One can

always convert to the original row values by the transformation:

$$y_i = N_2(i)(max - min) + min \tag{3.4}$$

Other methods of normalization of data are [48], Min-max normalization which performs a linear transformation on the original data, the z-score normalization which does normalization based on the mean and standard deviation of the attribute, and normalization by decimal scaling which normalizes by moving the decimal point of values of the attribute, where the number of decimal points moved depends on the maximum absolute value of the attribute.

There has always been an interest in understanding the inter-relationship between elements in a data set. Data mining is about solving problems by analyzing data already present in a database. Data mining is defined as the process of discovering patterns in data. The process must be automatic or (more usually) semi-automatic. Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarise the data in novel ways that are both understandable and useful to the data owner [46], [48], [52]. The DNA sequence can be seen as a data set made of different measures of particular interest.

A data set is a collection of objects with each object associated with a set of P attributes (sometimes called measurements). A collection of M objects can be represented by an $M \times P$ matrix where rows correspond to objects (individuals, entities,

cases, records, etc) and columns correspond to attributes (variables, features, fields, etc). A data cube [48] allows data to be modelled and viewed in multiple dimensions and helps to reduce the search space. In our case, the DNA sequence can be modelled as a cuboid. We relate the DNA sequence ACGT as 0, 1, 2, 3, 4-dimensional cuboid, where the 2 dimensional cuboids are the 2-gram, 3 dimensional cuboids are the 3-gram and so on. There are 2^p cuboids for p dimensions. The schematics of the DNA cuboid reduction is shown in Figure 3.1

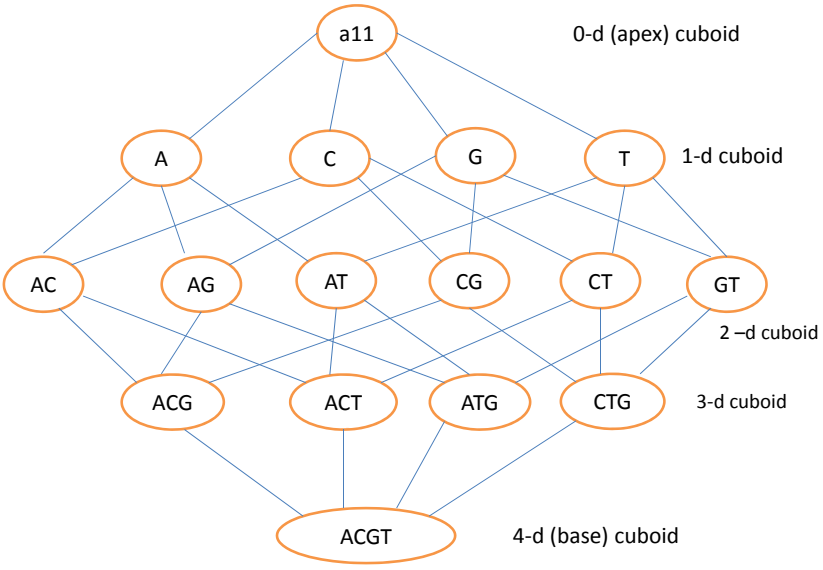


Figure 3.1: DNA sequence as lattice of cuboids

Closely related to data mining is machine learning which can be interpreted either as building sophisticated machines that are capable of learning or as mechanizing the process of learning [104]. A classical definition of "learning" as concerns computers is given in [108].

Definition 6. *A computer program is said to learn from experience E with respect to some class of tasks T and performance measure F if its' performance at tasks in T as measured by F, improves with experience E.*

Learning has many classification of which one is instance-based. Instance-based learning methods are conceptually straightforward approaches to approximating real or discrete valued target functions [108]. Example of instance-based learning methods are Nearest Neighbours and locally weighted regression. Learning in these algorithms consists of simply storing the presented training data. When a new query instance is encountered, a set of similar related instances is retrieved from memory and used to classify the new query instance. In general instance-based methods share three key properties [108]. The first is that they are "lazy" learners in that they defer the decision of how to generalize beyond the training data until a new query instance is observed. Secondly, they classify new query instances by analyzing similar instances while ignoring instances that are very different from the query. The last of the key properties is that they represent instances as real valued functions in an n-dimensional

Euclidean space. There are also the case-based reasoning classifiers which, instead of storing the training tuples as points in the Euclidean space, store the tuples or "cases" for problem solving as complex symbolic descriptions [48]. Advantages and disadvantages of these instance and case based learning methods are elaborated in many texts [108], [111], [120]. Other classification methods include Genetic algorithms which was introduced by John Henry Holland in 1970 [41], which attempts to incorporate ideas of natural evolution into classification. Genetic algorithms has led to other areas like genetic programming [51], [114], which has been applied in microarray data analysis [112].

3.2 Artificial Neural Networks

Neural Network learning methods provide a robust approach to approximating real-valued, discrete-valued and vector-valued target functions [108]. The study of artificial neural networks has been inspired in part by the observation that biological learning systems are built of very complex webs of interconnected neurons [46], [104], [108]. The ideas from artificial neural networks have led to computational analysis of human DNA sequence [3], study of working conditions [21], single base pair discrimination of terminal mismatches [45], biological phenomena through computational

intelligence [50], human donor and acceptor sites prediction [57], predicting global solar radiation in A1 Ain City, UAE [62], coding region recognition and gene identification [66], power prediction analysis [72], predicting transmembrane domains of proteins [77], prediction of nucleotide sequences using genomic signals [80], [81]. Many other advantages of neural networks in data mining and bioinformatics are outlined in [127].

Artificial Neural Network (ANN) architecture consists of many neurons organized in layers. Each neuron in a layer is connected to all the neurons of the next layer. There are three types of layers: input, hidden and output layers. In general, ANN's can be graphs with many types of structures: acyclic or cyclic, directed or undirected.

The most common and practical ANN approach is based on back propagation algorithm. The back propagation algorithm assumes the network is a fixed structure that corresponds to a directed graph, possibly containing cycles. Learning corresponds to choosing a weight value for each edge in the graph. Although certain types of cycles are allowed, the vast majority of practical applications involve acyclic forward networks. [92].

The schematics of an artificial neuron with one input and bias is shown in Figure 3.2. Scalar inputs v_i are transmitted through connections that multiply their strength by the scalar weight x_i to form the product $x_i v_i$ which is also a scalar. All the weighted

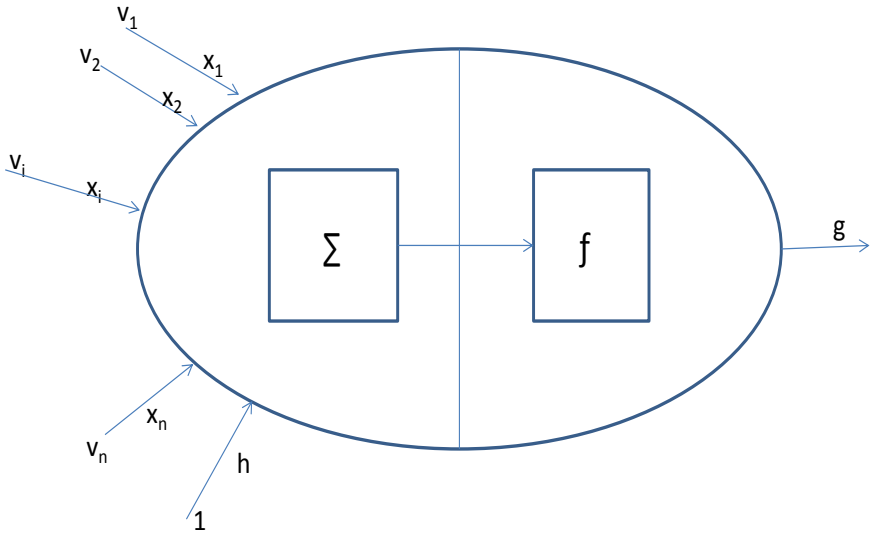


Figure 3.2: Schematics of an artificial neuron

inputs $x_i v_i$ are added plus the scalar bias h. The result is the argument of the transfer function **f** which produces the output **g** given by

$$\mathbf{g} = f(\sum x_i v_i + h). \tag{3.5}$$

The scalar weights x_i and biases h_i are adjustable scalar parameters of the neuron. The idea of an artificial neural network is the adjustment of such parameters so that

the network exhibits some desired or interesting behaviour and this adjustment of such parameters are done by the use of transfer functions like the hard-limit (step), linear and the sigmoid function [30]. The sigmoid function is a popular and useful nonlinear transfer function.

3.2.1 Learning Rules (LR) and Training

Learning rules are procedures for modifying the weights and biases of a network i.e. method of deriving the next changes that might be made in an ANN. It could be supervised or unsupervised. In supervised learning the network is provided with a set of examples (training set) of proper network behaviour (pairs of input and known/correct output (target)). As the inputs are applied to the network, the calculated network outputs are compared to the targets. The learning rule is then used to adjust the weights and biases of the network in order to move the network outputs closer to the target. In unsupervised learning, weights and biases are modified in response to network inputs only [30].

3.2.2 Perceptron learning rule

ANN systems are of many types. One is based on a unit called perceptron. A perceptron takes a vector of real valued inputs, calculates a linear combination of

these inputs, outputs a 1 if the result is greater than some threshold and -1 otherwise. Henceforth, I use w_i as weights and x_i as inputs. Given inputs x_1 through x_n, the output $o(x_1, x_2, ..., x_n)$ computed by the perceptron is [108], [111],

$$o(x_1, ..., x_n) = \begin{cases} 1 & \text{if } w_0 + w_1 x_1 + ... + w_n x_n > 0 \\ -1 & \text{otherwise} \end{cases} \tag{3.6}$$

where each w_i is a real valued constant (weight) that determines the contribution of input x_i to the perceptron output. The threshold $-w_0$ is the quantity that the weighted combination of inputs $w_1 x_1 + ... + w_n x_n$ must surpass in order for the perceptron to output 1. Because of this linearity assumption of a simple perceptron rule, using the step function to convert the weighted sum of the inputs into a 0/1 prediction is not plausible. The step function itself is not continuous. I need a function that is similar in shape but continuous and differentiable. To overcome this, I make use of the sigmoid (logistic) function which computes the output [15], [30], [108] as

$$o = \sigma(\overrightarrow{w} . \overrightarrow{x}) \tag{3.7}$$

where

$$\sigma(y) = \frac{1}{1 + e^{-y}} \tag{3.8}$$

The output of the sigmoid function ranges from 0 to 1, increasing monotonically with its input. The logistic function is also called a squashing function of a unit since

it maps very large input domain to a small range of output, hence my choice of it. Also, the derivative of the sigmoid function can be expressed in terms of the output function.

3.2.3 Backpropagation algorithm

I make use of the backpropagation algorithm which employs gradient descent to attempt to minimize the squared error between the network output values and the target values for this outputs [108]. Backpropagation learns by iteratively processing a data set of training tuples, comparing the networks prediction for each tuple with actual known target value. The target value may be the known class label of a continuous value for prediction. The backpropagation algorithm (Rumelhart and McClelland, 1986) is used in layered feed-forward ANNs. This means that the artificial neurons are organized in layers and send their signals "forward", and the errors are propagated backwards. The network receives inputs by neurons in the input layer and the output of the network is given by the neurons on an output layer. There may be one or more intermediate hidden layers. The backpropagation algorithm uses supervised learning which means that I provide the algorithm with examples of the inputs and outputs I want the network to compute, and the error (difference between actual and expected results) is calculated. The idea of the backpropagation algorithm

is to reduce this error, until the ANN learns the training data. The training begins with random weights and the goal is to adjust them so that the error will be minimal. The goal of the training process is to obtain a desired output when certain inputs are given. Since the error is the difference between the actual and the desired output, the error depends on the weights and we adjust the weights in order to minimize the error [108]. There are other architectural forms of neural networks which include radial basis function network which has the same architecture as multilayer perceptron but uses Gaussian, splice or various quadratic functions as the transfer function [46], [66]. Kohonen self organizing maps is another form of neural network architecture. They have only an input layer with no hidden or output layer. Input layer units connect to a grid of discrete units. The input vectors are mapped to one of the grid points by computing the Euclidean distances for each grid point to decide on the closest point which matches the input [66].

3.2.4 Derivation of the backpropagation rule for multi-layer networks

I define the error function F_d for the output of each neuron (individual training example) d as follows [108]

$$F_d(\overrightarrow{w}) = \frac{1}{2}(t_d - o_d)^2 \tag{3.9}$$

where t_d and o_d are the target value and unit output value for training example d. Stochastic gradient descent iterates over the training examples d in D, where D is the set of training examples at each iteration altering the weights according to the gradient with respect to $F_d(\overrightarrow{w})$. I take the square of the differences between the output and the desired target because it will be always positive and will be greater if the differences are big, and less if the differences are small.

In other words, for each training example d, every weight w_{ji} is updated by adding to it $\triangle w_{ij}$ where

$$\triangle w_{ji} = -\eta \frac{\partial F_d}{\partial w_{ji}} \tag{3.10}$$

where F_d is the error on training example d and η is the learning rate. The error of the network will simply be the sum of the errors of all the neurons in the output layer given by

$$F_d(\overrightarrow{w}) = \frac{1}{2} \sum_{k \in outputs} (t_k - o_k)^2 \tag{3.11}$$

where outputs is the set of output units in the network, t_k is the target value of unit k for training example d and o_k is the output of unit k given training example d. I make use of the following notations for our subsequent derivations [108]:

- x_{ji} = the i-th input to unit j
- w_{ji}= the weight associated with the i-th input to unit j

- $net_j = \sum_{i=0}^{n} x_{ji} w_{ji}$ (the weighted sum of inputs for unit j)
- o_j= the output computed by unit j
- t_j= the target output for unit j
- σ = the sigmoid function
- outputs = the set of units in the final layer of the network
- Downstream(j) = the set of units whose immediate inputs include the output of unit j.

The activation function of the artificial neurons in ANNs implementing the backpropagation algorithm is a weighted sum (the sum of the inputs x_{ji} multiplied by their respective weights w_{ji}) given by:

$$net_j(\overrightarrow{w}.\overrightarrow{x}) = \sum_{i=0}^{n} x_{ji} w_{ji} \tag{3.12}$$

To derive an expression for $\frac{\partial F_d}{\partial w_{ji}}$, note that the weight w_{ji} can influence the rest of the network only through net_j. Hence I can write

$$\begin{aligned} \frac{\partial F_d}{\partial w_{ji}} &= \frac{\partial F_d}{\partial net_j} \frac{\partial net_j}{\partial w_{ji}} \\ &= \frac{\partial F_d}{\partial net_j} x_{ji} \end{aligned} \tag{3.13}$$

I now consider the 2 cases for the convenient expression for $\frac{\partial F_d}{\partial net_j}$: the case where unit j is an ouput unit for the network and the case where j is an internal unit.

3.2.5 Case 1: Training rule for output unit weights

As already observed, just as w_{ji} can influence the network only through net_j, net_j can influence the network only through o_j [108]. By chain rule, I write

$$\frac{\partial F_d}{\partial net_j} = \frac{\partial F_d}{\partial o_j}\frac{\partial o_j}{\partial net_j} \tag{3.14}$$

Consider the first term in Equation 3.14, I have

$$\frac{\partial F_d}{\partial o_j} = \frac{\partial}{\partial o_j}\frac{1}{2}\sum_{k\in outputs}(t_k - o_k)^2 \tag{3.15}$$

The derivatives $\frac{\partial}{\partial o_j}(t_k - o_k)^2$ will be zero for all output units k except when $k = j$. I drop the summation over the output units and set $k = j$.

Hence,

$$\begin{aligned}\frac{\partial F_d}{\partial o_j} &= \frac{\partial}{\partial o_j}\frac{1}{2}(t_j - o_j)^2 \\ &= -(t_j - o_j)\end{aligned} \tag{3.16}$$

Next I consider the second term in Equation 3.14. Since $o_j = \sigma(net_j)$, the derivative $\frac{\partial o_j}{\partial net_j}$ is just the derivative of the sigmoid function which equals $\sigma(net_j(1 - \sigma(net_j))$. Therefore

$$\begin{aligned}\frac{\partial o_j}{\partial net_j} &= \frac{\partial \sigma net_j}{\partial net_j} \\ &= o_j(1 - o_j).\end{aligned} \tag{3.17}$$

Substituting Equations 3.17 and 3.16 into Equation 3.14, I have

$$\frac{\partial F_d}{\partial net_j} = -(t_j - o_j)o_j(1 - o_j) \tag{3.18}$$

Combining Equation 3.18 with Equations 3.10 and 3.13, I have the stochastic gradient descent rule for output units as

$$\begin{aligned} \triangle w_{ji} &= -\eta \frac{\partial F_d}{\partial w_{ji}} \\ &= \eta(t_j - o_j)o_j(1 - o_j)x_{ji}. \end{aligned} \tag{3.19}$$

3.2.6 Case 2: Training rule for hidden unit weights

The derivation of the training rule for w_{ji} for the case where j is an internal or hidden unit in the network must take into account the indirect ways in which w_{ji} can influence the network outputs and hence F_d, the error on training d, summed over all output units in the network [108]. I refer to the set of all units immediately downstream of unit j in the network and denote this set of units by Downstream(j). Note that net_j can influence the network outputs and F_d only through the units in Downstream(j).

I have

$$\begin{aligned}
\frac{\partial F_d}{\partial net_j} &= \sum_{k \in downstream(j)} \frac{\partial F_d}{\partial net_k} \frac{\partial net_k}{\partial net_j} \\
&= \sum_{k \in Downstream(j)} -\delta_k \frac{\partial net_k}{\partial net_j} \\
&= \sum_{k \in Downstream(j)} -\delta_k \frac{\partial net_k}{\partial o_j} \frac{\partial o_j}{\partial net_j} \\
&= \sum_{k \in Downstream(j)} -\delta_k w_{kj} \frac{\partial o_j}{\partial net_j} \\
&= \sum_{k \in Downstream(j)} -\delta_k w_{kj} o_j (1 - o_j)
\end{aligned} \tag{3.20}$$

Rearranging terms and using δ_j to denote $-\frac{\partial F_d}{\partial net_j}$, I have

$$\delta_j = o_j(1 - o_j) \sum_{k \in Downstream(j)} \delta_k w_{kj}$$

and

$$\triangle w_{ij} = \eta \delta_j x_{ji}$$

This is precisely the general rule for updating internal unit weights in arbitrary acyclic directed graphs [46], [78], [104], [108], [120].

Chapter 4

Data Formating and Simulations

4.1 Sequence Encoding Schema

The sequence encoding schema is used to convert DNA sequences (character strings) into input vectors (numbers) of the neural network classifier. A good encoding scheme should satisfy the basic coding assumption so that similar sequences are represented by 'close' vectors [16], [19]. In the encoding, the original DNA sequence string can be represented by different alphabet sets including: set A, the 4-letter nucleotide, set E, Purines (R) versus Pyrimidines (Y): R=A, G; Y=C, T. It can also be encoded by Strong (S) versus weak (W) hydrogen bonding: S=C, G; W=A, T. Another encoding which is of less physiochemical significance is Keto (K) and aMino (M): K =T, G and

M = A, C. Different n-gram encoding methods are named by a two character code: the first character is a letter designating the alphabet set and the second character is a digit representing the size or length of the n-gram [47], [120]. Table 4.1 gives a summary of the merged forms of nucleotide monomers.

Molecule	Size	Grouping
DNA	4	Adenine(A) vs. Cytosine(C) vs. Guanine(G) vs. Thymine (T)
DNA	2	Purines(R) vs. Pyrimidines(Y): R = A, G; Y = C, T
DNA	2	Strong(S) vs. Weak(W) hydrogen bonding: S= C, G; W = A, T
DNA	2	Less physiochemical significance, Keto(K)=T, G vs. aMino(M)= A, C

Table 4.1: Merged alphabets of nucleotide monomers

4.1.1 Neural network architecture with n-gram

Python programming language [63,64] is used to extract and count the occurrences of patterns of n-consecutive bases (n-grams). N-gram patterns consist of n-consecutive residues and are extracted from sequence strings in a sliding window fashion. A feed-forward neural network with three layers, namely, an input layer, one hidden layer and output layer is used for the signal intensity prediction. The number of

nodes in the input layer is 4 and 16 for n = 1, 2 respectively. Due to the influence of the length of nucleotide in the nucleotide sequence hybridization intensity [1], we replace the nucleotide sequences with their n-gram counts (ratios). Hence, the prediction task has been simplified by reducing the set of input values. Specifically, every single nucleotide has been replaced by their probability. The same is done on the dinucleotides. Tables 4.2 and 4.3 show the percentages (ratios) from our dataset of nucleotides and dinucleotides respectively.

Nucleotides	A	C	G	T
ratio	0.310	0.311	0.131	0.248

Table 4.2: The nucleotide percentages (ratios)

Dinucleotides	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
Ratio	0.098	0.089	0.049	0.074	0.091	0.105	0.026	0.089	0.038	0.043	0.025	0.025	0.084	0.073	0.030	0.061

Table 4.3: The dinucleotide percentages (ratios)

Experimentation is done with different number of neurons in the hidden layer that give an optimal prediction performance. The output node layer has in our case 4 nodes reflecting our choice of sequence signals to predict. The schematics of DNA neural

network architecture is shown in Figure 4.1. The DNA sequence is first converted by a sequence encoding schema into neural network input vectors (ratios of n-gram). The neural network then predicts those intensities according to the sequence information embedded in the neural interconnections after network training.

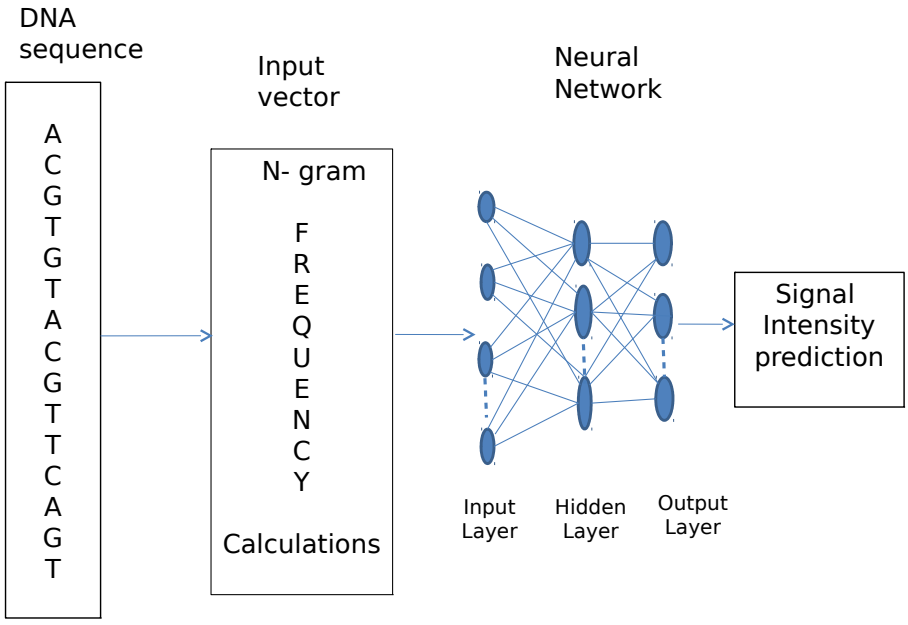


Figure 4.1: A neural network system for signal intensity prediction.

The counts of n-gram patterns (frequency) are scaled to fall between 0 and 1 and used as input vectors for the neural network, with each unit of the vector representing the n-gram pattern. The size of the input vector i.e. the number of input units

for each n-gram method is M^n, where M is the size of the alphabet and n is the n-gram pattern. The n-gram method has several advantages which include mapping sequences of various lengths into input vectors of the same length. It also provides certain representation invariances with respect to residue insertion and deletion and is independent of the apriori recognition of certain specific patterns.

The major drawback of the n-gram method is that the size of the input vector tends to be large. This indicates that the size of the weight matrix i.e. the number of neural interconnections would also be large because the weight matrix size equals w where w = input size × hidden size + hidden size × output size.

As an example, the tri-grams of amino acids would require 20^3 or 8000 input units. Accepted statistical techniques and current trends in neural networks favor minimal architectures i.e. those with fewer neurons and interconnections for better generalization capability [3], [124].

4.1.2 Matlab neural network simulator

A neural network developed by Mathworks is the major tool for our analysis [107]. Our choice of Artificial Neural Network (ANN) for our task is necessitated by the fact that ANNs have the ability to learn from data, either in a supervised or an unsupervised manner and can be used in tasks such as regression, classification, clustering

and other forms of learning tasks. The number of input nodes is the number of independent variables in the problem. The number of hidden layers and/or nodes is a measure of the nonlinearity of the function and has to be determined from the data. The number of output nodes is the number of dependent variables. The activation function is a simple function that sets the value of a given node based on the total input into it. Two common choices of activation functions are linear and logistic. A linear activation function for all nodes reduces the Neural Network to linear regression while a logistic function is a nonlinear but smoothed step function bound between 0 and 1. I make use of the logistic function for the analysis.

The Matlab neural network simulator normally partitions the input data into three parts (percentages) for training purposes which are training, say 60 percent, validation, 20 percent and testing, 20 percent. Training is used for computing the gradient and updating the networks weights and biases. Validation is used to decide when to stop the training process to avoid overfitting. Overfitting is a serious problem in neural networks where the network so created memorizes the training data, rather than learning the law that governs them [30], [66]. This makes it impossible to categorize new data and hampers generalization. Testing is a separate data set which is used to test the trained neural network at the end to determine whether it has generalized the training data set accurately. It is important that the testing data do not participate

in the training process.

4.1.3 Algorithmic steps on the DNA profile

The algorithmic steps for our data manipulation are as follows:

1: Compute n-gram profiles of the DNA data set using Python programming language.

2: Calculate the nucleotide and dinucleotide frequencies of these profiles.

3: Do substitution of the nucleotides and dinucleotide strings with their respective frequencies.

Do the following on the intensity profiles:

4: Calculate the highest and lowest value along each row.

5: Do normalization along each row using

$$N(i) = \frac{y_i - min}{max - min}$$

where y_i is the actual value of the attribute i, max and min are the maximum and minimum values along each row.

6: Repeat step 5 for every row of intensity profile.

7: Combine results obtained from step 1 to step 6.

8: Extract every 26th line [1] from the data set after the operations above.

[1] To avoid subsequence overlap and random match

9: Use matlab subroutines to do performance, confusion matrix, regression and ROC evaluations.

The modified Affymetrix data is shown in Appendix A with the nucleotides and dinucleotides replaced by their respective ratios (n-gram counts) with row-wise normalized intensity values.

4.1.4 Cross validation

A k-fold cross validation randomly partitions an initial data into k mutually exclusive subsets $F_1, F_2, ..., F_k$, each of approximately equal size (partitions) [46], [48], [108]. Each of the folds in turn is used for testing and the remainder is used for training. This makes it possible that at the end of the exercise, every partition has been used the same number of times for training and exactly once for testing. The error associated with each run on the different training set is recorded and the k error extimates are averaged to yield an overall error extimate. I did a 10-fold cross validation on our data set to see if the results are consistent since it has been shown that 10 is about the right number of folds to get the best extimate of error [46]. Algorithmic steps for performance of a k-fold cross validation can be listed as follows:

1: Arrange the data set in a random order.

2: Divide the data set into k-folds of approximately equal sizes.

3: For i=1,...,k

a: Train the classifier using all the data set that do not belong to i-th data partition

b: Test the classifier on the examples in i-th data partition

c: Compute the accuracy or error associated with each of the iterations

4: Total accuracy or error is the sum of all the correct or wrong classification errors divided by the number of runs.

4.2 Steps to using the Matlab simulator

Neural Network Toolbox in Matlab, [107] which is a set of tools that include GUIs which stands for graphical user interfaces, wizards and functions is used. The toolbox can be used to fit a function, do regression analysis, recognize patterns/classification and cluster data. The neural network and simulation pane are shown in Figure 4.2. As expected, for it to work with large data sets like excel, one has to format the data file to be readable by Matlab and this problem is overcome easily since Matlab has a way of converting an excel file into a .mat file for easy analysis. Again, the intensity profile of the DNA strands are normalized row-wise. After converting the excel file to a .mat file for Matlab neural network simulator, the following steps are done:

1. Load the data say $A1$ = positional information of the n-gram values or input

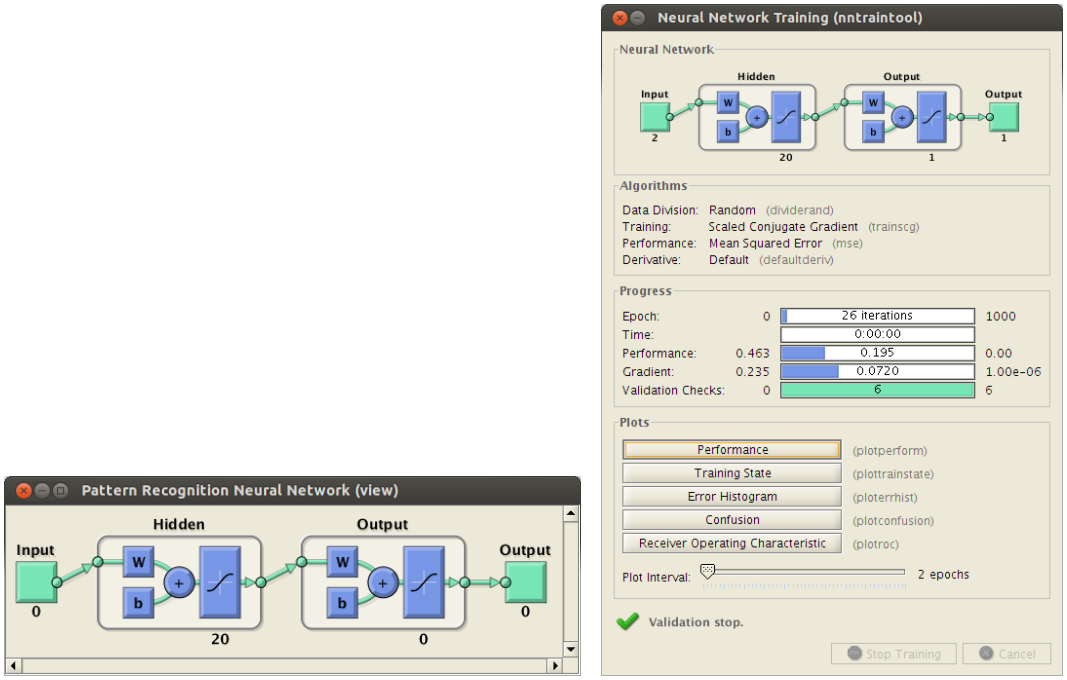


Figure 4.2: Created network and simulation network with 20 neurons in the hidden layer

and $A2 =$ which is the normalized intensities or target into the Matlab prompt using: load *.mat where * represents A1 or A2.

2. Create a pattern recognition network which is a feed-forward network with sigmoid function in both the hidden layer and output layer using:

   net=newpr(A1, A2, n) or net= patternnet(n)

   or a regression network with net=newfit(A1, A2, n)

   where n is the number of neurons in the hidden layer

3. Train the network: Defaults are scaled conjugate gradient algorithm for pattern recognition and Levenberg-Marquardt algorithm for regression. Both randomly divide the input and target vectors into three sets: 60 percent used for training, 20 percent for validation that the network is generalizing and to stop training before overfitting, 20 percent used as an independent test of network generalization. The syntax is net= train(net, A1 ,A2)

4. Compute the output and error with [net, tr, y, e] =train(net, A1, A2)

   where net is the network created, tr are the evaluation parameters, y is output, e is the mean square error (MSE) which is the average squared error between the network outputs y and the target output A2 given by

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y - A2)^2 \tag{4.1}$$

4.2.1 Data evaluation functions

In Matlab neural networks, there are functions that help check whether results obtained from our dataset are consistent and make sense. Again, this data set is adopted from the Cambridge Reference Sequence with ascension number NC_012920. Three of them which I will be making mainly use of are:

Performance : This is a plot of the training, validation and test errors. It shows the

mean square error MSE dynamics in a logarithmic scale. The training MSE is always decreasing and the least. Validation and test MSE are of more interest and are supposed to be similar for a near perfect training. Training on the data set normally stops when there is a consistent increase in the validation error for a given number of iterations (epochs). The best performance is taken from the epoch with the lowest validation error. Figure 4.3 shows a typical performance plot.

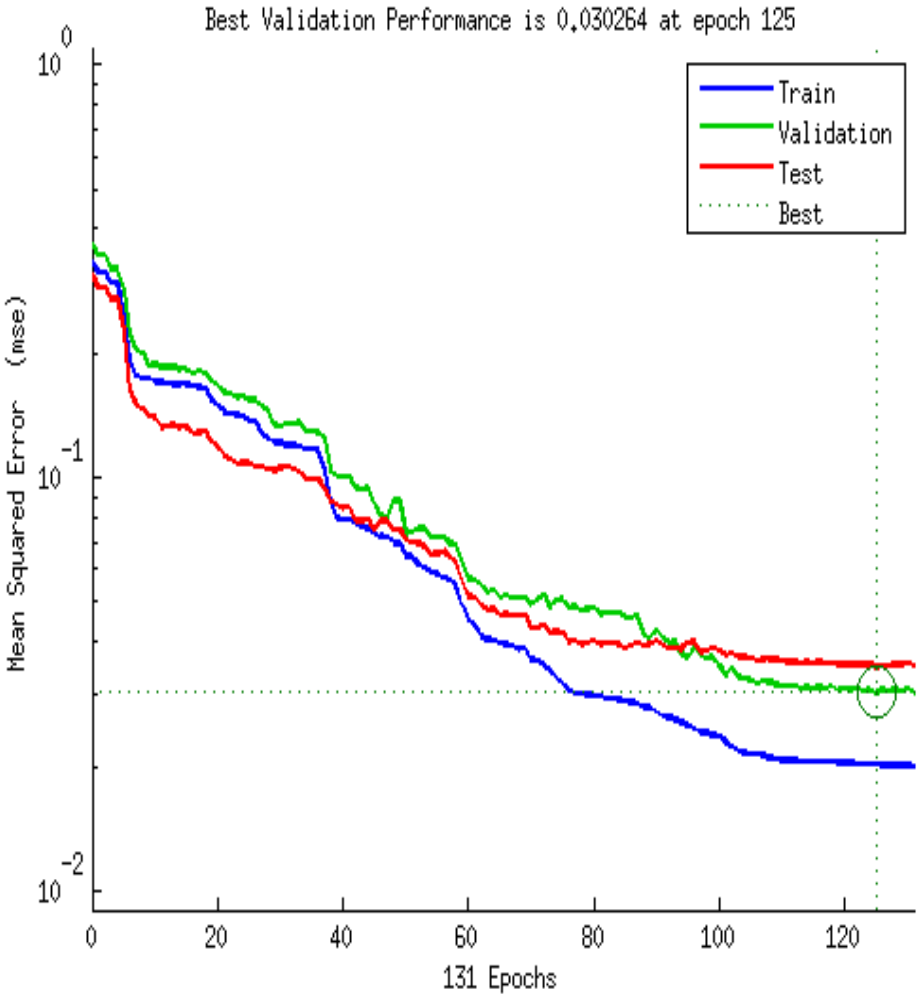


Figure 4.3: A typical performance plot showing training, validation and test errors in terms of MSE.

Confusion Matrix : This is a 2-dimensional matrix with a row and column for each class. Each matrix element shows the number of test examples for which the actual class is the row and the predicted class is the column. Good results correspond to large numbers down the main diagonal. The diagonal (green cells) in each table show the number of cases that were correctly classified. The off-diagonal (red cells) show the misclassified cases. Blue cells in the bottom right show the total percent of correctly classified cases (in green text) and the total percent of misclassified cases (in red text).

Figure 4.4 shows a confusion matrix with 4 tables each displaying the network response for the training, validation, testing and all datasets.

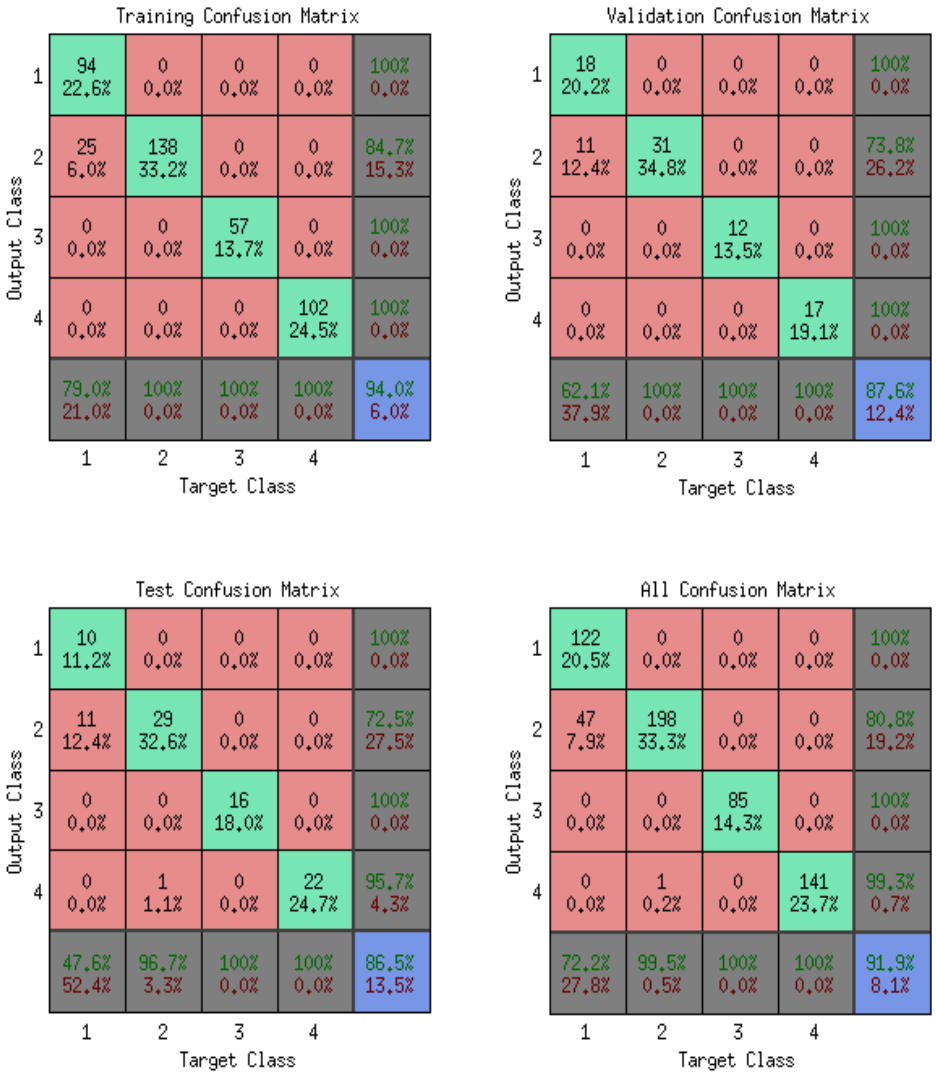


Figure 4.4: A typical confusion matrix showing various types of errors that occurred for the final trained network

Regression : This performs a linear regression analysis between the network outputs and the corresponding targets. The solid line represents the best fit linear regression line between outputs and targets. In an ideal situation, i.e. with zero error, the points are placed on the target=output line. High regression values are indication of good results. The scatter plot is helpful in showing that certain data points have poor fits. Figure 4.5 shows a regression plot.

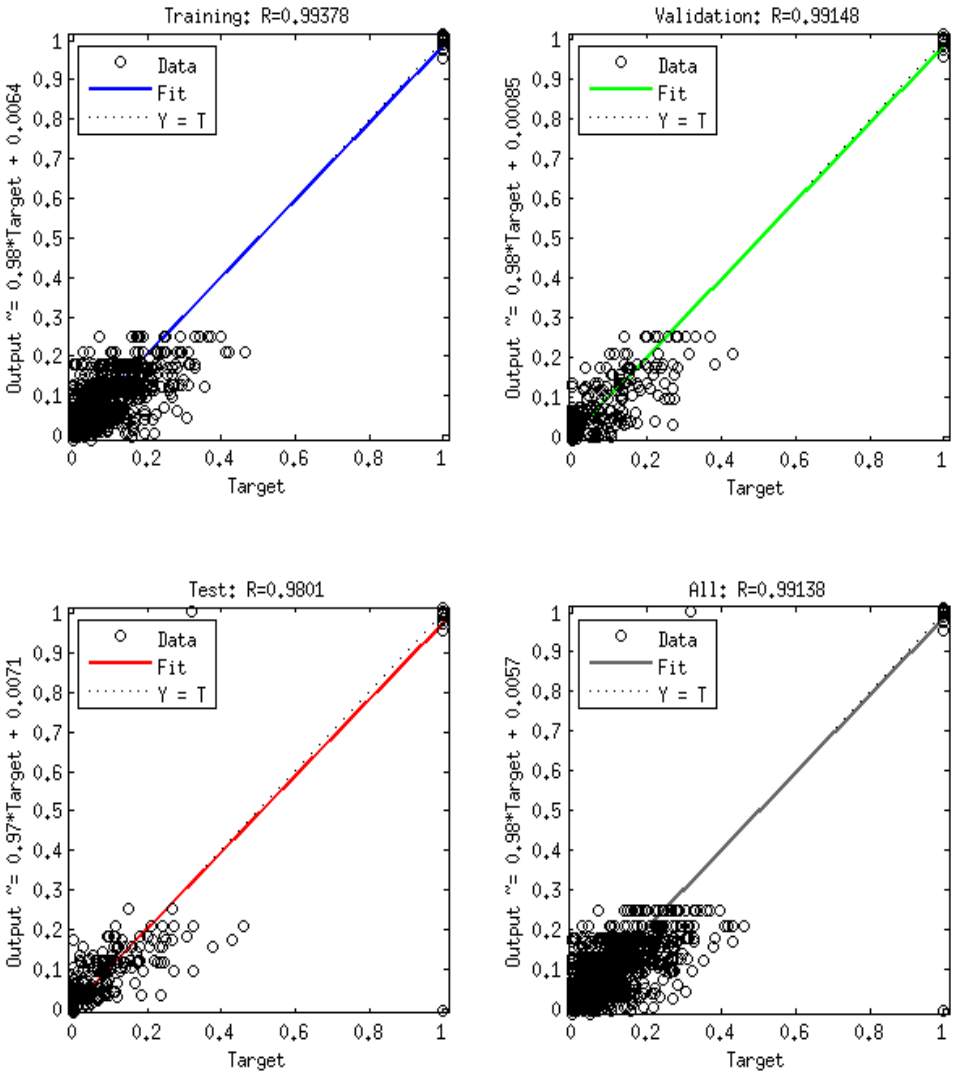


Figure 4.5: A typical regression plot

The fourth data evaluation function is the Receiver Operating Characteristic, ROC.

Receiver Operating Characteristic, ROC : ROC curve shown in Figure 4.6 is another form of visualization and analysis of the quality of our network. It is a plot of true positive rate (sensitivity) versus the false positive rate (1-specificity). The colored lines in each axis represent the ROC curves for each categoy of the problem. The ROC always goes through the origin and through $(1, 1)$. A good test would show points in the upper-left corner.

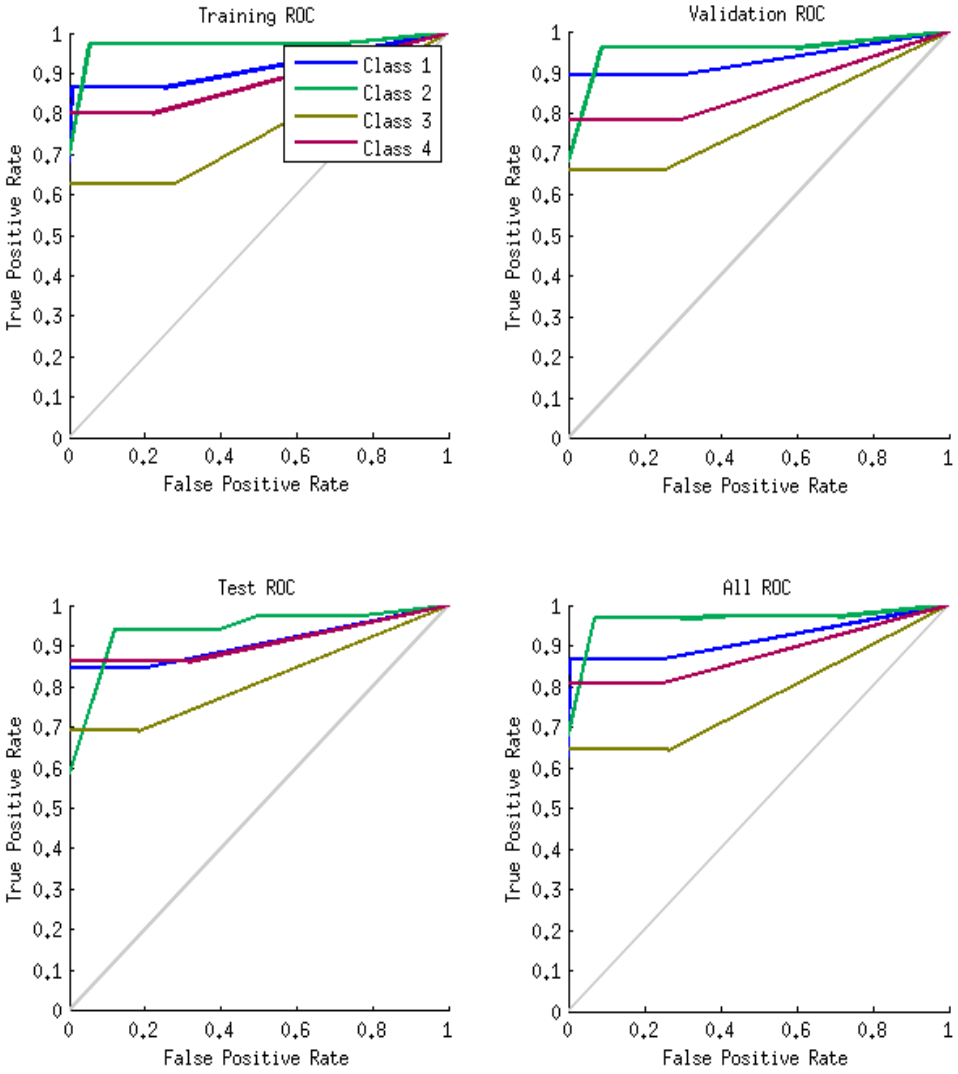


Figure 4.6: A typical ROC plot

4.3 Results

Our dataset is made of 15,453 rows and 6 columns where 2 of the columns are the n-grams for n= 1,2 and the other 4 columns represent the normalized intensities for Adenine, Cytosine, Guanine and Thymine. I extract every 26th line of the dataset which reduces the dataset to 594 rows (lines) respectively. I use 1-grams and 2-grams independently to predict the intensities for the four nucleotides ACGT. I also use a combination (composition) of the 1-grams and 2-grams to repeat my analysis. Neural networks are good at fitting functions. It is a generally held belief that a fairly simple neural network can fit any practical function. The regression value R, so computed by the neural network determines how robust the prediction is. The higher the R value the better and a smaller MSE in terms of performance implies good prediction. We compare the performances of the networks with 1-gram and 2-gram with different number of neurons in the hidden layer. The number of neurons in the hidden layer has been varied between 20 and 40 with step size 5 as a matter of choice and hopefully to find the optimal network architecture.

4.3.1 Analysis with 26th line (row) using Regression toolkit

The 1-gram performance plots for ACGT with 20 and 40 neurons in the hidden layer are shown in Figure 4.7 and Figure 4.8 respectively. Training stopped when the validation error increased for six iterations (by default). The best validation performance of 0.003209 occurred at iteration (epoch) 11 and 0.0031945 at epoch 4 for the two configurations respectively. Hence the epochs for the best validation are chosen using the above criteria i.e. when the validation error inceased for six iterations. This same interpretation applies to all the performance plots in this thesis. Both the test set error and the validation set error have similar characteristics which shows that the results are reasonable.

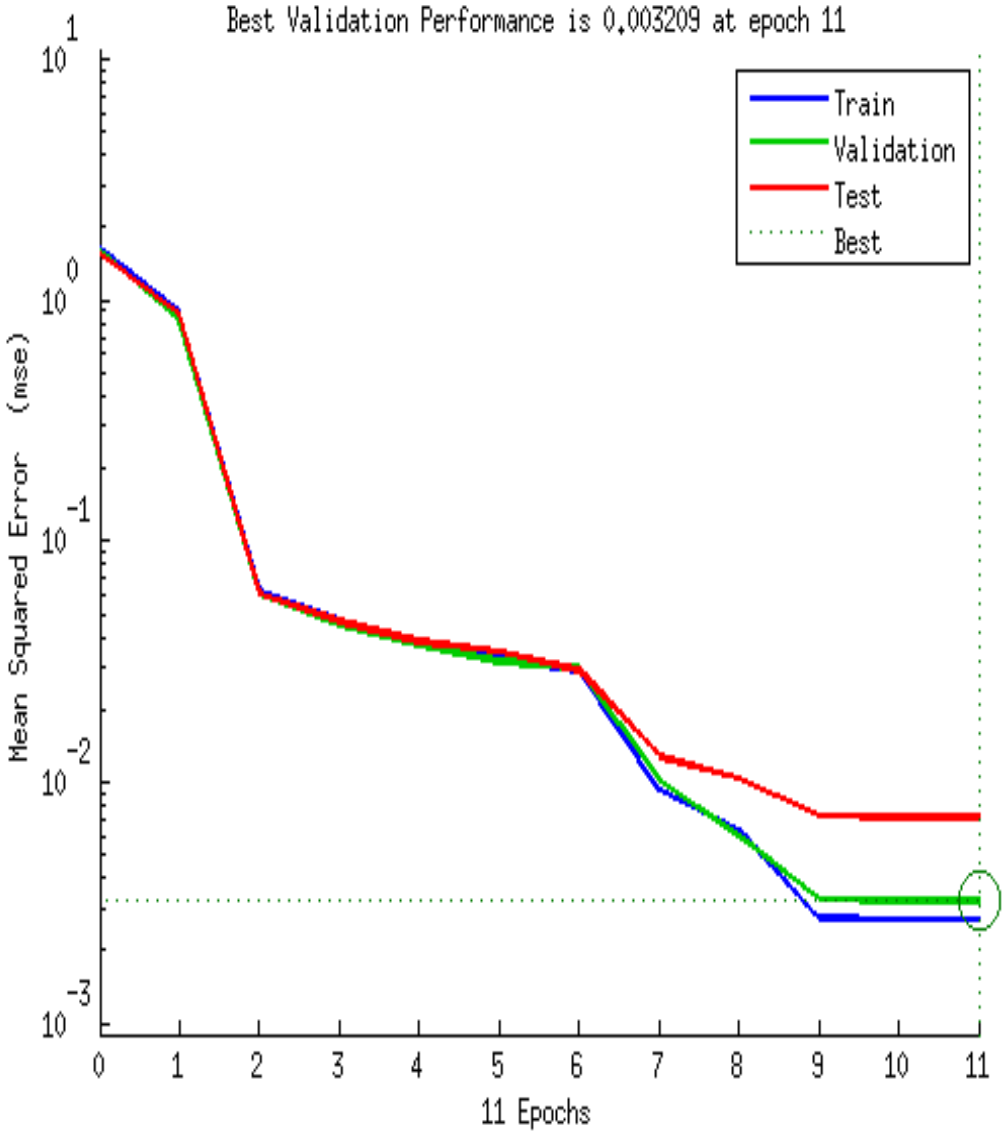


Figure 4.7: Performance plot with 20 neurons

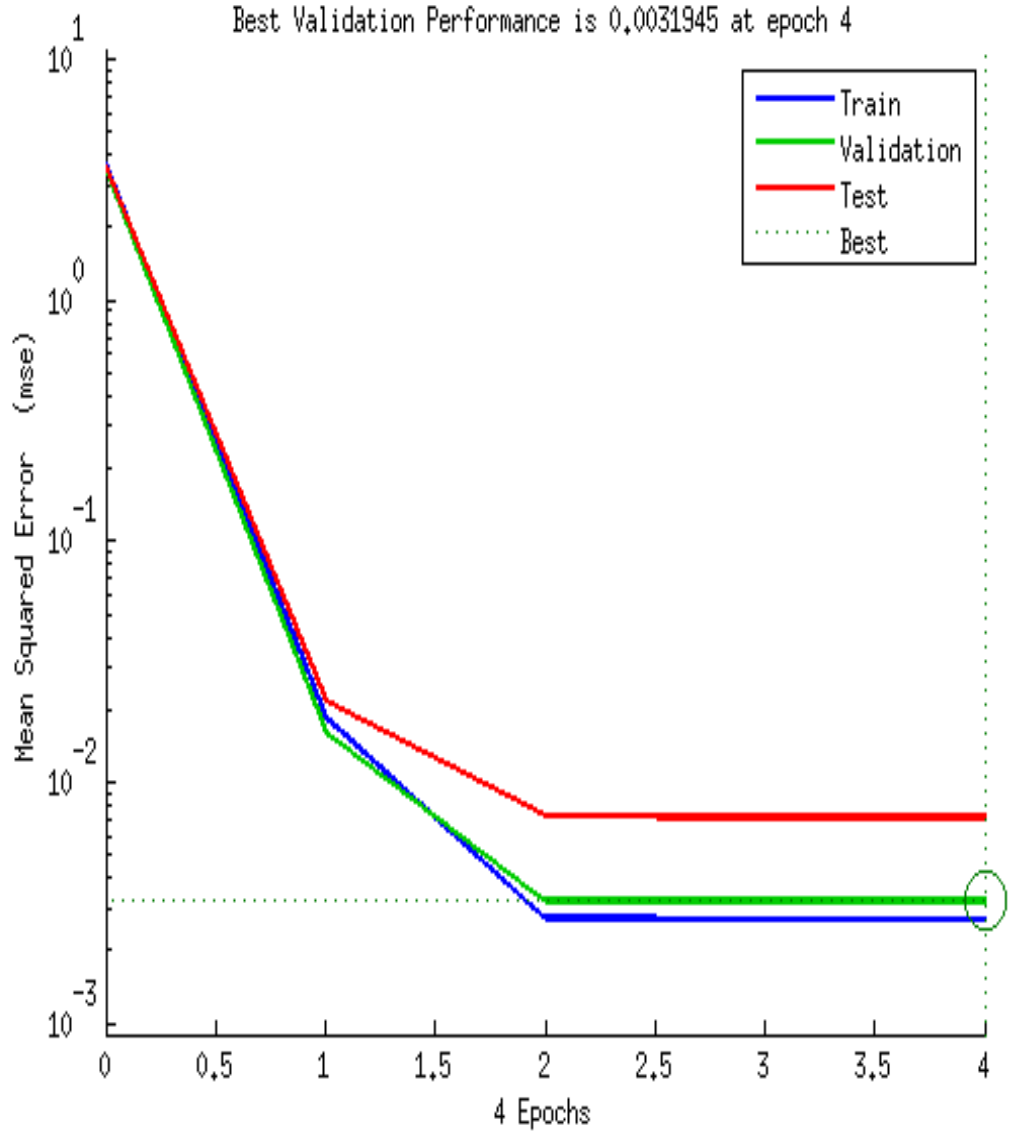


Figure 4.8: Performance plot with 40 neurons

The corresponding 1-gram regression plots are shown in Figure 4.9 and Figure 4.10 respectively. The output tracks the targets very well for training, testing and validation. The R-values are over 0.98 for the total response.

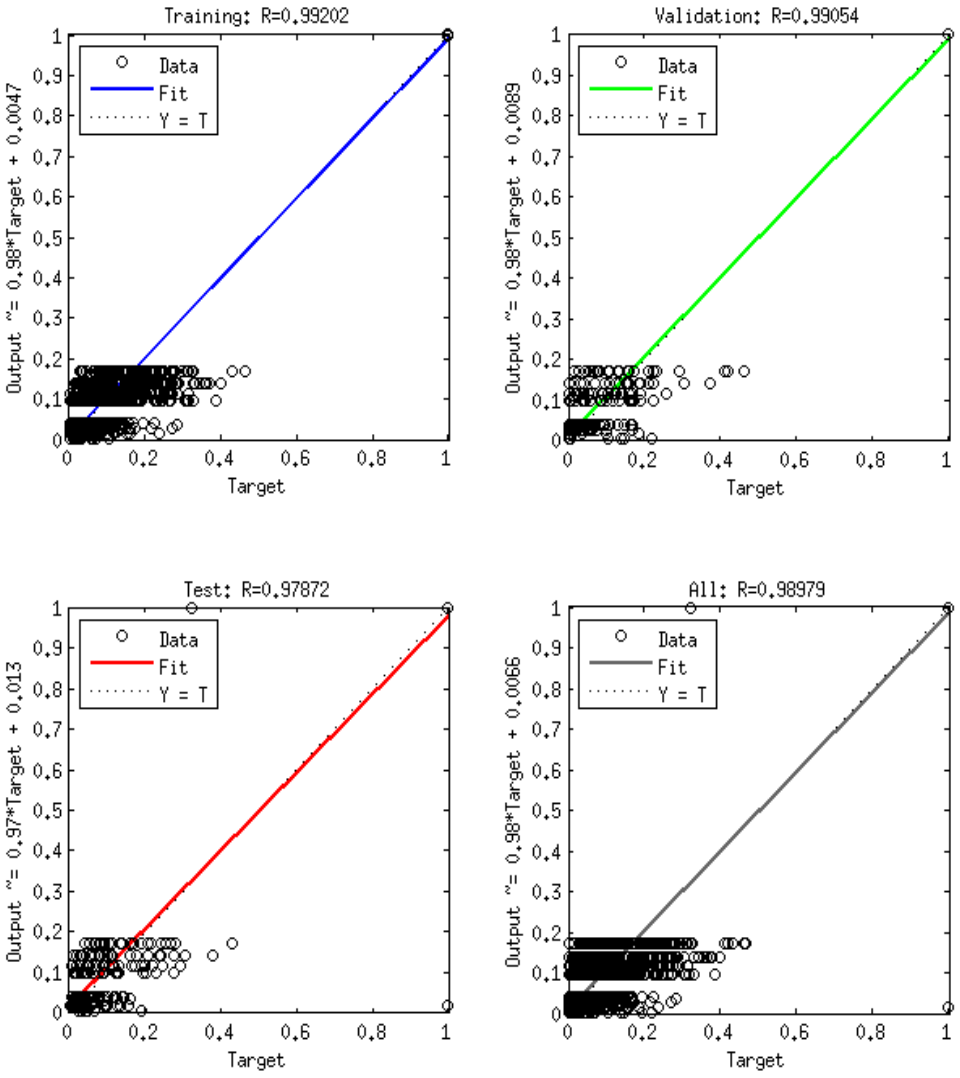


Figure 4.9: Regression plot with 20 neurons

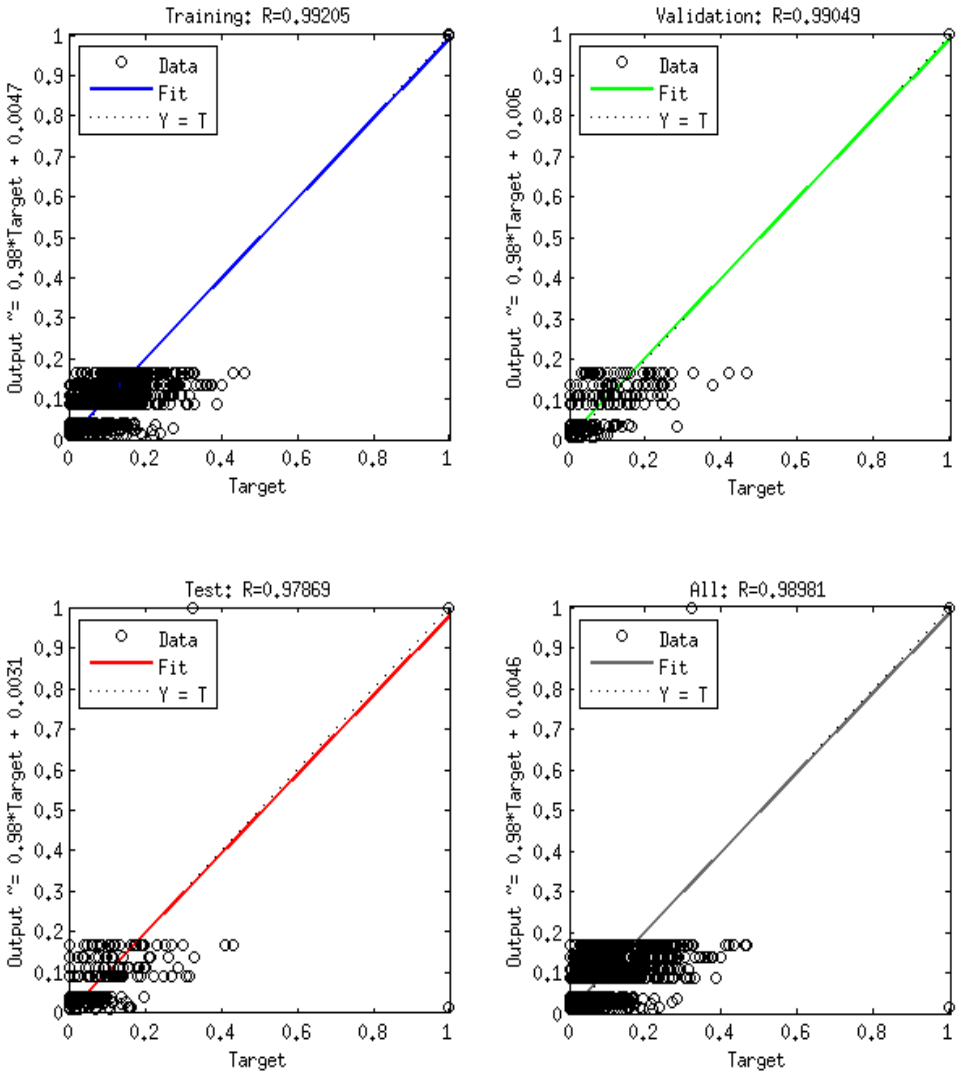


Figure 4.10: Regression plot with 40 neurons

The 2-gram performance plots for ACGT with 20 and 30 neurons in the hidden layer are shown in Figure 4.11 and Figure 4.12 respectively. We note the best validation performances of 0.027319 and 0.02278 respectively for those different number of neurons in the hidden layer.

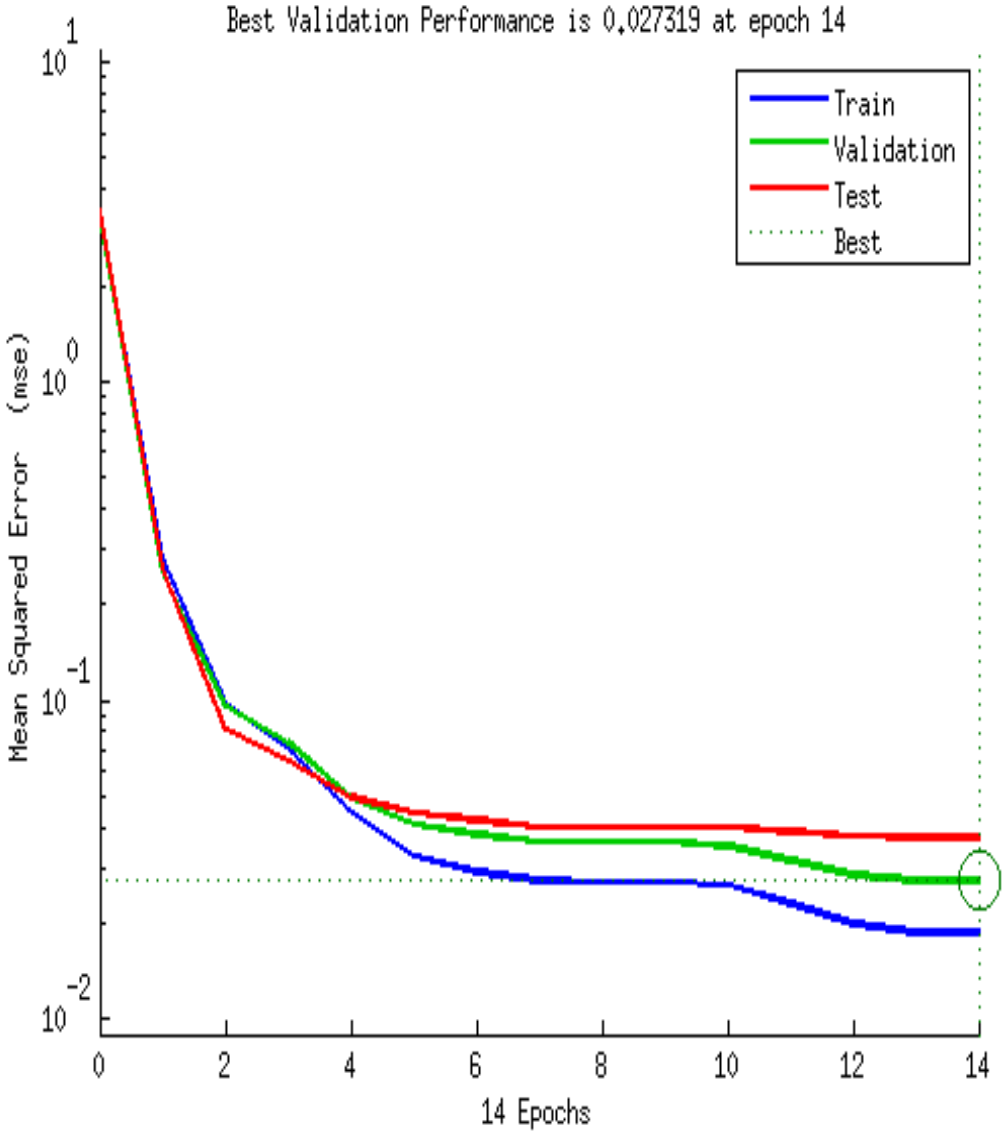


Figure 4.11: Performance plot with 20 neurons

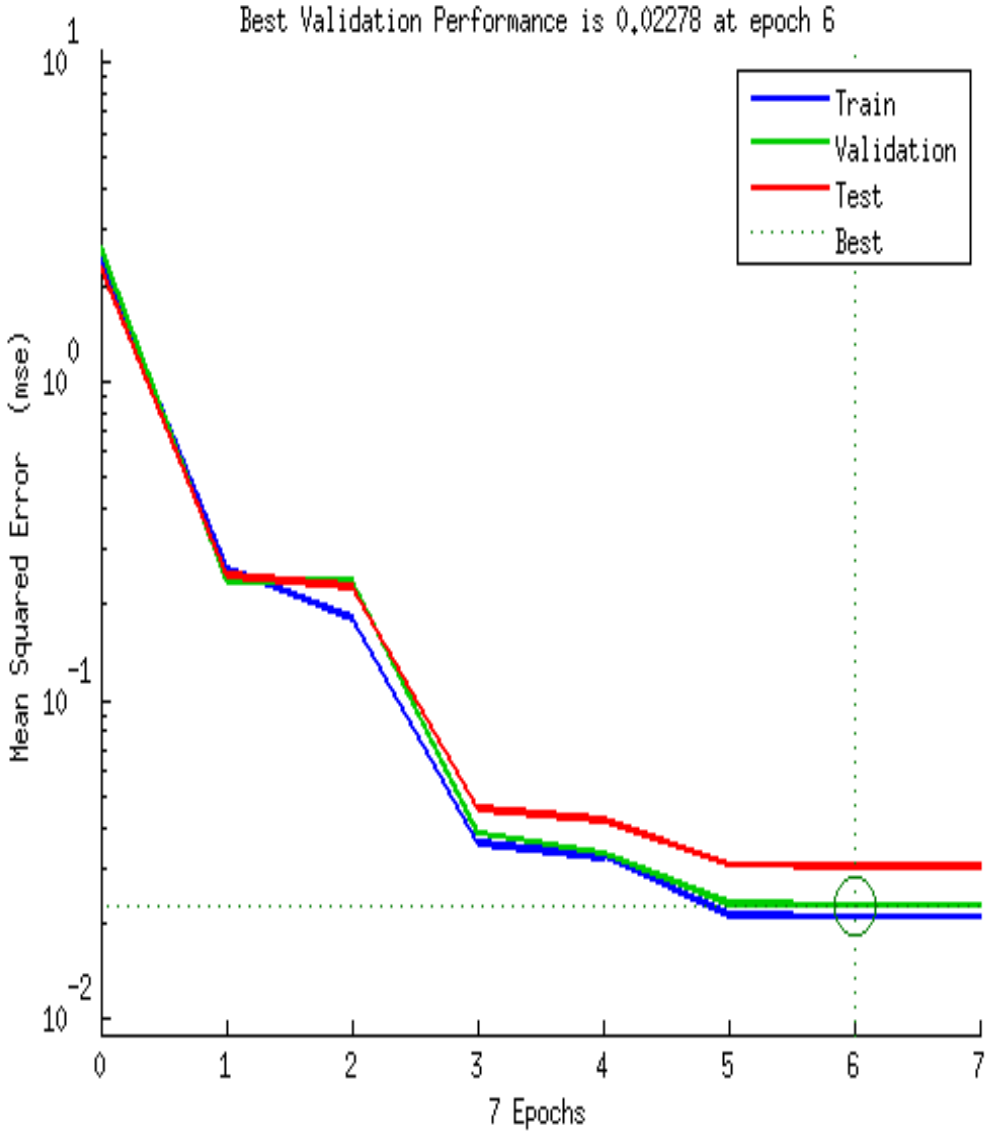


Figure 4.12: Performance plot with 30 neurons

The corresponding 2-gram regression plots are shown in Figure 4.13 and Figure 4.14 respectively. Again, the R-values are over 0.93 for the total response. The R-values here are worse than the ones gotten when we used 1-gram.

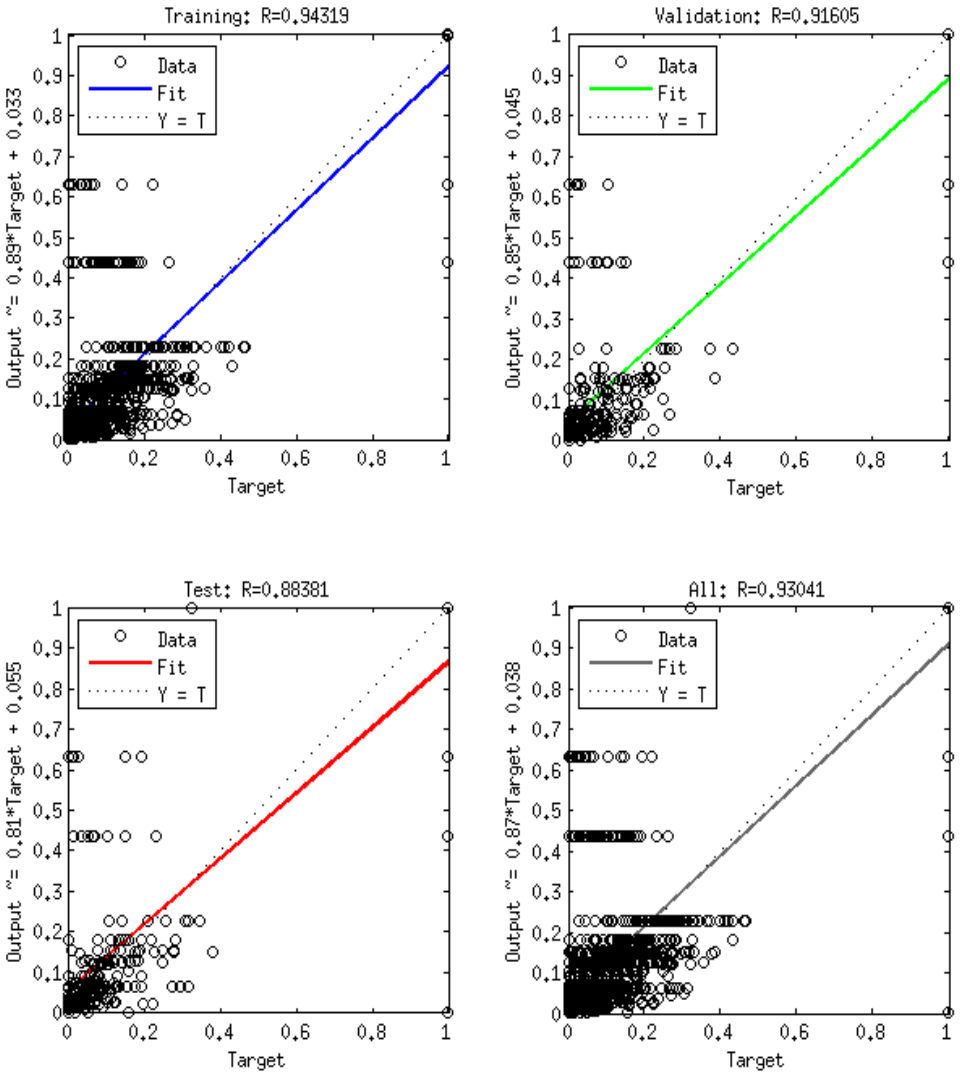


Figure 4.13: Regression plot with 20 neurons

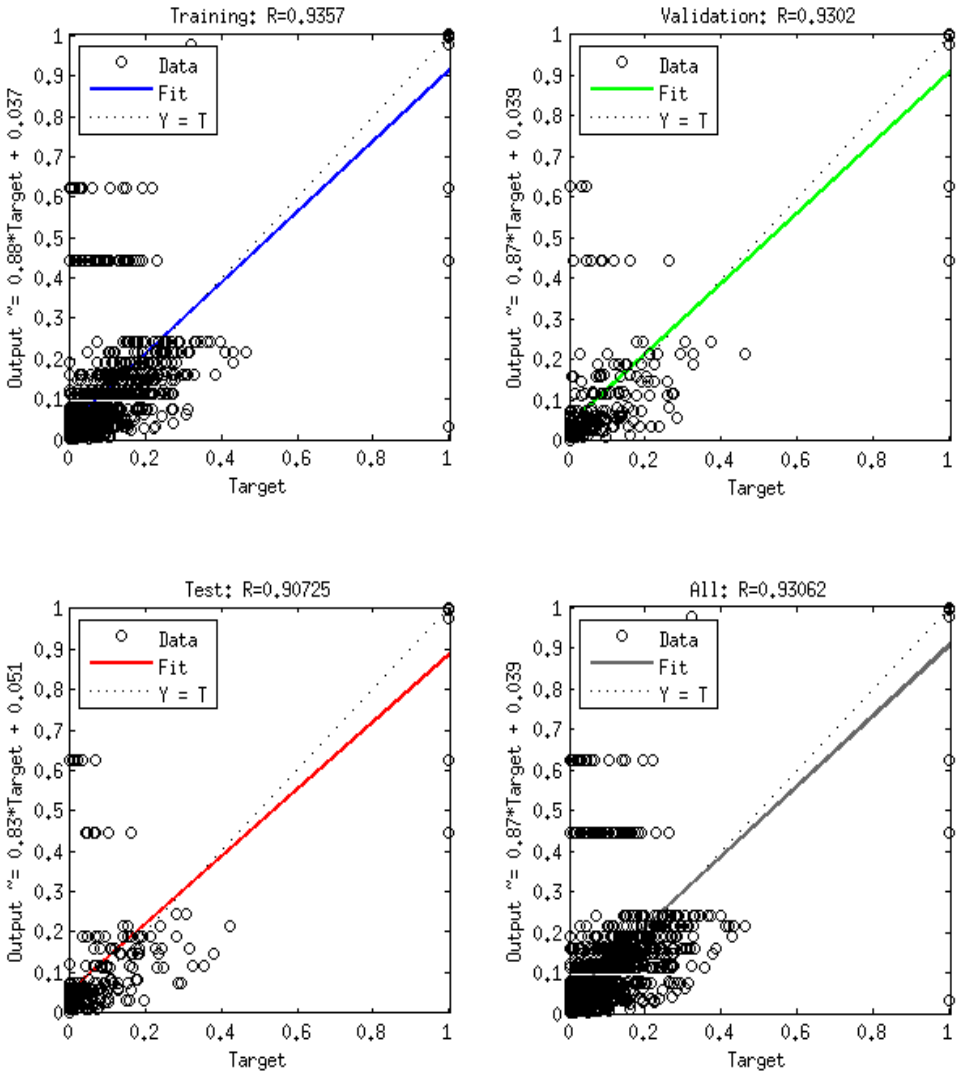


Figure 4.14: Regression plot with 30 neurons

The 1-2-gram performance plots for ACGT with 20, 25 and 40 neurons in the hidden layer are shown in Figures 4.15, 4.16 and 4.17 respectively. Again, the best validation performances of 0.0028494 occurred at epoch 32, 0.002663 at epoch 56 and 0.0025253 at epoch 45 for the three configurations respectively. These errors are significantly low. Both the test set error and the validation set error have similar characteristics which shows that the results are reasonable.

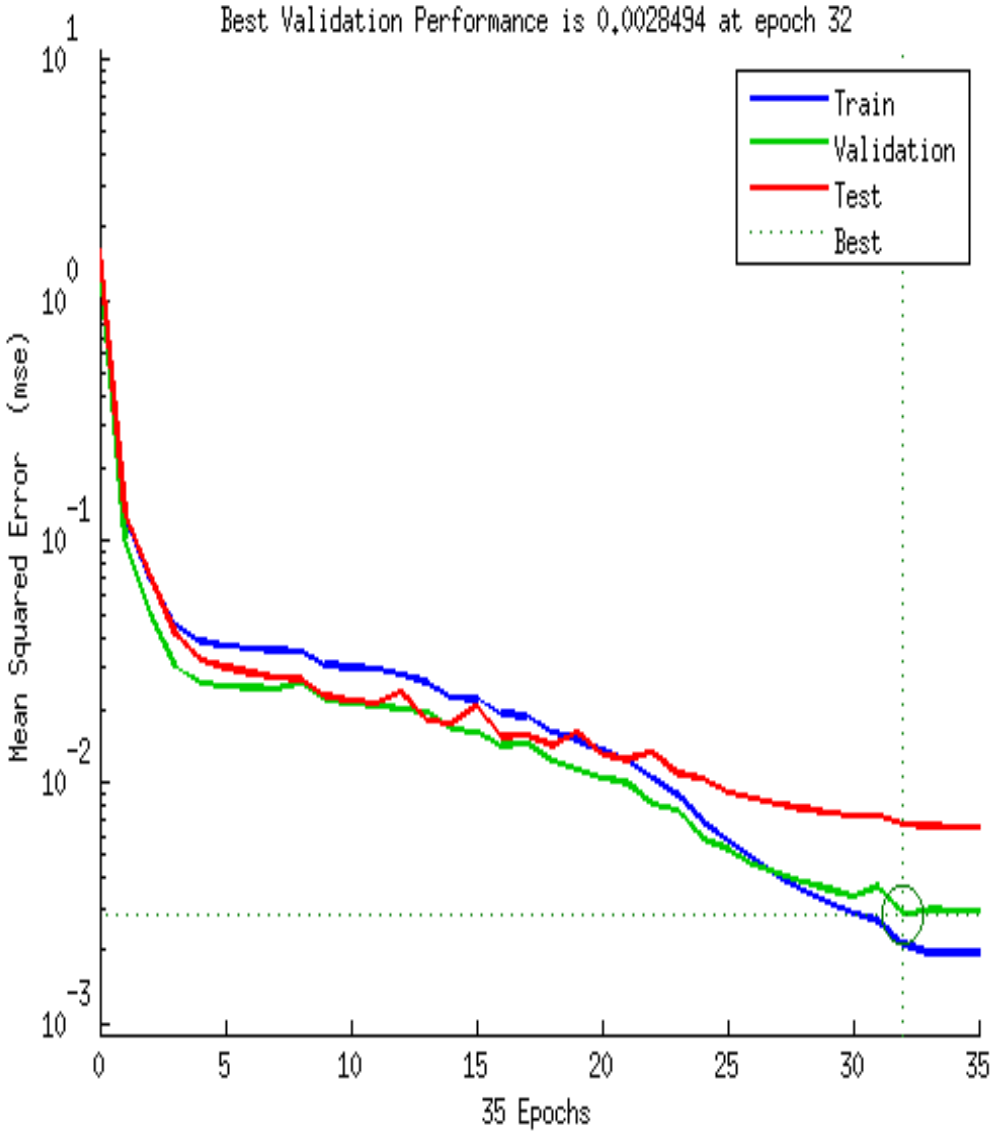


Figure 4.15: Performance plot with 20 neurons

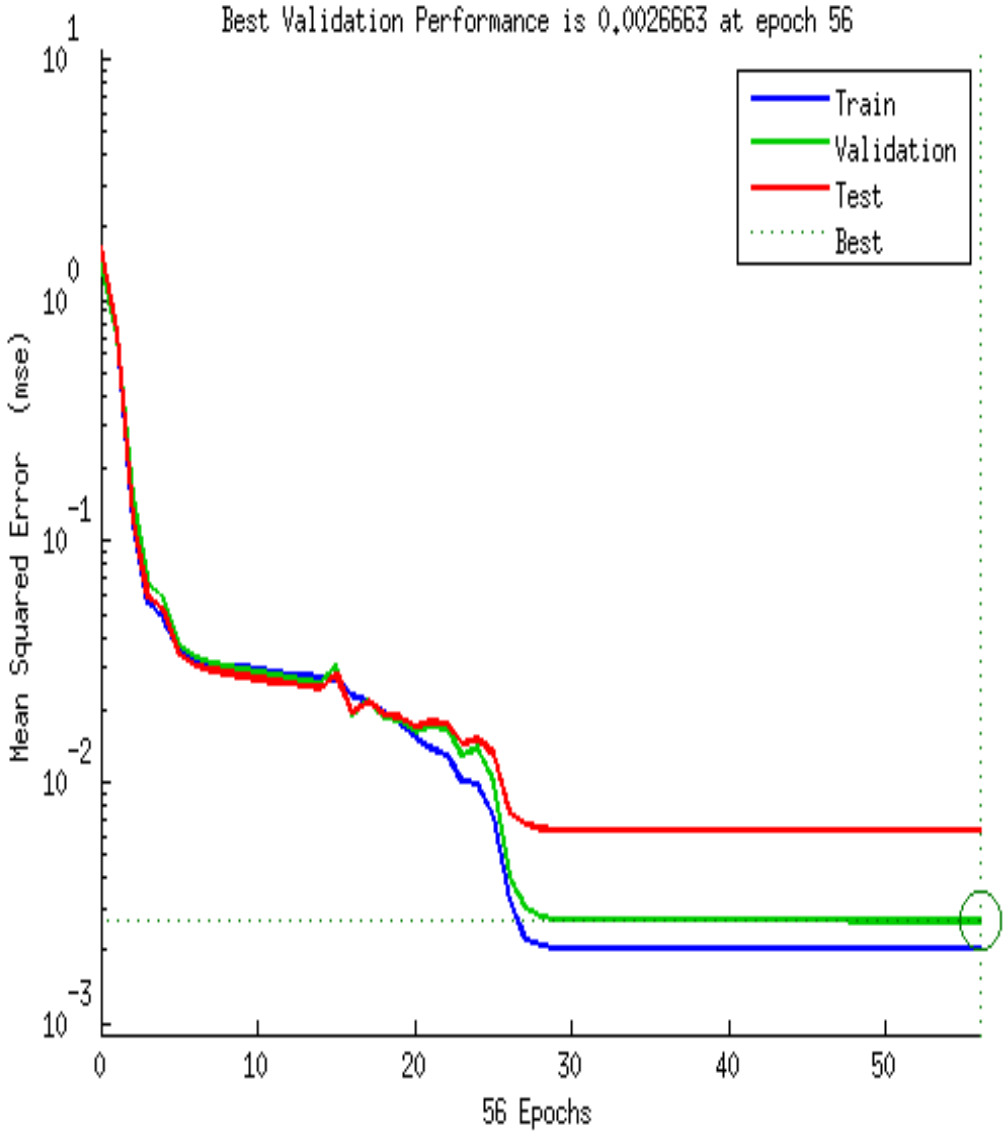


Figure 4.16: Performance plot with 25 neurons

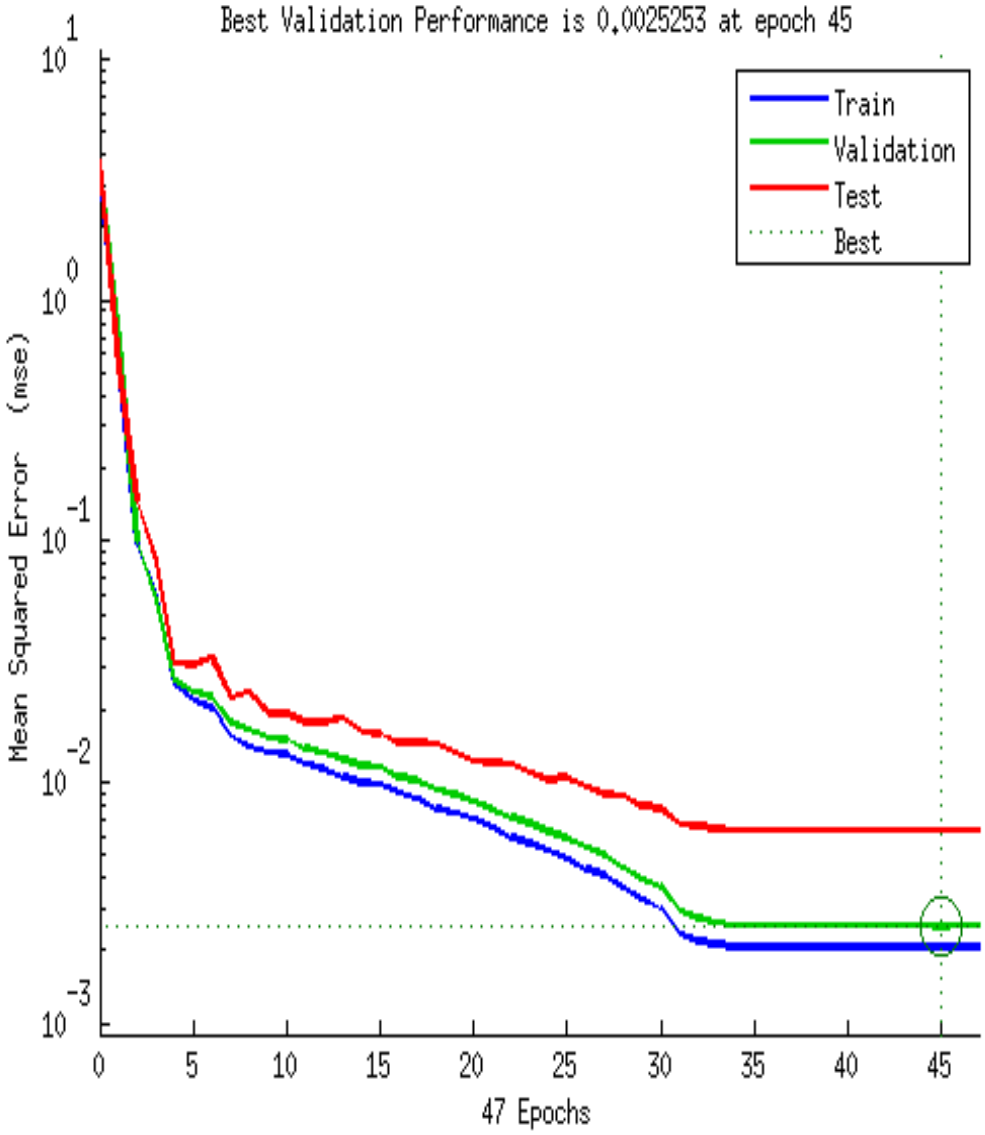


Figure 4.17: Performance plot with 40 neurons

The corresponding 1-2-gram regression plots for ACGT with 20, 25 and 40 neurons in the hidden layer are shown in Figures 4.18, 4.19 and 4.20 respectively. We notice higher regression values for the three configurations when we used this composition as compared with using either 1 or 2-gram independently.

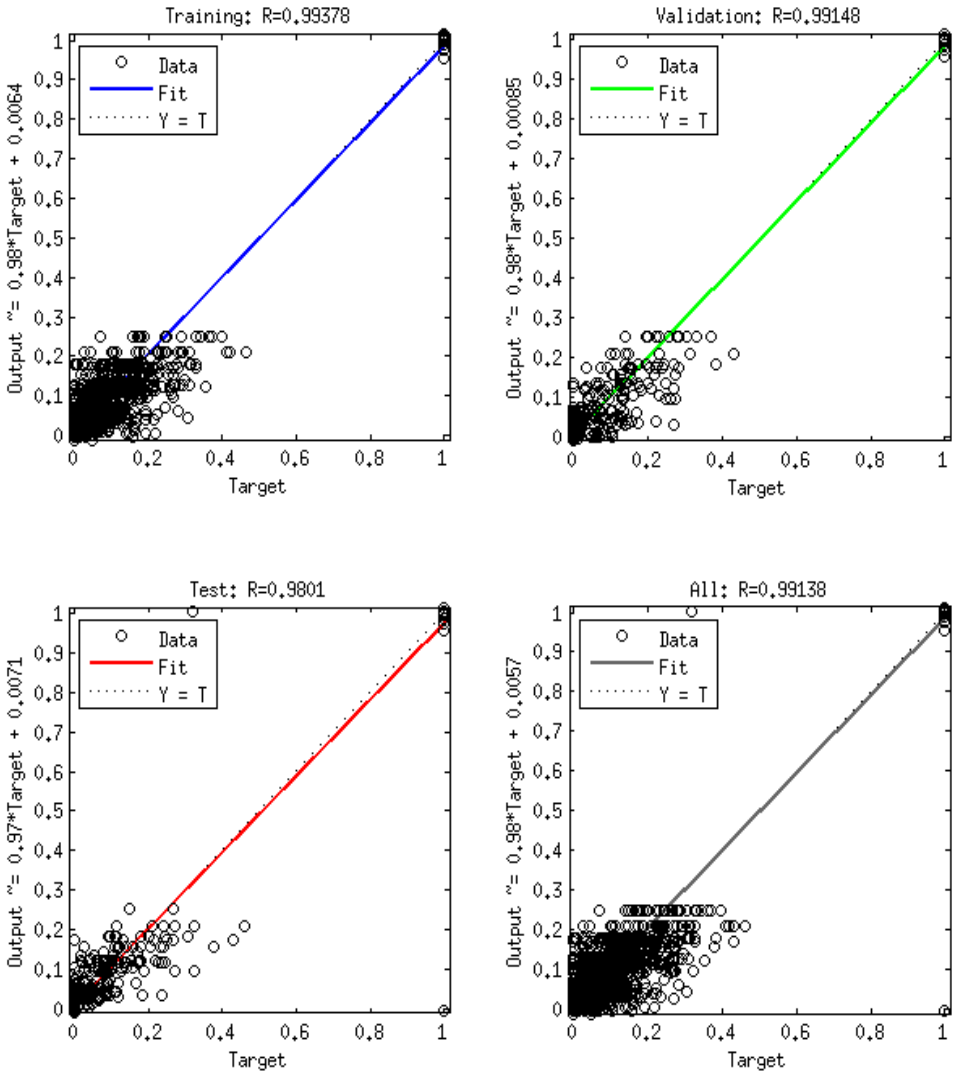


Figure 4.18: Regression plot with 20 neurons

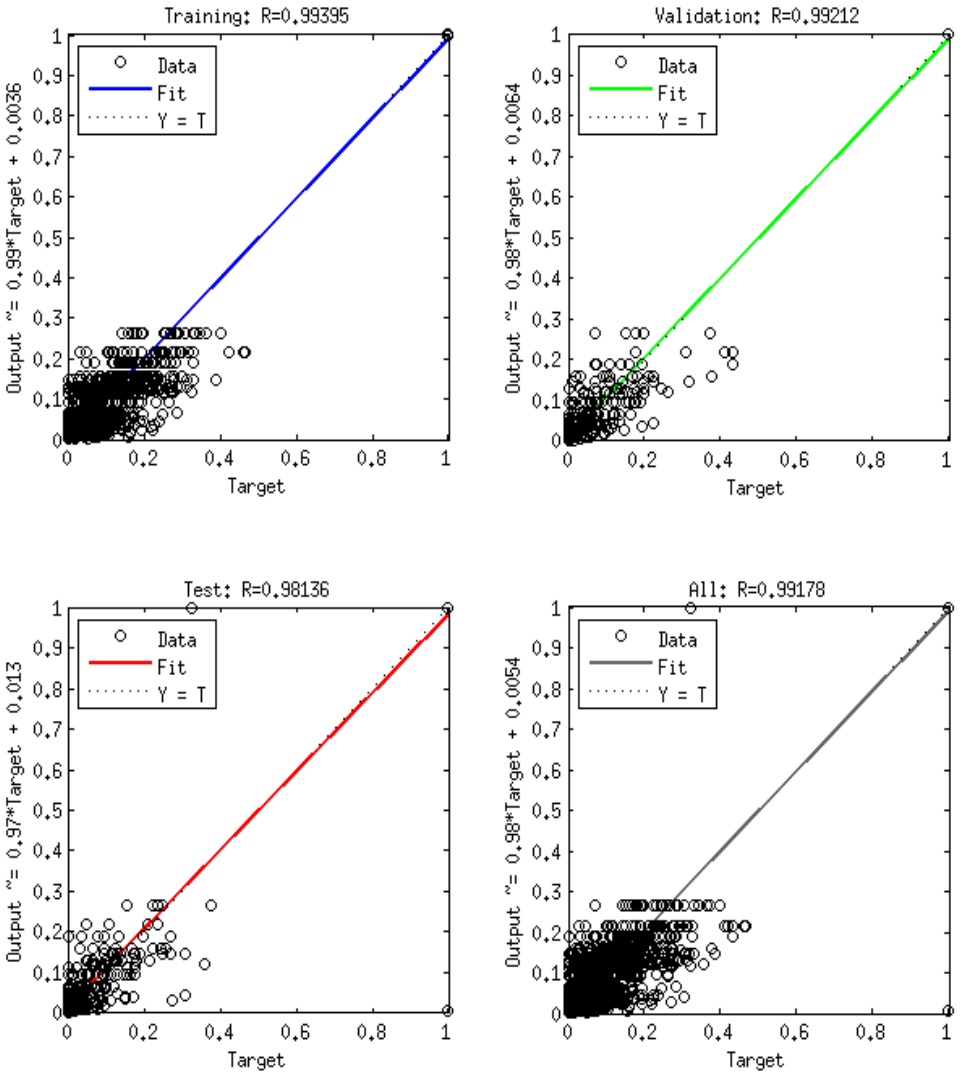


Figure 4.19: Regression plot with 25 neurons

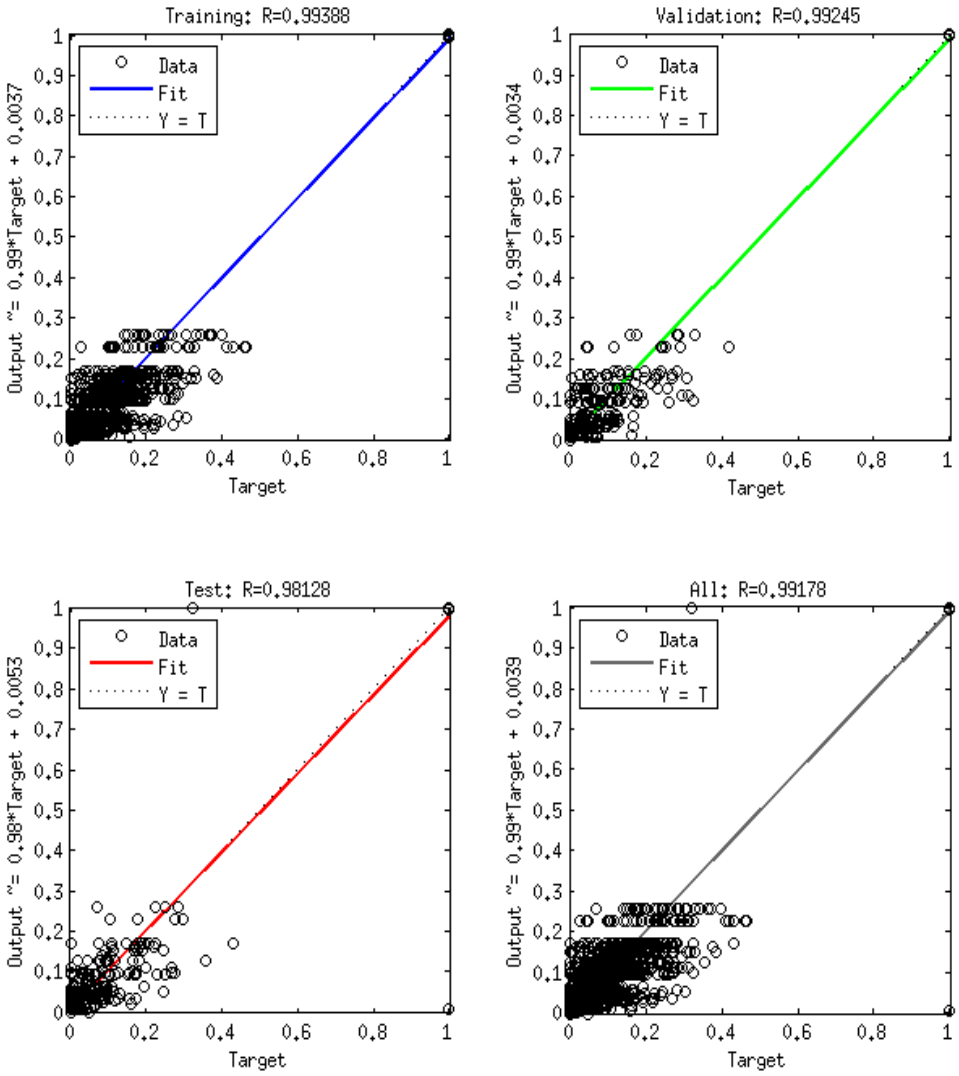


Figure 4.20: Regression plot with 40 neurons

4.3.2 Use of pattern recognition toolkit in Matlab

We make use of the pattern recognition network which is a feed-forward network with sigmoid functions in both the hidden layer and output layer with 20 and 40 neurons in one hidden layer. Plots of performance, confusion matrix and ROC curves were obtained with 1-gram, 2-gram and 1-2-gram composition to predict the signal intensities of the ACGT using 26th line (row).

Figures 4.21 and 4.22 are respectively, performance plots using 1-gram with 20 and 40 neurons in the hidden layer.

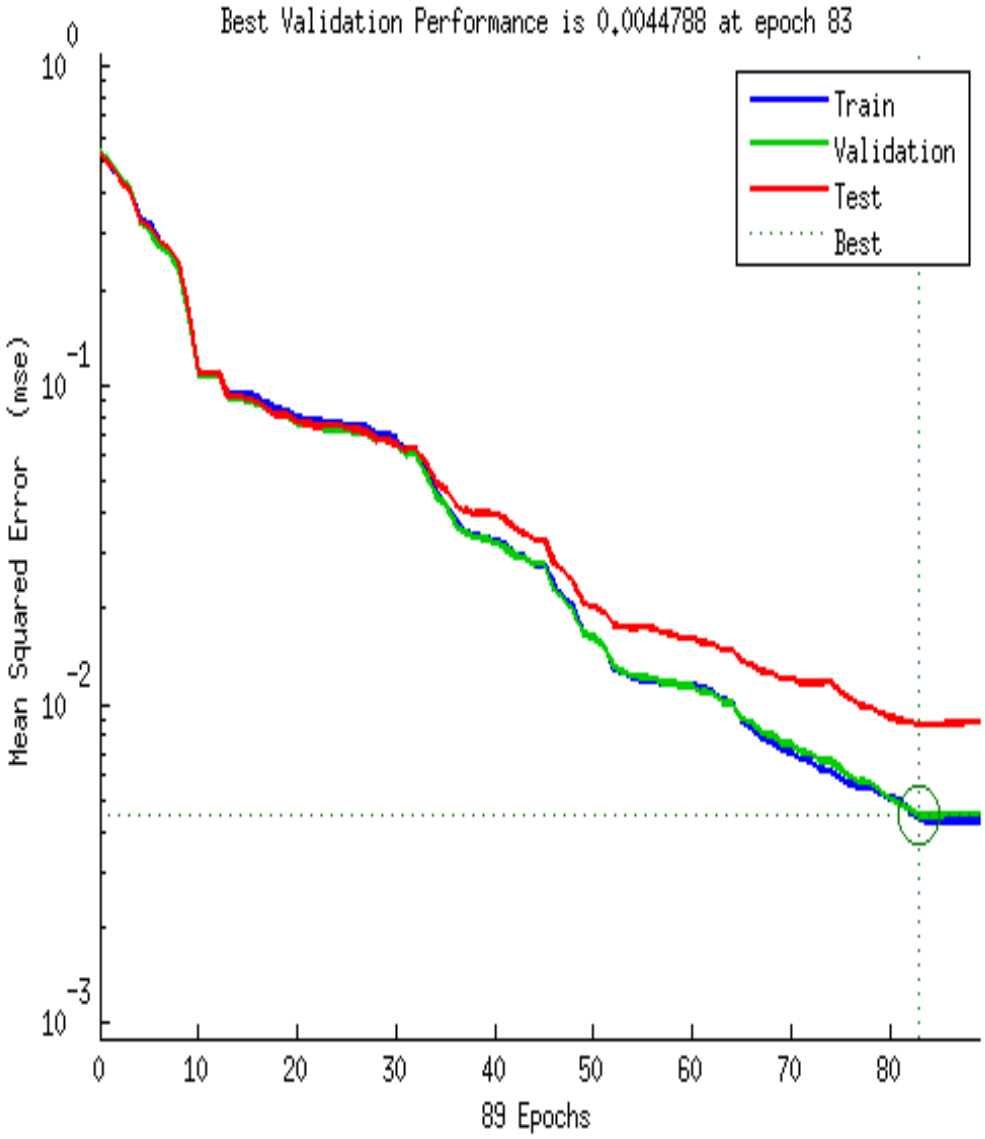


Figure 4.21: Performance plot with 20 neurons

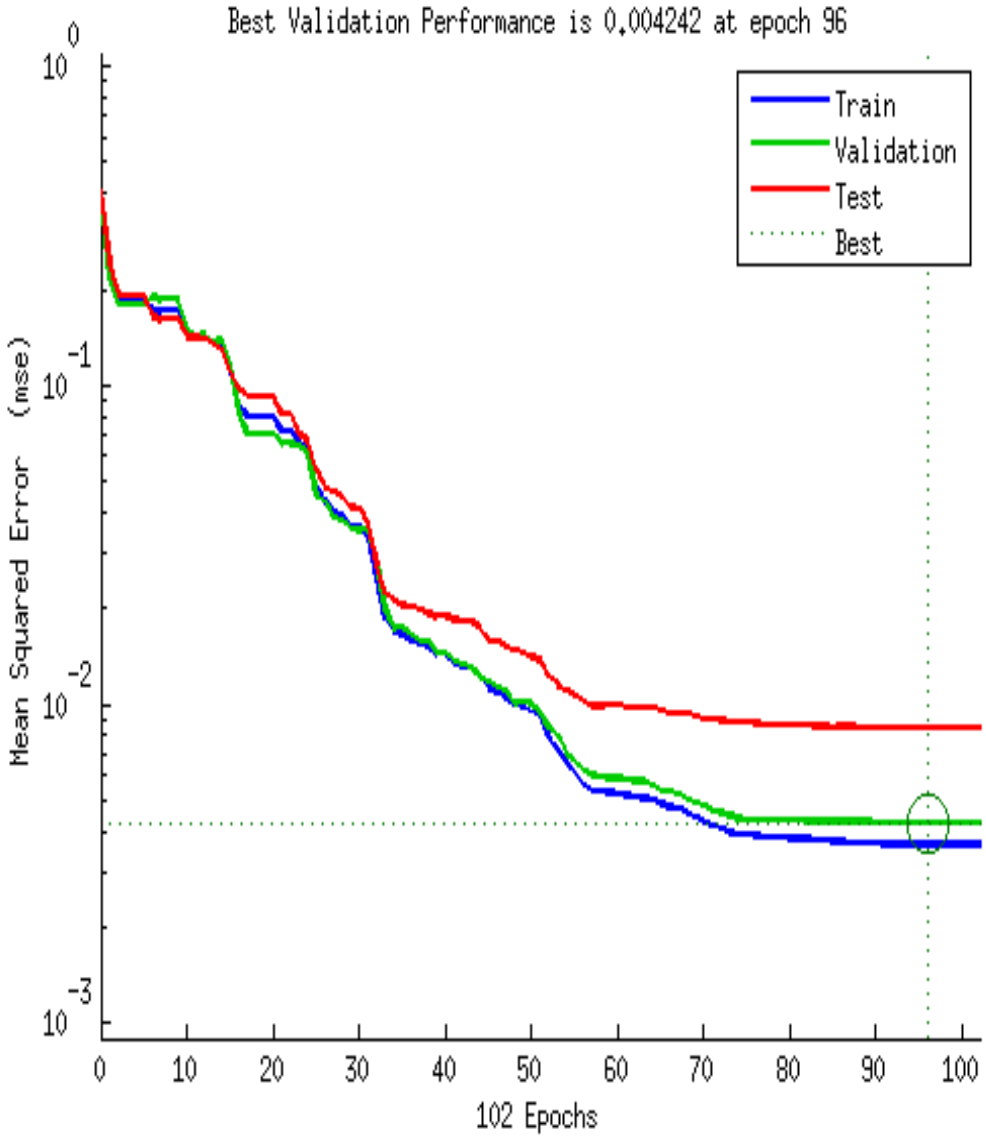


Figure 4.22: Performance plot with 40 neurons

The corresponding confusion matrix plots are shown in Figures 4.23 and 4.24.

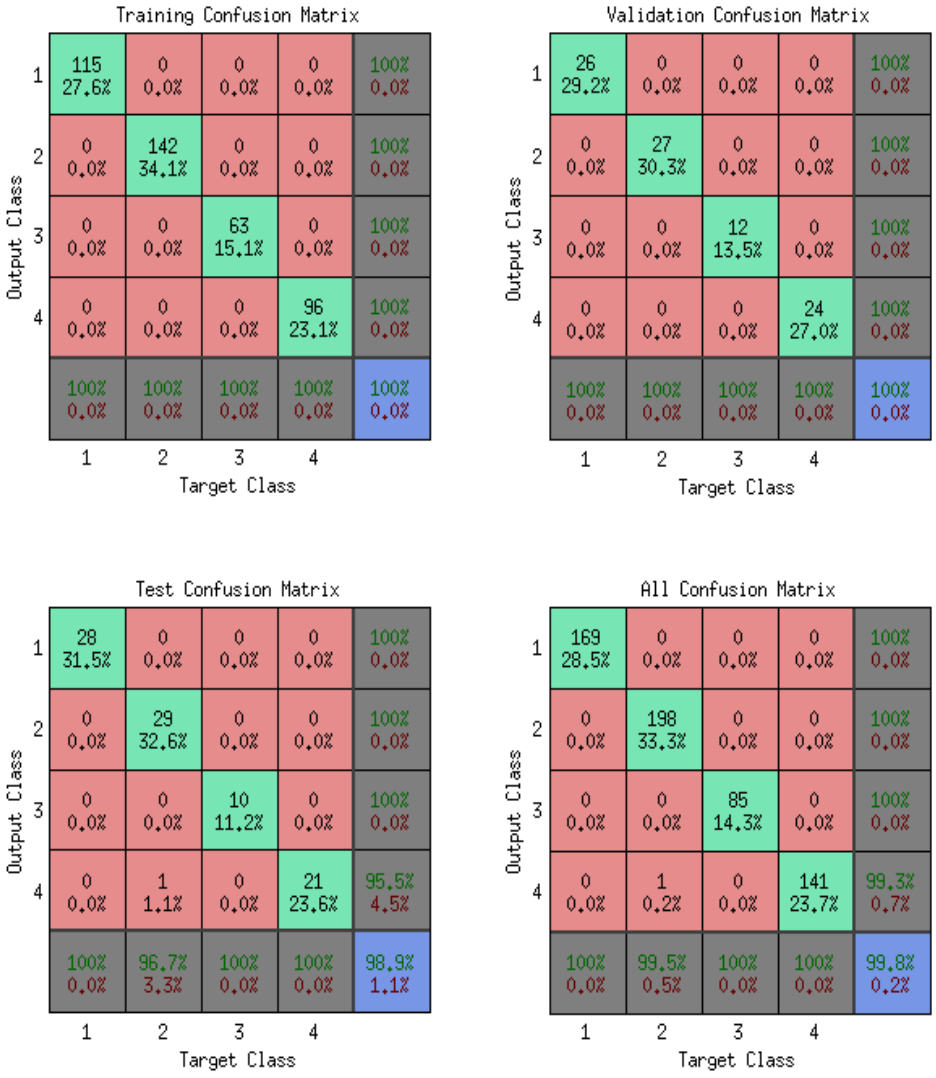


Figure 4.23: Confusion matrix with 20 neurons

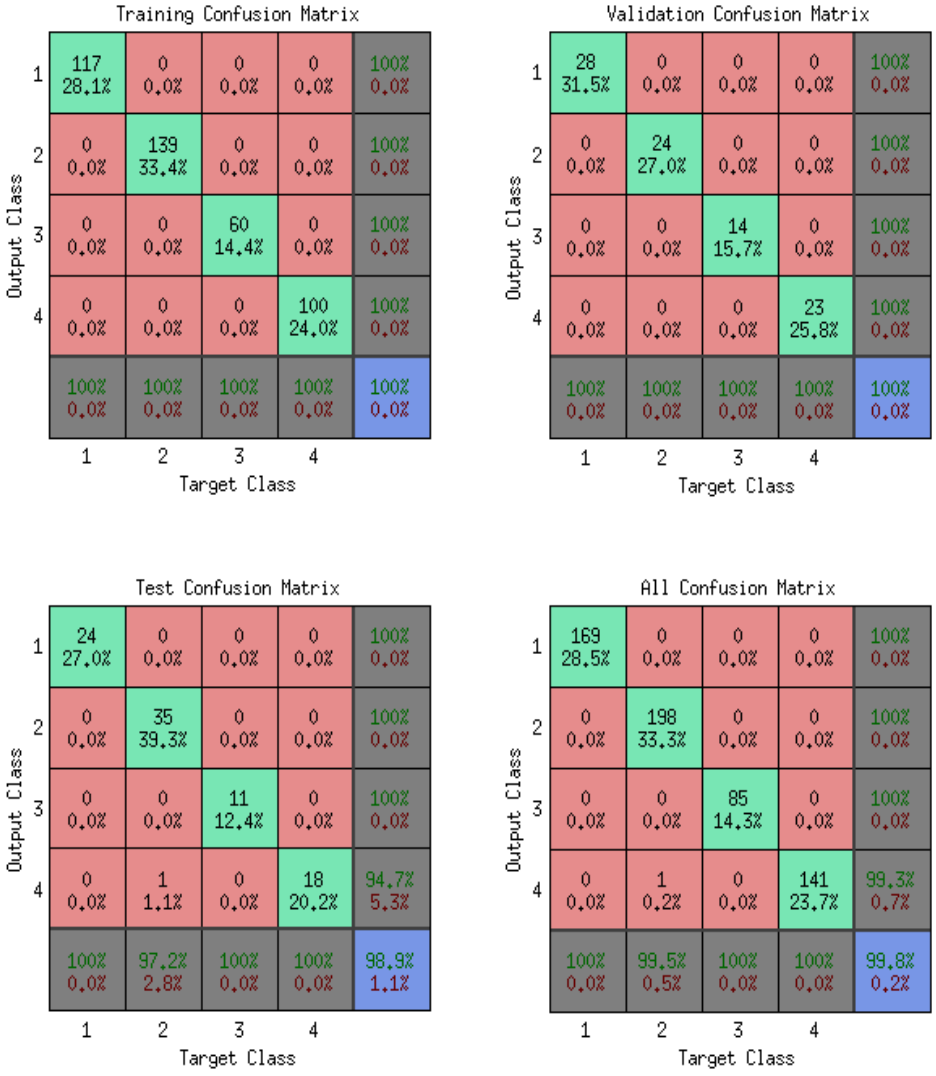


Figure 4.24: Confusion matrix with 40 neurons

The corresponding ROCs are shown in Figures 4.25 and 4.26.

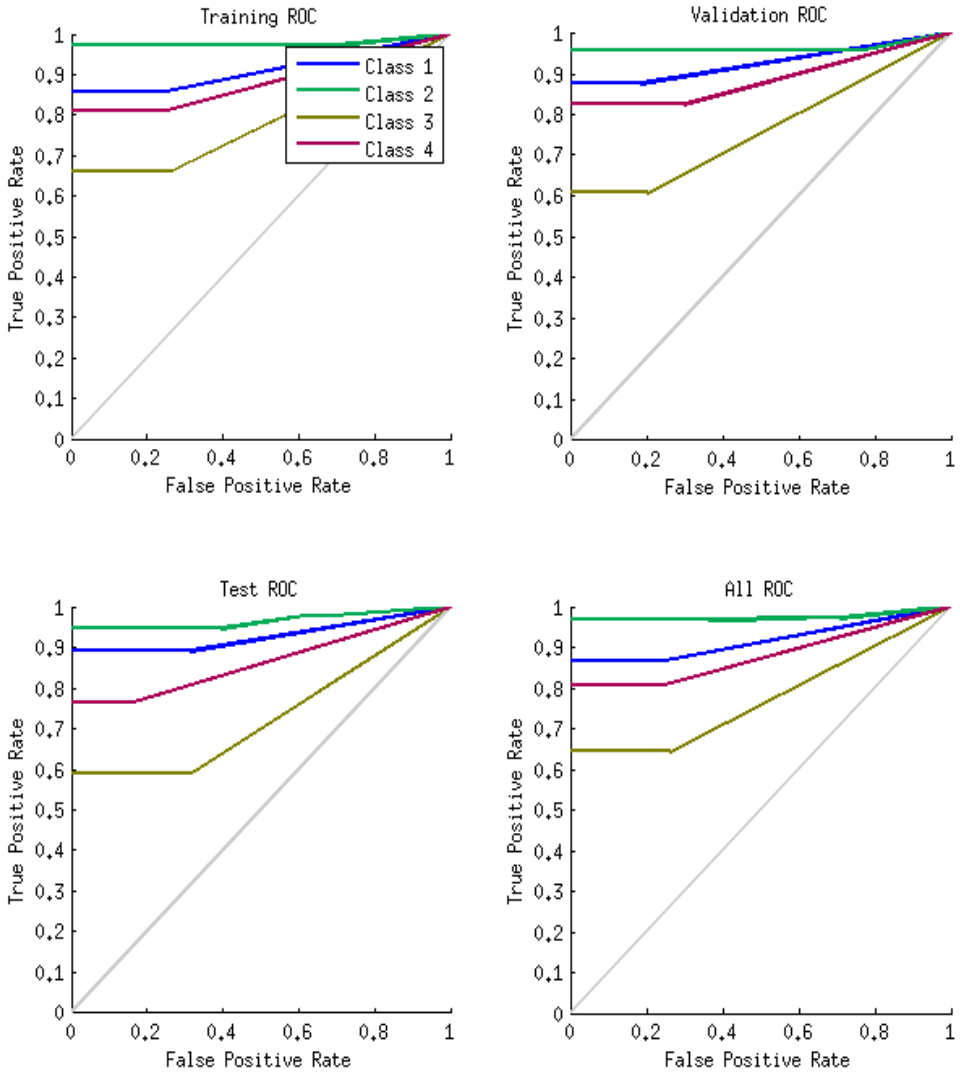


Figure 4.25: ROC with 20 neurons

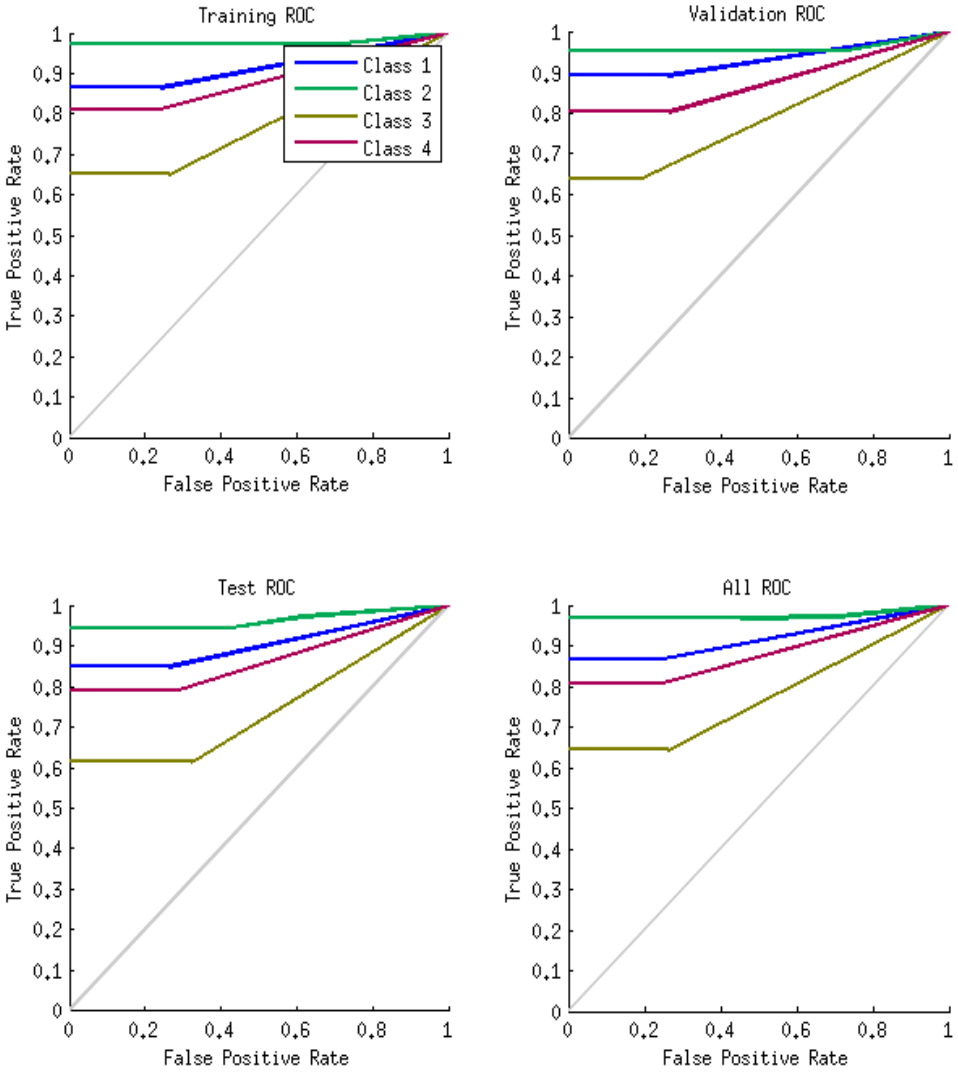


Figure 4.26: ROC with 40 neurons

Figures 4.27 and 4.28 are performance plots using 2-gram with pattern recognition toolkit with 20 and 40 neurons in the hidden layer.

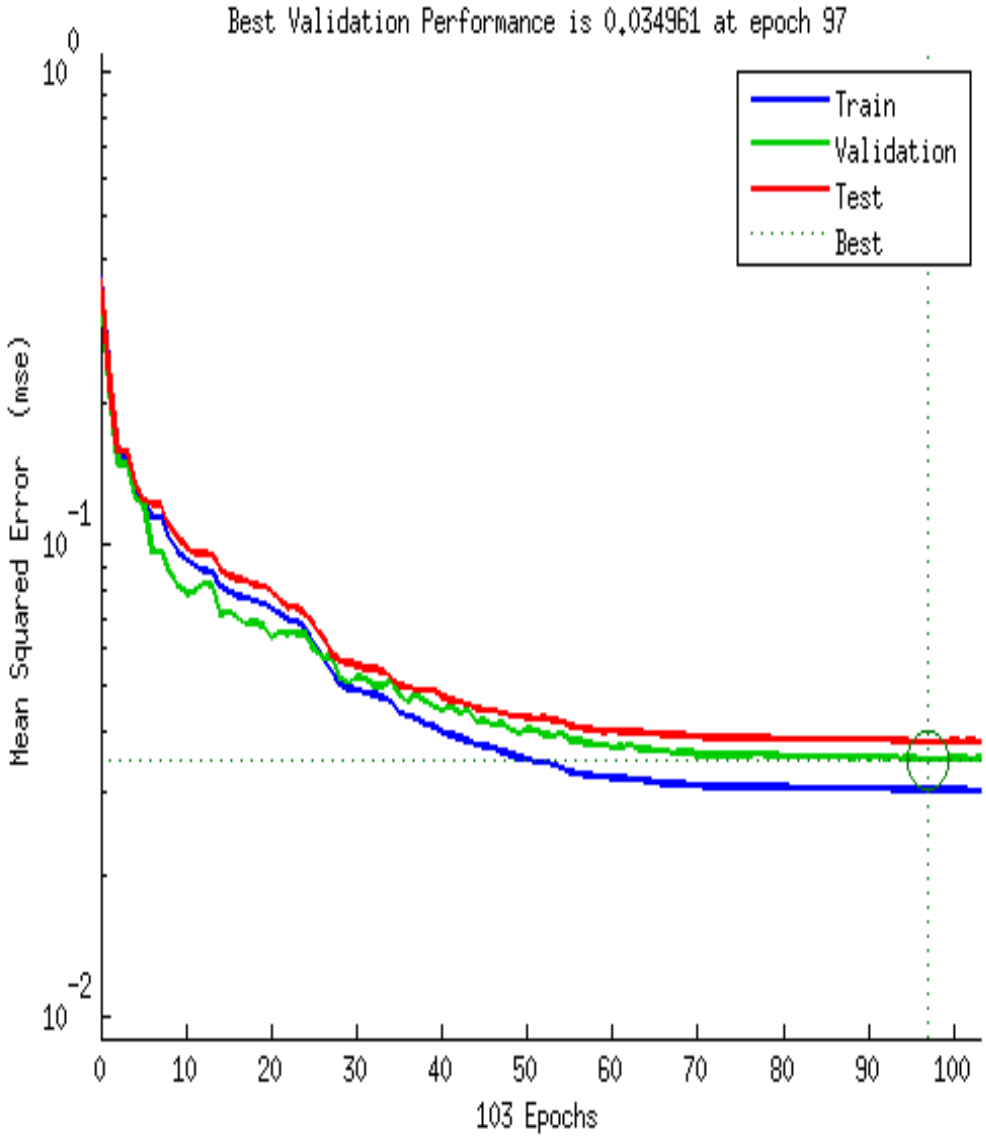


Figure 4.27: Performance plot with 20 neurons

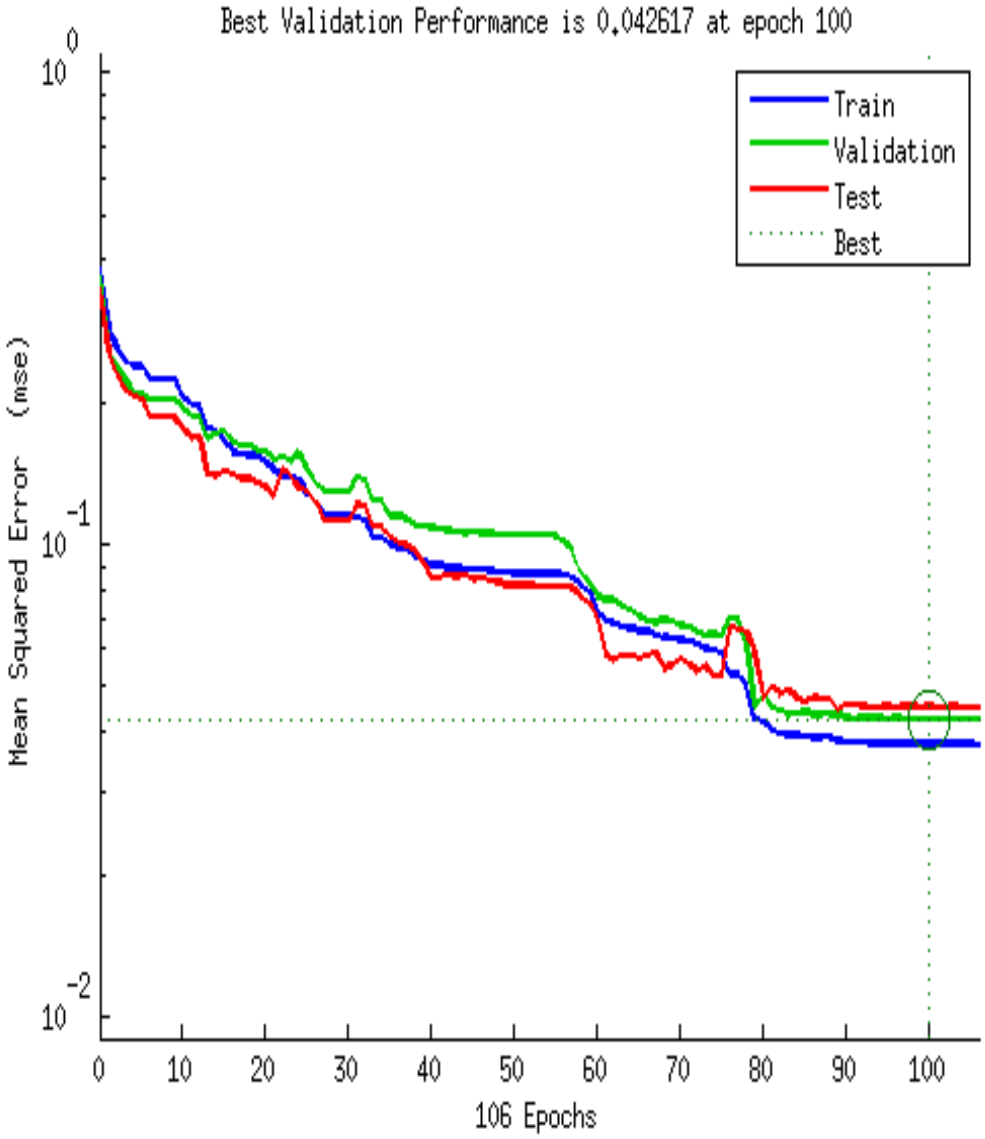


Figure 4.28: Performance plot with 40 neurons

The corresponding confusion matrix plots are shown in Figures 4.29 and 4.30

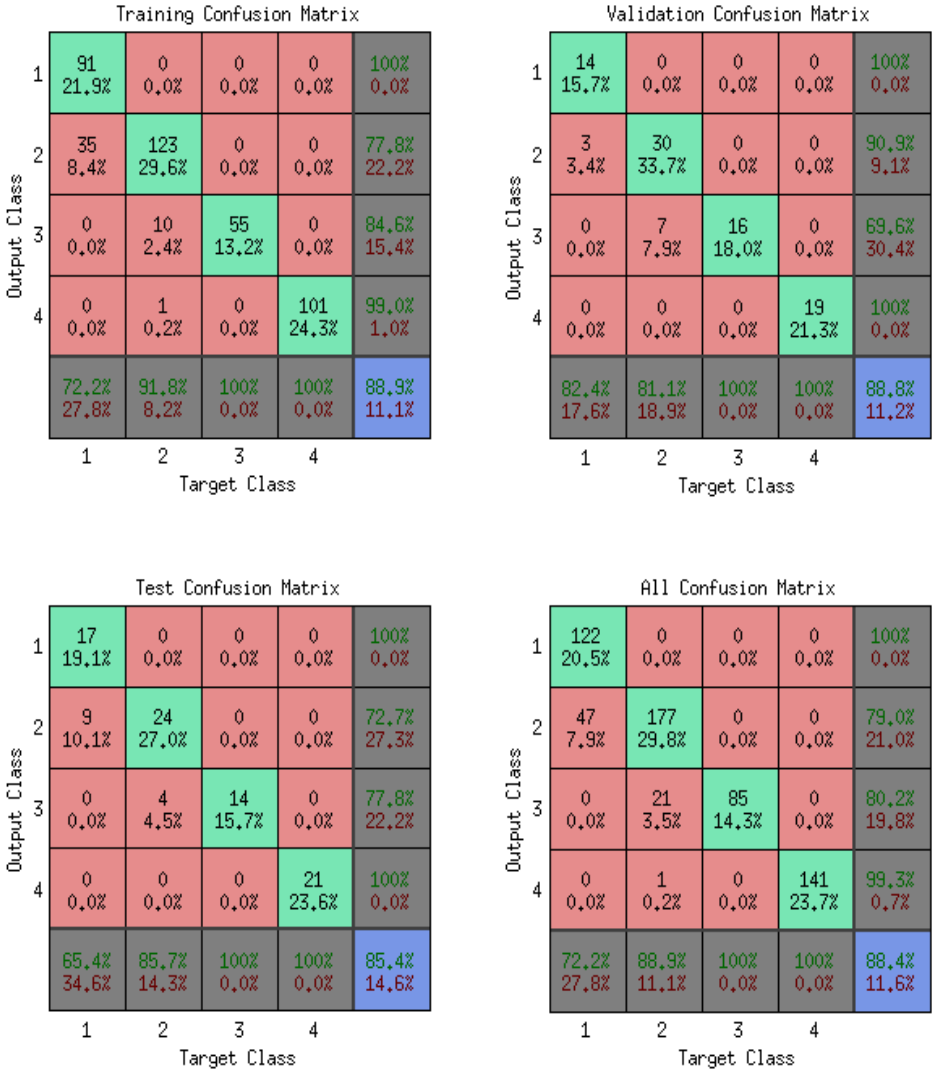


Figure 4.29: Confusion matrix with 20 neurons

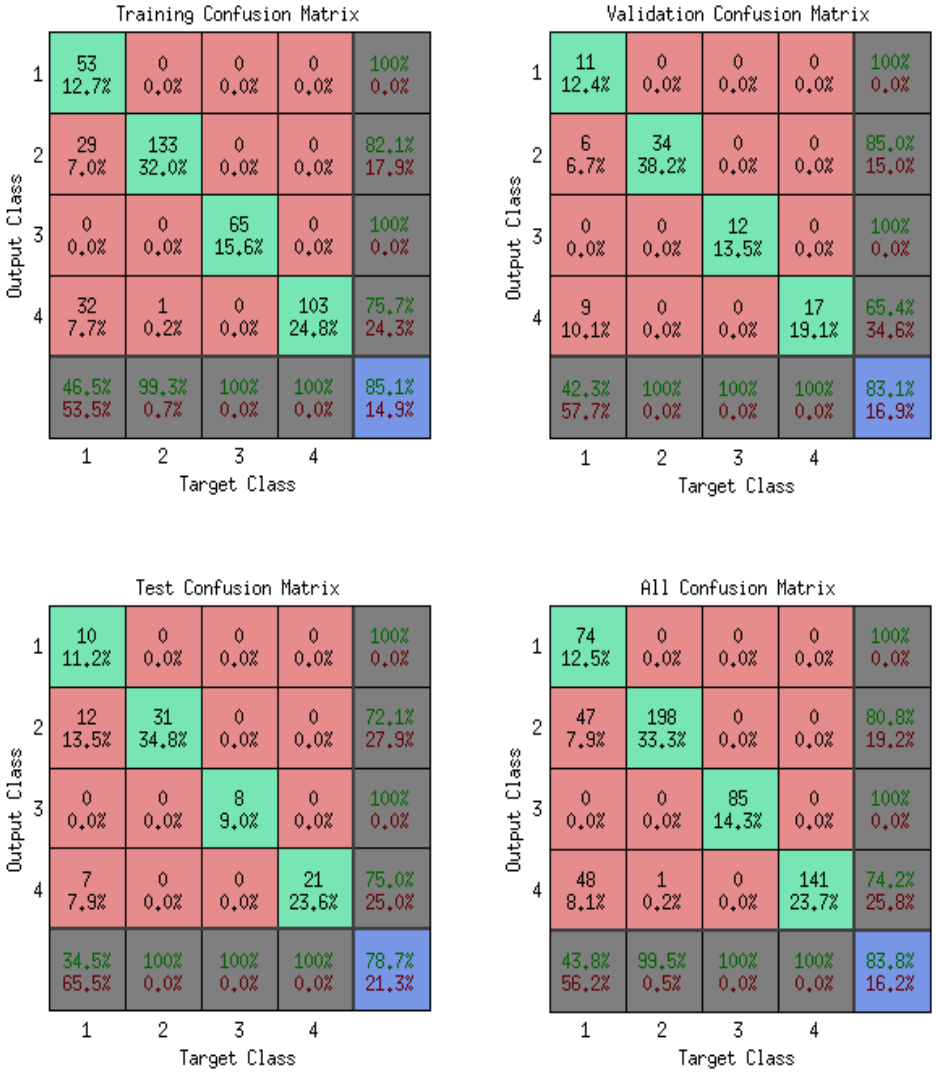


Figure 4.30: Confusion matrix with 40 neurons

The corresponding ROCs are shown in Figures 4.31 and 4.32

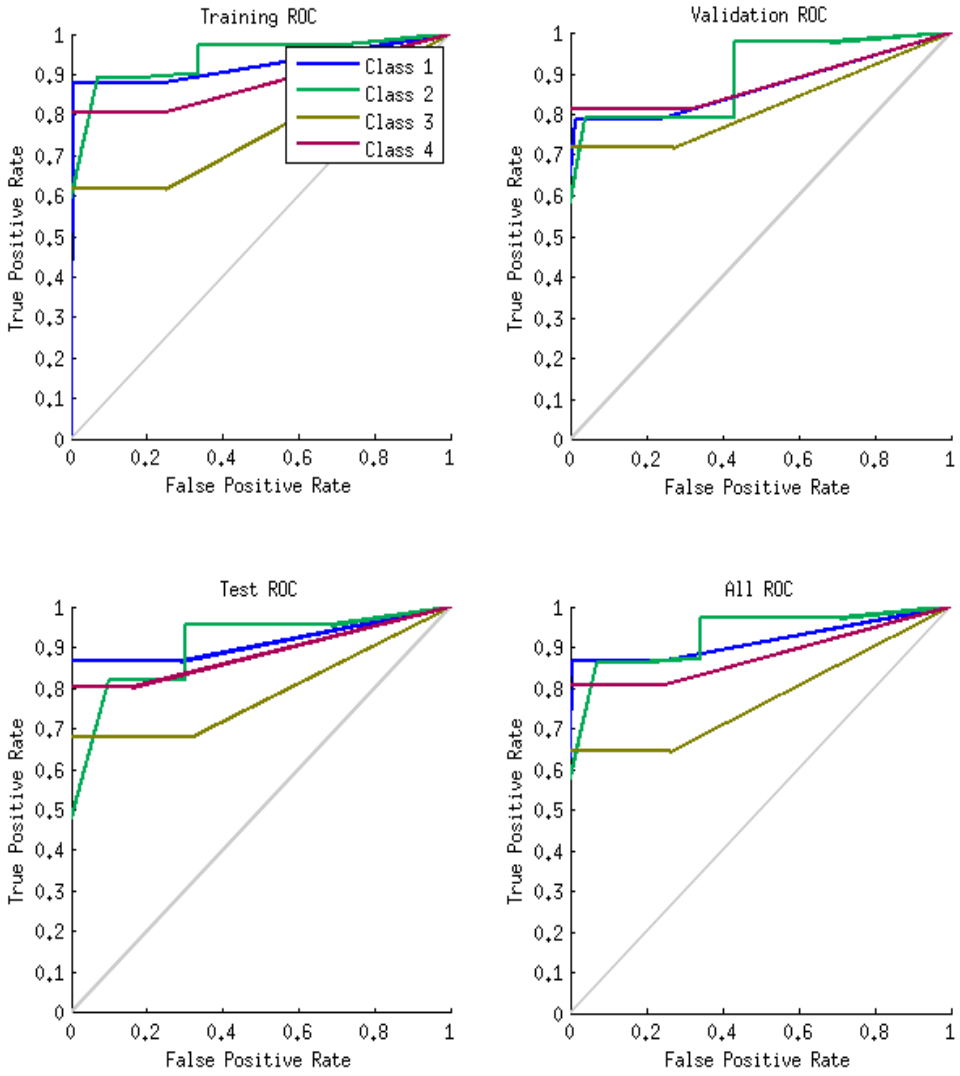


Figure 4.31: ROC with 20 neurons

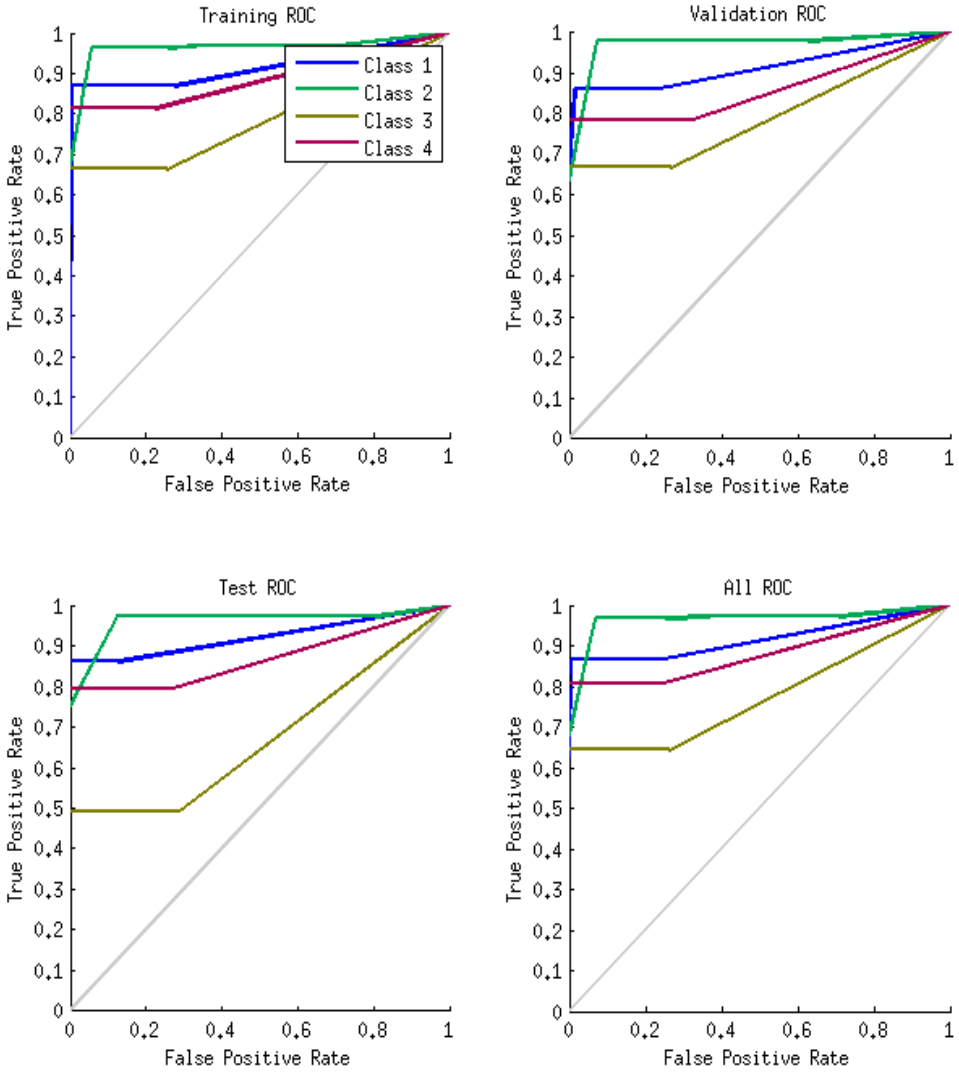


Figure 4.32: ROC with 40 neurons

The 1-2-gram performance plots for ACGT with 20 and 40 neurons in the hidden layer are given in Figures 4.33 and 4.34 respectively

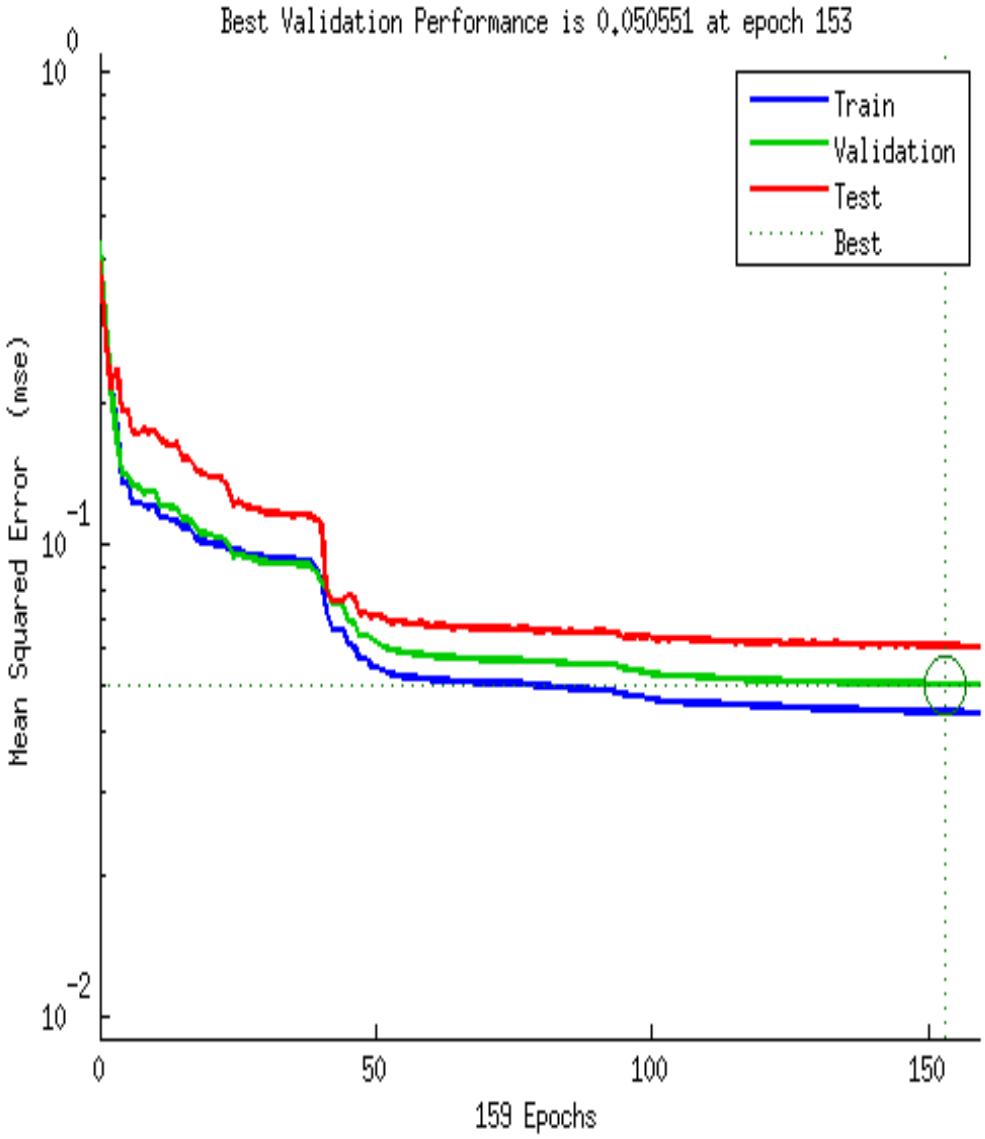


Figure 4.33: Performance plot with 20 neurons

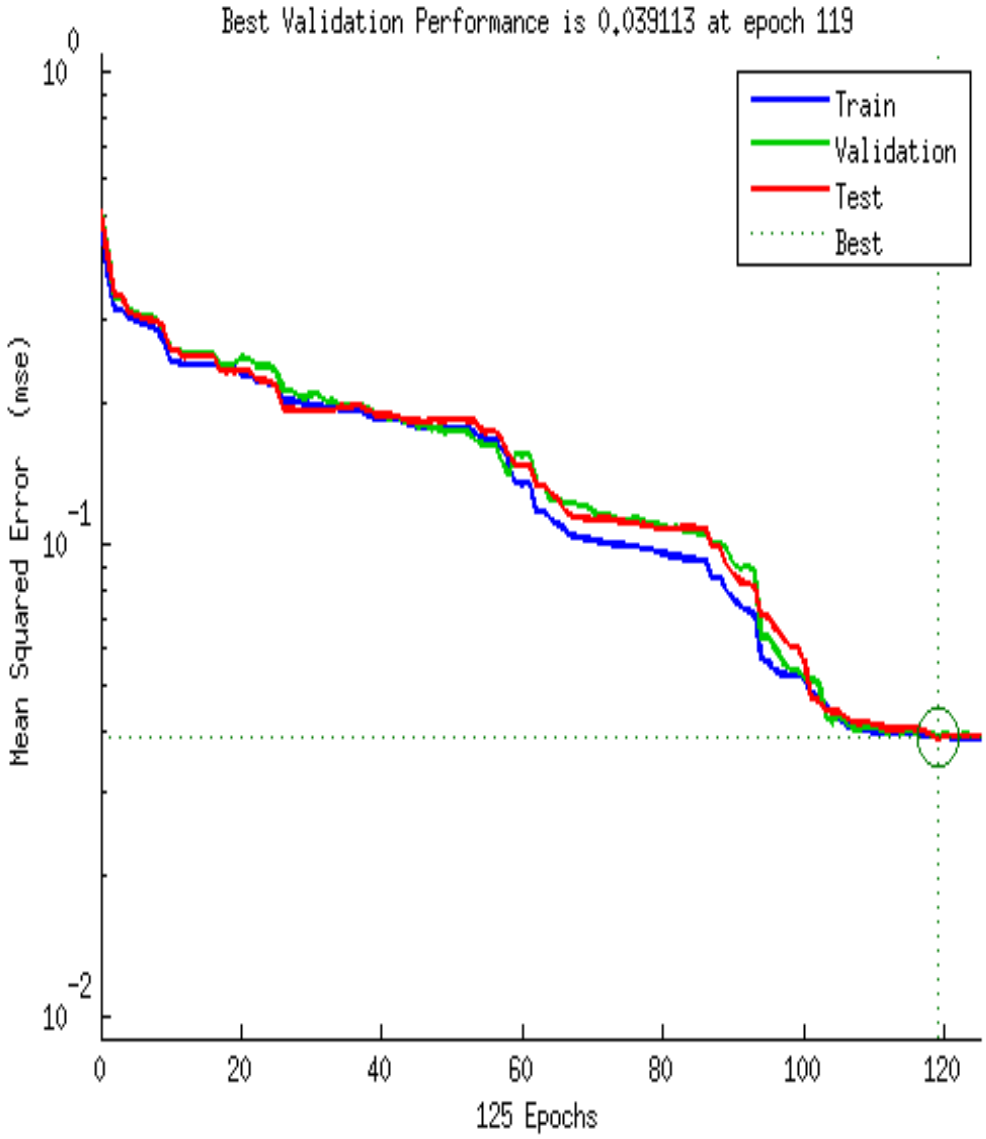


Figure 4.34: Performance plot with 40 neurons

The corresponding confusion matrix plots are shown in Figures 4.35 and 4.36

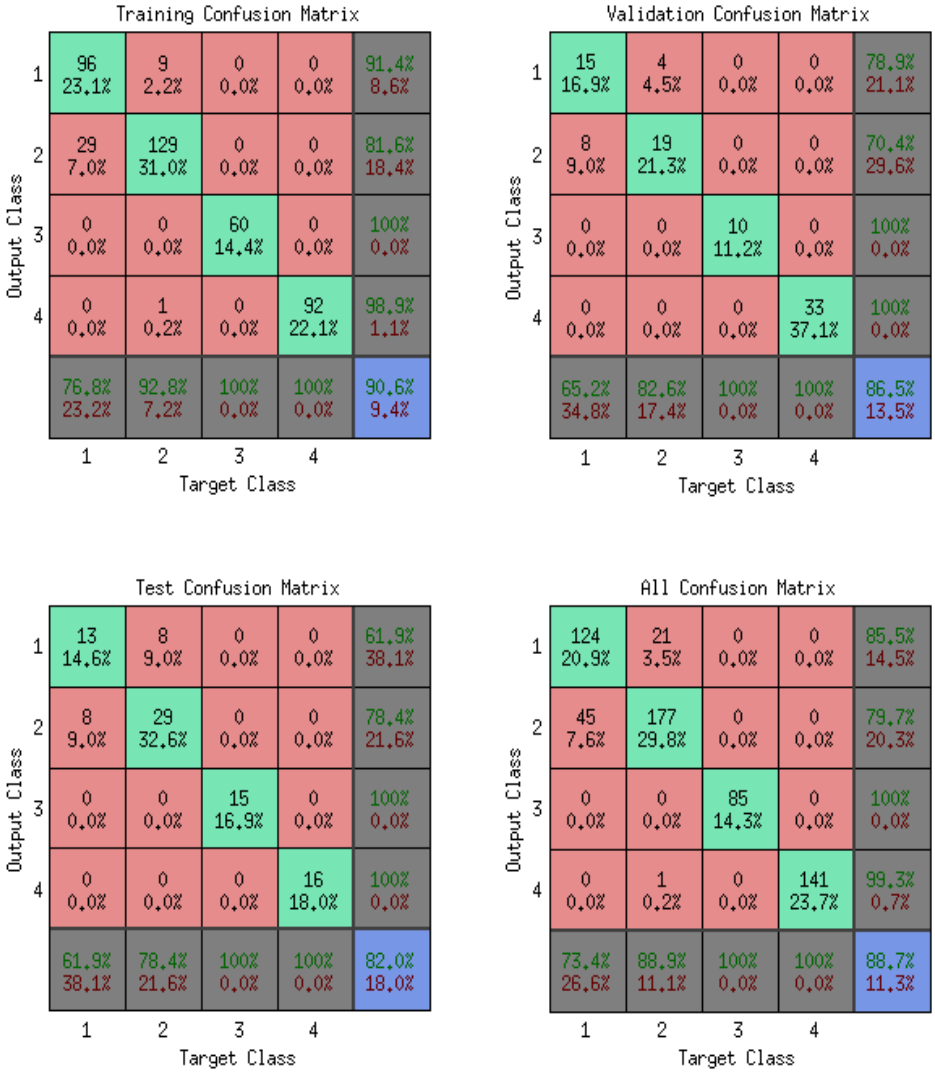


Figure 4.35: Confusion matrix with 20 neurons

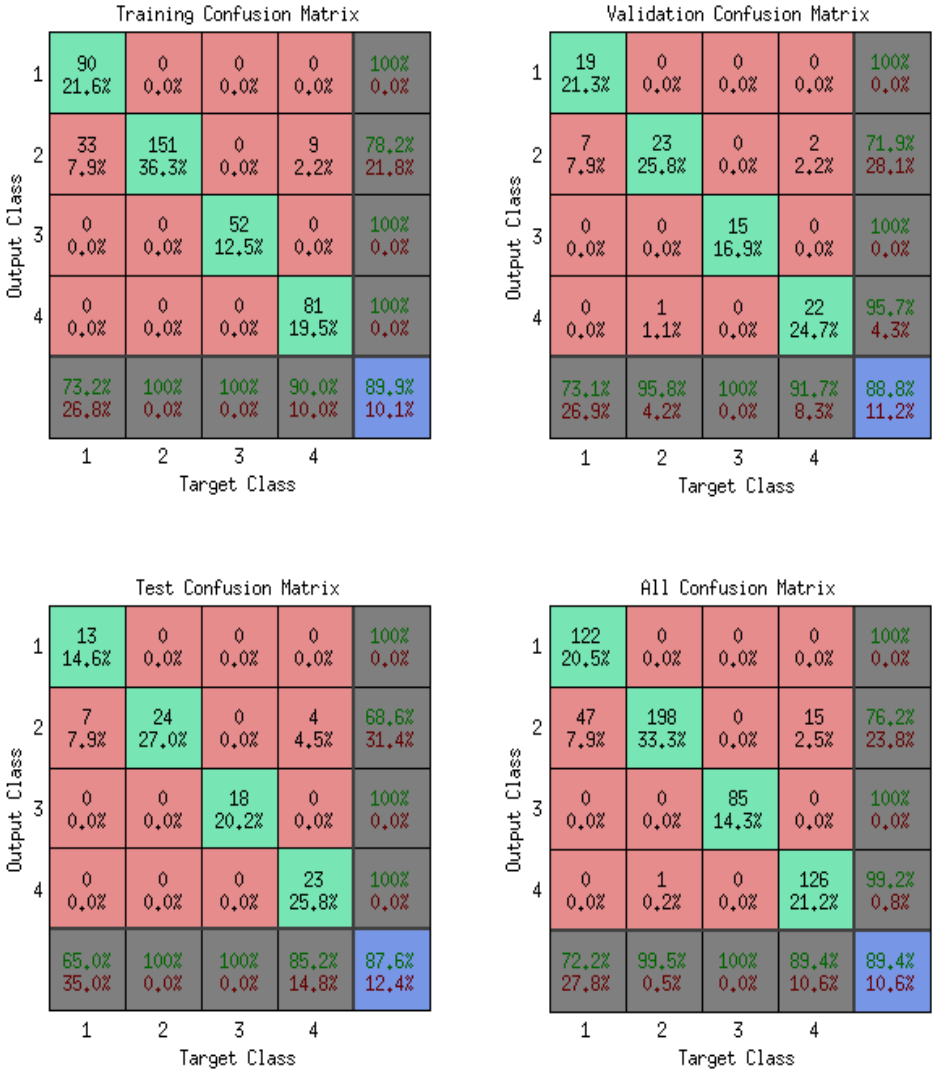


Figure 4.36: Confusion matrix with 40 neurons

The corresponding ROCs are shown in Figures 4.37 and 4.38

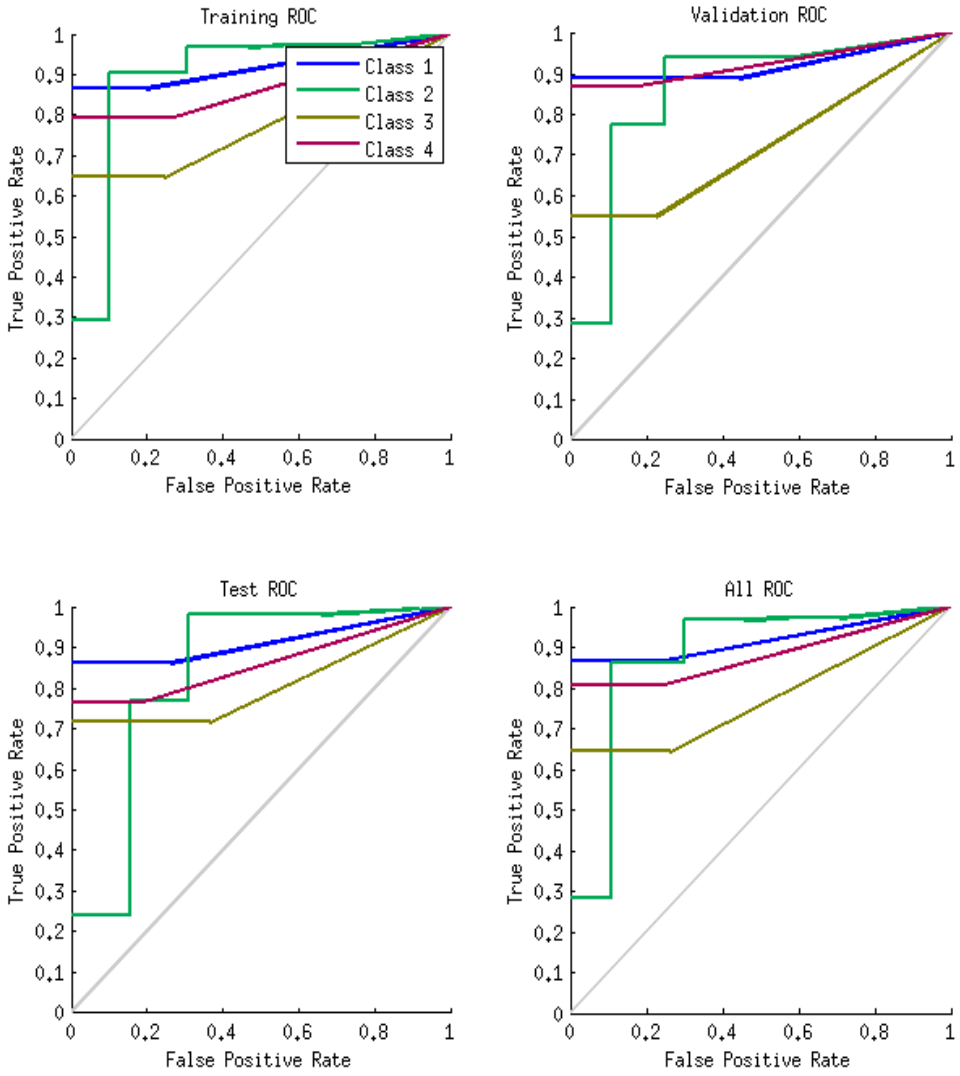


Figure 4.37: ROC with 20 neurons

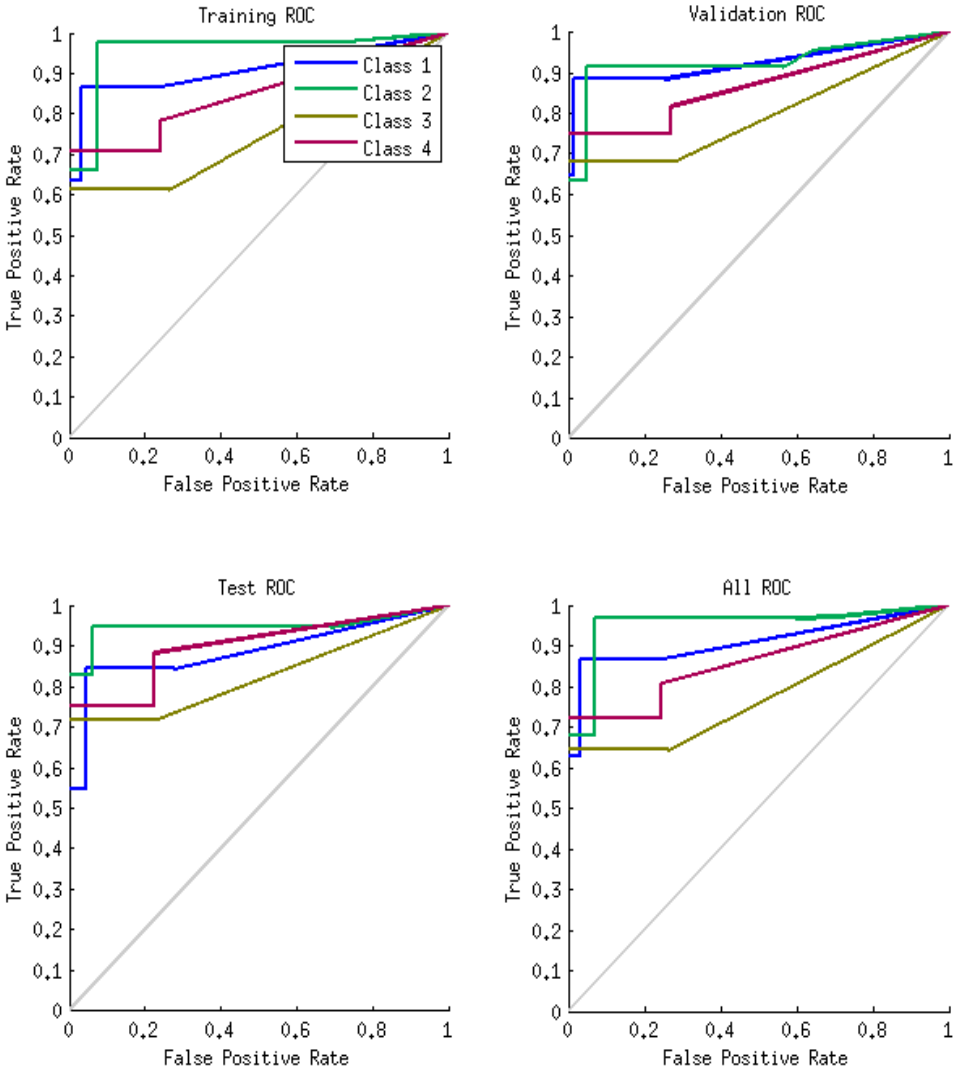


Figure 4.38: ROC with 40 neurons

Chapter 5

Data evaluation and Conclusion

Table 5.1 gives a summary of the regression values extracted from 1-gram outputs for ACGT and their averages in terms of training, validation and testing using every 26th line (row) with Matlab regression toolkit. The table shows maximum regression values (in bold) which corresponded to training set which is consistent with the expected result.

Table 5.2 gives a summary of the regression values and their averages using 2-gram. The table shows maximum regression values (in bold) for ACGT again corresponding to the training set.

No. of neurons	Best perf. values	Training	Validation	Testing
20	0.003209	0.99202	0.99054	0.97872
25	0.003173	0.99788	0.99055	0.98080
30	0.003137	0.99211	0.99070	0.97821
40	0.003195	0.99205	0.99049	0.97869
Averages	0.003178	**0.99194**	0.99057	0.97911

Table 5.1: Regression values with 1-gram

No. of neurons	Best perf. values	Training	Validation	Testing
20	0.027319	0.94319	0.91605	0.88381
25	0.025913	0.93830	0.93036	0.90905
30	0.022780	0.93570	0.93020	0.90725
40	0.022379	0.93779	0.93157	0.89763
Averages	0.024598	**0.93875**	0.92705	0.89944

Table 5.2: Regression values with 2-gram

Table 5.3 gives a summary of the regression values and their averages using 1-2-gram composition. Again, the table shows the maximum regression values (in bold) for ACGT corresponding to the training set.

No. of neurons	Best perf. values	Training	Validation	Testing
20	0.002849	0.99378	0.99148	0.98010
25	0.002666	0.99395	0.99212	0.98136
30	0.003130	0.99420	0.99078	0.98123
40	0.002525	0.99245	0.98128	0.97869
Averages	0.002793	**0.99395**	0.99171	0.98059

Table 5.3: Regression values with 1-2-gram

Using regression toolkit, I observed from Tables 5.1, 5.2 and 5.3 that the best regression values in terms of training, testing and validation were gotten when we used the 1-2-gram composition.

Table 5.4 is a summary of confusion matrix values for ACGT with 1-gram using pattern recognition toolkit. With varying number of neurons in the hidden layer ranging from 20 to 40, the top values in the rows represent percentages of the training, validation and test sets that were classified correctly while the bottom values represent the misclassified percentage.

No. of neurons	20	25	30	40
Train	100 0	100 0	100 0	100 0
Validation	100 0	100 0	100 0	100 0
Test	98.9 1.1	98.9 1.1	98.9 1.1	98.9 1.1

Table 5.4: Confusion matrices for ACGT with 1-gram

Summary of confusion matrices with 2-grams are shown in Table 5.5, with varying number of neurons in the hidden layer. The top values in the rows represent the percentages of training, validation and test sets that were classified correctly while the bottom values represent the misclassified percentage.

No. of neurons	20	25	30	40
Train	88.9	94.0	92.3	85.1
	11.1	6	7.7	14.9
Validation	88.8	87.6	92.1	83.1
	11.2	12.4	7.9	16.9
Test	85.4	86.5	89.9	79.2
	14.6	13.5	10.1	21.3

Table 5.5: Confusion matrices for ACGT with 2-gram

Table 5.6 is a summary of confusion matrices with 1-2-gram. With varying number of neurons in the hidden layer, the top values in the rows represent the percentages of training, validation and test sets that were classified correctly while the bottom values represent the misclassified percentage.

No. of neurons	20	25	30	40
Train	90.6	89.7	100	89.9
	9.4	10.3	0	10.1
Validation	86.5	89.9	100	88.9
	13.5	10.1	0	11.2
Test	82.0	80.9	98.9	87.6
	18	19.1	1.1	12.4

Table 5.6: Confusion matrices for ACGT with 1-2-gram

5.1 Analysis of the results

For the performance, a comparison of Figure 4.7 for 1-gram and Figure 4.15 for 1-2-gram showed a mean square error reduction from 0.003209 to 0.0028494 when we used 20 neurons in the hidden layer. The same thing was applicable for 1-2 gram

using 40 neurons in the hidden layer. The error reduced from 0.0031945 to 0.0025253 as shown in Figures 4.8 and 4.17. Comparison with other different number of neurons in the hidden layer showed the same trend, as the number of neurons increased.

The regression values computed also showed the same trend going from 0.98979 (98.98 percent) to 0.99138 (99.14 percent) with some outliers. These outliers in the regression plot could be from variations in the size of signal intensities recorded as some had higher values than others and hence affected the normalized values too. The confusion matrix did not perform as well as performance and regression measurements. Also, the ROC plots were not so promising when compared with the regression values obtained using the regression toolkit. A good ROC plot should have the curves in the upper left corner of the plot.

I did a 10-fold cross validation on the dataset using only the 1-2-gram of 26th row (line). Simulation to check the consistency of the results we had generated using the regression toolkit showed a performance value of 0.0164 which implies a performance accuracy of 99.36 percent. In Table 5.3 the average regression value (not shown) gotten from the 1-2-gram simulation was 0.991655 or 99.17 percent. The 10-fold cross validation even resulted in better classification.

Using the regression analysis toolkit, the test and validation errors had similar characteristics i.e. green and red lines in most of the performance plots shown in this

work. These resulted in small mean square errors (MSE's) which is a measure of performance. The training error (blue lines) were the least amongst them. These are some of the signs that our results are encouraging, although with slight offsets depending on the number of neurons in our hidden layer.

A look at Table 5.1 on page 136 shows the average regression value R in terms of training, validation and testing. The training value was 0.99194 (99.2 percent) using 1-gram. This value decreased to 0.93875 as shown in Table 5.2 on page 137 when we used the 2-gram but increased to 0.99395 when I used the 1-gram and 2-gram composition as shown in Table 5.3 on page 138.

A comparison of Tables [5.1-5.3] shows that better R-values were recorded when I used the 1-2-gram composition. This is due to the influence of the 2-gram values.

On the use of pattern recognition tookit, we got the best confusion matrix value of 99.8 percent for 1-2-gram when we used 30 neurons in the hidden layer as shown in Table 5.6. In general, I noticed in this thesis that better results were obtained when using the regression toolkit. The results of this study show that Artificial Neural Networks based n-gram model for prediction of normalized signal intensities is at least accurate based on the general nature of the curves and corresponding numerical values obtained with their attendant low mean square errors which is a measure of performance. Hence, we can use n-gram model to predict the signal intensities via

their normalized values from Affymetrix data. The result produced from this research can still be used if one wants to investigate individual nucleotide intensities along a given sequence.

This study also shows another way the signal intensity profile (measurements) of the DNA sequence as recorded by Affymetrix Genechip can be studied and analyzed. Useful prediction can then be made using this method of n-gram.

Appendix B shows the predicted normalized intensities with 20 neurons in the hidden layer using 1-grams.

Appendix C is the predicted normalized intensities with 30 neurons in the hidden layer using 2-grams.

Appendix D shows the predicted normalized intensities using 1-2-grams composition with 40 neurons in the hidden layer.

A comparison of Appendix A and Appendix D showed a 100 percent call accuracy for those nucleotides with 1 (one) as their normalized values. These are the nucleotides with the highest signal intensities across rows and have been correctly predicted using the n-gram ideas.

5.2 Future work

In this thesis, I have used mainly 1-gram and 2-gram to carry out analysis. One may improve upon these results if higher n-gram values and their different compositions are considered. An effort could also be made to get optimal number of neurons in the hidden layer that give maximal regression values and high confusion matrix values along the diagonals. An increase in regression value to say 0.999 is indicative of a much better prediction.

Other forms of normalization like Min-Max, Z-score and normalization by decimal scaling could also be explored to compare results. Another important aspect in any future research is to look into and analyze the network architecture to know which n-gram combinations give the best values thereby causing a reduction in mean square error which will in turn increase the percentage of well classified dataset in the confusion matrix. We got a maximum confusion matrix (correctly classified) value of 99.8 using the 1-2-gram composition with 30 neurons in the hidden layer. The Receiver Operating Characteristics (ROC) curves will also be improved by this. An effort can be made at improving the predictive accuracies for the individual nucleotides with 1-gram and 2-gram using the pattern recognition toolkit.

Again, we considered only one strand which could be the sense or antisense strand in

DNA parlance. Further simulations could be done using recorded intensities for both strands to know if there could be an improvement in the results that were obtained in this thesis. It will also be interesting to test our results with signal intensities from other DNA data set for consistency.

As a form of confirmation, other forms of numerical representations of DNA sequence mentioned in Chapter 2 can be tested to predict signal intensities recorded by Affymetrix Genechip to compare results. Further comparisons can be done using classifiers like Support Vector Machines and Kohenen self organizing maps to see if they offer better results than n-gram method.

Bibliography

[1] A. Abbaci et al: *DNA as building block for self-assembly of micro-components*, Quantum, Nano and Micro Technologies, Second International Conference on. pp 28-33, IEEE, *2008*.

[2] A. Bird: *C_pG islands as gene markers in the vertebrate nucleus*, Trends Genet, Vol 3, pp 342-347, *1987*

[3] A. Hatzigeorgiou and M. Megraw: *Computational Analysis of Human DNA Sequences: An Application of Artificial Neural Networks*, Nonconvex Optimization and Its Applications Vol. 85 *2006*.

[4] A. J. F. Griffins et al: *Modern Genetic Analysis: Integrating Genes and Genomes*, Second Edition, W. H. Freeman and Company, New York, *2002*.

[5] A. Jarvelin et al: *S-grams: Defining generalized n- grams for information retrieval*, Information Processing and Management, 43, 1005-1019, *2007*.

[6] A. K. Konopka and M. J. C. Crabbe: *Compact Handbook in Computational Biology*, Marcel Dekker, New York, *2004*.

[7] A. Polanski and M. Kimmel: *Bioinformatics*, Springer, Heidelberg, *1998*.

[8] A. Robertson and P. Willett: *Applications of N-grams in textual information systems*, Journal of Documentation Vol 54, No. 1, *1998*.

[9] A. S. Nair and S. P. Sreenadhan: *A coding measure scheme employing electron-ion interaction pseudopotential (EIIP)*, Bioinformation 1(6): 197-202, *2006*.

[10] A. S. Nair and T. Mahalakshmi: *Visualization of genomic data using internucleotide distance signals*, Proceedings of IEEE Genomic Signal Processing 408, *2005*.

[11] A. Tomovic et al: *N -gram based classification and unsupervised Hierarchical Clustering of Genome Sequences*, Computer methods and programs in biomedicine 81.2, 137-153, *2006*

[12] B. R. King et al: *Application of discrete Fourier intercoefficient difference for assessing genetic sequence similarity*, EURASIP Journal on Bioinformatics and Systems Biology, Vol. 1, pp 1-12, *2014*.

[13] B. Y. M. Cheng and J. G. Carbonell: *Combining n-grams and Alignment in G- Protein Coupling Specificity Prediction*, Carnegie Mellon Univesity, Research Showcase, *2007*.

[14] C. E. Shannon: *A Mathematical Theory of Communication*, Bell System Technical Journal, Vol. 27, pp 379-423, *1948*.

[15] C. Gershenson: *Artificial Neural Networks for beginners*, arXiv preprint cs/0308031 *2003*.

[16] C. H. Wu et al: *Neural Networks for Full- Scale Protein Sequence Classification: Sequence Encoding with Singular Value Decomposition*, Machine Learning, 21, 177-193, *1995*.

[17] C. Jaloth and R. Mahajan: *Analysis of n-Gram based human Promoter Recognition*, IJERA, Vol.2, Issue 6, pp 247-254 *2012*.

[18] C.K. Peng et al: *Analysis of DNA sequences using methods of statistical physics*, Physica A, Elsevier Science B.V, 249: 430-438, *1998*.

[19] C. Wu et al: *Neural Networks For Molecular Sequence Classification*, ISMB-93 Proceedings, *1993*.

[20] D. Anastassiou: *Genomic signal processing*, IEEE Signal Processing Magazine, 18: 8–20 *2001.*

[21] D. C. Anghel et al: *A matlab neural network application for the study of working condition*, In Advanced Materials Research, Vol. 837, pp. 310-315, *2014*

[22] Dennise D. Dalma-Weishausz et al: *The Affymetrix Genechip Platform: An Overview*, Methods in Enzymology, vol 410, Elsevier Inc, *2006*

[23] D. Gusfield: *Algorithms on Strings, Trees and Sequences*, Cambridge University Press, New York, *1997.*

[24] D. Polychronopoulos: *Analysis and classification of constrained DNA elements with n-gram graphs and Genomic signature*, A1CoB, LNBI 8542, pp 220-234, *2014.*

[25] E. Borrayo et al: *Genomic Signal processing methods for computation of alignment free distances from DNA sequences*, PLOS one Vol 9, issue 11, *2014.*

[26] E. G. Goodaire and M. M. Parmenter: *Discrete Mathematics with Graph Theory*, Third Edition, Prentice Hall, New Jersey, *2006.*

[27] E. Hamori and J. Ruskin: *H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences.* J. Biol. Chem. 258(2):1318-27. *1983.*

[28] E. Southern: *United Kingdom patent application*, GB8810400, *1988*.

[29] E. Ukkonen: *Approximate string-matching with q-grams and maximal matches*, Theoretical Computer Science, 191-211, *1992*.

[30] F. Kokkoras and K. Paraskevopoulos: *Artifial Neural Networks*, www.ihu.edu.gr, *2015*

[31] G. Giannakopoulos et al: *Summarization system evaluation revisited: N-gram graphs*, ACM Trans. Speech Lang process, 5(3), 139, *2008*.

[32] G. Kondrak: *N-Gram Similarity and Distance*, LNCS 3772, pp. 115-126, *2005*.

[33] G. K. Zipf: *Selective studies and the principle of relative frequency in language*, Cambridge MA, Harvard University Press, *1932*.

[34] G. Liu and Y. Luan: *Identification of Protein Coding Regions in the Eukaryotic DNA Sequences based on Maple Algorithm and Wavelet Packets Transform*, Abstract and Applied Analysis, Vol 2014, Hindawi Publishing Corporation, *2014*.

[35] H. Herzel et al: *Interpreting correlations in biosequences*, Physica A 249, 449-459 *1998*.

[36] H. Herzel and I. GroBe: *Measuring correlations in symbol sequences*, Physica A 216 pp 518-542, *1995*.

[37] H. K. Kwan and S. B. Arniker: *Numerical representation of DNA sequences* Electro/Information Technology. IEEE International Conference, Windsor, 307-310 *2009*.

[38] H. K. Kwan et al: *Novel methodologies for spectral classification of exon and intron sequences*, Eurasip Journal on Advances in Signal Processing, 2012.1: pp 1-14. *2012*.

[39] H. Saberkari et al: *A fast algorithm for Exonic Regions Prediction in DNA Sequences*, Journal of Medical Signals and Sensors, Vol. 3, Issue 3, *2013*.

[40] H. T. Chang: *DNA sequence visualization* Advanced Data Mining Technologies in Bioinformatics, Idea Group Publishing 4: 63-84, *2006*

[41] *http://en.wikipedia.org/wiki/John Henry Holland*, June 29, 2015.

[42] *http://www.affymetrix.com*, June, 30 *2015*.

[43] *http://www.uvm.edu/ cgep/Education/Sequence.html*, October, 30 *2015*

[44] H. U. Osmanbeyoglu and M. K. Ganapathiraju: *N- gram analysis of 970 microbial organisms reveals presence of biological language models*, BMC Bioinformatics 12:12, *2011*.

[45] H. Urakawa et al: *Single-Base-Pair discrimination of Terminal Mismatches by using oligonucleotide microarrays and neural network analysis*, Applied and Environmental Biology, Vol. 68, No. 1, pp. 235-244, *2002*.

[46] I. H. Witten et al: *Data Mining: Practical Machine Learning Tools and Techniques*, Third Edition, Elsevier, MA, USA, *2011*.

[47] J. Bao et al: *An improved alignment-free model for DNA sequence similarity metric*, BMC Bioinformatics 15:321, *2014*.

[48] J. Han and M. Kamber: *Data Mining, Concepts and Techniques*, Second Edition, Morgan Kauffman Publishers, New York, *2006*.

[49] J. K. Vries et al: *The relationship between n-gram patterns and protein secondary structure*, Proteins, Wiley Interscience, *2007*.

[50] J. Pandey and K. Tripathi: *On the investigation of Biological Phenomena through Computational Intelligence*, Global Journal of Computer Science and Technology, Vol. 14, Issue 1 Version 1.0, *2004*.

[51] J. R. Koza: *On the programming of computers by means of natural selection*, MIT, *1992*.

[52] J. Tang: *Lecture Notes on Data Mining CS 6762*, Memorial University, Newfoundland Canada, Fall Semester, *2014*.

[53] L. K. Buehler and H. H. Rashidi: *Bioinformatics Basics, Applications in Biological Science and Medicine*, Second Edition, Taylor and Francis, New York, *2005*.

[54] L. Pray: *Discovery of DNA structure and function: Watson and Crick*, Nature Education, 1(1): 100, *2008*.

[55] M. Abo-Zahhad et al: *A Novel Circular Mapping Technique for Spectral Classification of Exons and Introns in Human DNA Sequences*: International Journal of Information Technology and Computer Science (IJITCS) 6, no. 4 :19, *2014*.

[56] M. Abo-Zahhad et al: *Genomic Analysis and Classification of Exon and Intron Sequences Using DNA Numerical Mapping Techniques*, I.J. Information Technology and Computer Science, Vol.8, 22-36, *2012*.

[57] M. Abo-Zahhad et al: *Integrated model of DNA sequence numerical representation and artificial neural network for human donor and acceptor sites prediction*, Int. Journal of Inf. Tech and Computer Science, 51-57, *2014*.

[58] M. Akhtar et al: *Signal Processing in Sequence Analysis: Advances in Eukaryotic Gene Prediction*, IEEE Vol. 2, no. 3, *2008*.

[59] M. Akhtar et al: *On DNA numerical representations for period-3 based exon prediction* In Proc. of IEEE Workshop on Genomic Signal Processing and Statistics (GENSIPS), 1-4, *2007*.

[60] M. Ganapathiraju et al: *BLMT, Statistical Sequence Analysis using N-grams*, Appl. Bioinformatics, 3,: 193-200, *2004*.

[61] M. Gardiner-Garden and M. Frommer: *C_pG islands in vertebrate genomes* J. Mol Bio, Vol 196 pp 261-282, *1987*.

[62] M. H. Al Shamisi et al: *Using Matlab to develop Artificial Neural Network models for predicting global solar radiation in Al Ain City- UAE*, INTECH Open Access Publisher, *2011*.

[63] M. Lutz: *Programming Python*, O' Reilly, Third Edition, Beijing, *2006*

[64] M. L. Model: *Bioinformatics programming using Python*, O'Reilly, Beijing, *2009*

[65] M. Masso: *Sequence-Based Prediction of HIV-1 Coreceptor Usage: Utility of n-grams for representing gp120 V3 Loops*, ACM, 978-1-4503-0796, *2011*.

[66] M. M. Hapudeniya: *Artificial Neural Networks in Bioinformatics*, Sri Lanka Journal of Bio-Medical Informatics, I(2): 104-111, *2010*.

[67] M. Pop et al: *Genome Sequence Assembly: Algorithms and Issues*, Computer 35, no(7), pp. 47-54.*2002*.

[68] M. S. Waterman: *Introduction to Computational Biology: Maps, Sequences and Genomes*, Chapman and Hall, New York, *1995*.

[69] M. Yan et al: *A new Fourier transform approach for protein coding measure based on the format of Z curve*, Bioinformatics, 14: 685–690, *1998*.

[70] N. A. Campbell et al: *Biology*, Fifth Edition, Addison Wessley Longman, Inc, New York, *1999*.

[71] N. Chakravarthy et al: *Autoregressive modeling and feature analysis of DNA sequences*, EURASIP Journal of Genomic Signal Processing, 1: 13-28, *2004*.

[72] N. Mahamad et al: *Power prediction analysis using Artificial Neural Network in MS Excel*, Latest trends in renewable energy and environmental informatics, ISBN 978-1-61804-175-3, *2013*.

[73] N. S. Mitic et al: *Could n-gram analysis contribute to genomic island determination?*, Journal of Biomedical Informatics, 41, 938-943, *2008*.

[74] N. Shalaby: *Lecture notes on Graph Theory*, Memorial University of Newfoundland, Canada, Winter Semester, *2012*.

[75] O. Weiss and H. Herzel: *Correlations in Protein Sequences and Property Codes*, J. Theor. Biol., 190, 341-353, *1998*.

[76] P. A. Pevzner: *Computational Molecular Biology, An Algorithmic Approach*, MIT Press, Massachusetts, *2000*.

[77] P. Aloy et al: *TransMem: A neural network implemented in excel spreadsheets for predicting transmembrane domains of proteins*, CABIOS Vol 13 no. 3, *1997*.

[78] P. Baldi and S. Brunak: *Bioinformatics: The Machine Learning Approach*, A Bradford Book, MIT, *2001*.

[79] P. Bernaola- Galvan et al: *Study of statistical correlations in DNA sequences*, Gene 300, 105-115, *2002*.

[80] P. Cristea et al: *Prediction of Nucleotide Seqences by using genomic signals*, 8th WSEAS Int. Conf. on Neural Networks, Sofia Bulgaria, *2008*.

[81] P. Cristea et al: *Application of neural networks, PCA and feature extraction for prediction of nucleotide sequences by using genomic signals*, 9th Symposium on Neural Network Applications in Electrical Engineering, NEUREL, *2008*.

[82] P. D. Cristea: *Conversion of nucleotides sequences into genomic signals*. J. Cell. Mol. Med., April- 6:279-303, *2002*.

[83] P. D. Cristea: *Genetic signal representation and analysis* In Proc. SPIE Inf. Conf. Biomedical Optics, 77–84, *2002*.

[84] P. J. Russell: *iGenetics: A Molecular Approach*, Third Edition, Benjamin Cummings, New York, *2010*.

[85] P. Kumar et al: *Using Subsequence Information with KNN for Classification of Sequential Data*, ICDCIT, LNCS, pp 536-546, *2005*.

[86] Q. G. Sural et al: *Similarity between Euclidean and cosine angle distance for nearest neighbour queries*, SAC 1232-1237, *2007*.

[87] R. Drmanac, et al: *Sequencing of megabase plus DNA by hybridization: theory of the method*, Genomics, 4: 114-128, *1989*.

[88] R. F. Voss: *Evolution of long-range fractal correlations and 1/f noise in DNA base sequences* Phys. Rev. Lett. 68: 3805–3808, *1992*.

[89] R. I. Mubark et al: *Different Species Classifier based on Haemoglobin Sequences*, Biomed Proceedings 21, pp 279-281, *2008*.

[90] R. Jiang and H. Han: *Segmentation of short human exons based on spectral features of double curves*, International Journal of Data Mining and Bioinformatics 2 (1),15-35 *2008*.

[91] R. Jiang and H. Han: *Studies of Spectral properties of short genes using the wavelet subspace Hilbert-Huang transform(WSHHT)*, Physica A, 387, 4223-4247, *2008*.

[92] R. K. Jena et al: *Soft Computing Methodologies in Bioinformatics*, European Journal of Scientific Research, Vol. 26, No.2, *2009*.

[93] R. M. Indury and M. S. Waterman: *A new algorithm for DNA sequence analysis*, Journal of Computational Biology, Vol. 2, no. 2, pp 291-306, *1995*.

[94] R. S. Pujari et al: *Intrusion Detection using Text Processing Techniques with Binary- Weighted Cosine Metric*, Int. Journal. of Information Security, Springer - Verlag, *2004*

[95] R. Zhang and C. T. Zhang: *Z curves. An Intuitive Tool, for Visualizing and Analyzing the DNA sequences*, J. BioMol. Struct. Dyn. 11: 767-782, *1994*.

[96] S. Aluru: *Handbook of Computational Molecular Biology*, Chapman and Hall, CRC, New York, *2006*.

[97] Sams String Metrics: http://www.dcs.shef.sc.uk/ sam/stringmetrics.html.

[98] S. Datta and A. Asif: *A Fast DFT Based Gene Prediction Algorithm for Identification of Protein Coding Regions*, ICASSP (5), pp 653-656, *2005*.

[99] S. Karlin and L. R. Cardon: *Computational DNA Sequence Analysis*, Annu. Rev. Microbiol. Vol 48: pp 619-54, *1994*.

[100] S. Logeswaran et al: *Computational identification of short initial exons*: Pattern Recognition in Bioinformatics, LNBI vol.4146, pp. 42-48, *2006*.

[101] S. M. Carr et al : *Phylogeographic genomics of mitochondrial DNA: Highly resolved patterns of intraspecific evolution and a multi-species, microarray-based DNA sequencing strategy for biodiversity studies*, Comparative Biochemistry and Physiology, Part D 3 1-11, *2008*

[102] S. M. C. Flynn: *Species specificity of DNA re-sequencing microarrays*: B.Sc Project, Department of Biology, MUN, NL, Canada, *2006*.

[103] S. M. C. Flynn and S. M. Carr: *Interspecies hybridization on DNA resequencing microarrays: efficiency of sequence recovery and accuracy of SNP detection in human, ape, and codfish mitochonrial DNA genomes sequenced on a human-specific Mitochip*, BMC Genomics, 8:339, *2007*.

[104] S. Mitra et al: *Introduction to Machine Learning and Bioinformatics*, CRC Press, London, *2008*.

[105] S. R. Maetschke et al: *A visual framework for sequence analysis using n-grams and spectral re-arrangement*, Bioinformatics, Vol. 26, No 6, pp 737-744, *2010*.

[106] T. Brown: *Introduction to Genetics, A Molecular Approach*, Garland Science, New York, *2012*

[107] The Mathworks: *www.mathworks.com*, June 20, 2015.

[108] T. M. Mitchell: *Machine Learning*, WCB McGraw-Hill, Boston, *1997*.

[109] T. S. Rani and R. S. Bapi: *Analysis of n-Gram based Promoter Recognition Methods and Application to Whole Genome Promoter Prediction*, In Silico Biology 9, S1-S16, *2009*.

[110] T. W. Fox and A. Carreira: *A Digital Signal Processing Method for Gene Prediction with Improved Noise Suppression*, EURASIP Journal on Applied Signal Processing, Vol. 1, 108-114, *2004*.

[111] U. Maulik et al: *Computer Intelligence and Pattern Analysis in Biological Informatics*, Wiley Series on Bioinformatics, *2010*.

[112] V. Bevilacqua et al: *Genetic Algorithms and Artificial Neural Networks in Microarray Data Analysis: A Distributed Approach*, Engineering Letters, 13:3, *2014*.

[113] W. Bains and G. Smith: *A novel method for nucleic acid sequence determination*, Journal of Theoretical Biology, 135: 303-307, *1988*.

[114] W. Banzhaf et al: *Genetic Programming: An Introduction, On the Automatic Evolution of Computer Programs and its Applications*, Morgan Kaufmann Publishers Inc, California, *1998*.

[115] W. F. Zhang and H. Yan: *Exon Prediction using empirical mode decomposition and Fourier transform of structural profiles of DNA sequences*, Pattern Recognition, 45, 947-955, *2012*.

[116] W. Li: *The study of correlation structures of DNA sequences: A critical review*, Computers Chem Vol. 21. No. 4, pp 257-271, *1997*.

[117] W. S. Klug and M. R. Cummings: *Essentials of Genetics*, Fifth Edition, Prentice Hall, New Jersey, *2005*.

[118] X. Cao et al: *Indexing DNA Sequences Using q-grams*, Database Systems for Advanced Applications. Springer Berlin Heidelberg, *2005*.

[119] X. Wang et al: *Efficient and Effective KNN Sequence Search with approximate n-grams*, Proceedings of the VLDB Endowment, Vol. 7, No 1, *2013*.

[120] X. Wu and V. Kumar: *The Top Ten Algorithms in Data mining*, CRC Press, London, *2009*

[121] X. Xia: *The effect of probe length and GC percent on Microarray Signal Intensity: Characterizing the Functional Relationship*, Int. Jour. of Systems and Synthetic Biology 1(2) pp 171-183, *2010*.

[122] Y. C. Huang et al: *Using n-gram analysis to cluster heartbeat signals*, Medical Informatics and Decision Making, 12: 64, *2012*.

[123] Y. Dang et al: *A KNN-based Learning Method for Biology Species Categorization*, LNCS 3610, pp. 956-964, *2005*.

[124] Y. Le Cun et al: *Optimal brain damage*; In advances in Neural Information Processing Systems 2. San Mateo, CA. *1990*.

[125] Y. Lysov et al: *DNA sequencing by hybridization with oligonucleotides*, doklady Academic Nauk USSR, 303: 1508-1511, *1988*.

[126] Y. Zhang and W. Chen: *A New Measure for Similarity Searching in DNA Sequences*, Match Commun, Math Comput. Chem., 65 pp 477- 488, *2011*.

[127] Z. S. Kovacheva: *Application of Neural Networks to data mining*, SQU Journal for Sciences, 12(2), 121-141, *2007*.

[128] Z. Volkovich et al: *The method of N- grams in large-scale clustering of DNA texts*, Pattern Recognition, 38, 1902-1912, *2005*.

[129] Z. Wang et al: *A Brief Review of Computational Gene Prediction Methods*, Geno. Prot. Bioinfo. Vol. 2 No. 4, *2004*.

Appendix A

1-gram and 2-gram values with normalized Affymetrix intensity measurements of the respective nucleotides for every 26th line.

CRS ref	Position	CRS ref (26)	1-gram values	2-gram values	A	C	G	T
c	26	a	0.310	0.098	1.000	0.000	0.103	0.125
a	52	c	0.311	0.026	0.040	1.000	0.000	0.025
c	78	t	0.248	0.084	0.061	0.000	0.192	1.000
a	104	g	0.131	0.043	0.047	0.015	1.000	0.000
g	130	c	0.311	0.091	0.060	1.000	0.000	0.073
t	156	c	0.311	0.089	0.262	1.000	0.000	0.043
t	182	a	0.310	0.049	1.000	0.000	0.170	0.069
t	208	c	0.311	0.091	0.244	1.000	0.000	0.193
a	234	c	0.311	0.091	0.058	1.000	0.000	0.119
t	260	t	0.248	0.084	0.009	0.000	0.137	1.000
g	286	c	0.311	0.089	0.105	1.000	0.094	0.000
t	312	t	0.248	0.061	0.072	0.000	0.104	1.000
a	338	a	0.310	0.074	1.000	0.000	0.118	0.275
g	364	g	0.131	0.025	0.074	0.030	1.000	0.000
c	390	c	0.311	0.105	0.119	1.000	0.000	0.119
t	416	g	0.131	0.025	0.016	0.000	1.000	0.051
t	442	a	0.310	0.098	1.000	0.000	0.088	0.094
a	468	t	0.248	0.084	0.105	0.000	0.241	1.000
c	494	c	0.311	0.091	0.095	1.000	0.000	0.021
c	520	g	0.131	0.043	0.000	0.000	1.000	0.035
t	546	a	0.310	0.074	1.000	0.000	0.135	0.108
c	572	t	0.248	0.084	0.080	0.000	0.209	1.000
c	598	g	0.131	0.038	0.065	0.000	1.000	0.043
t	624	a	0.310	0.049	1.000	0.000	0.157	0.343
c	650	a	0.310	0.089	1.000	0.140	0.126	0.000
a	676	c	0.311	0.089	0.128	1.000	0.004	0.000
a	702	c	0.311	0.091	0.020	1.000	0.000	0.044
a	728	a	0.310	0.049	1.000	0.000	0.109	0.179
g	754	a	0.310	0.049	1.000	0.000	0.105	0.266
c	780	c	0.311	0.091	0.083	1.000	0.000	0.061

a	806	c	0.311	0.091	0.050	1.000	0.000	0.001
a	832	a	0.310	0.089	1.000	0.000	0.072	0.025
t	858	a	0.310	0.049	1.000	0.000	0.056	0.190
a	884	t	0.248	0.030	0.001	0.000	0.083	1.000
c	910	a	0.310	0.089	1.000	0.148	0.000	0.222
a	936	c	0.311	0.089	0.043	1.000	0.043	0.000
c	962	c	0.311	0.089	0.021	1.000	0.001	0.000
t	988	t	0.248	0.073	0.000	0.013	0.238	1.000
g	1014	g	0.131	0.025	0.161	0.000	1.000	0.091
a	1040	c	0.311	0.089	0.162	1.000	0.019	0.000
a	1066	t	0.248	0.061	0.010	0.000	0.081	1.000
a	1092	g	0.131	0.043	0.094	0.054	1.000	0.000
a	1118	c	0.311	0.105	0.034	1.000	0.053	0.000
t	1144	t	0.248	0.061	0.000	0.052	0.157	1.000
g	1170	g	0.131	0.025	0.000	0.006	1.000	0.008
t	1196	g	0.131	0.043	0.000	0.041	1.000	0.008
t	1222	a	0.310	0.098	1.000	0.000	0.126	0.175
t	1248	a	0.310	0.098	1.000	0.000	0.127	0.098
a	1274	c	0.311	0.089	0.000	1.000	0.014	0.009
g	1300	t	0.248	0.084	0.041	0.000	0.148	1.000
a	1326	a	0.310	0.049	1.000	0.000	0.194	0.328
c	1352	t	0.248	0.084	0.000	0.157	0.191	1.000
g	1378	t	0.248	0.073	0.013	0.000	0.410	1.000
g	1404	t	0.248	0.061	0.014	0.000	0.044	1.000
g	1430	g	0.131	0.038	0.056	0.015	1.000	0.000
c	1456	g	0.131	0.038	0.017	0.000	1.000	0.023
t	1482	t	0.248	0.061	0.015	0.000	0.033	1.000
c	1508	t	0.248	0.061	0.000	0.161	0.131	1.000
a	1534	a	0.310	0.049	1.000	0.000	0.019	0.150
c	1560	a	0.310	0.098	1.000	0.000	0.049	0.068
a	1586	t	0.248	0.061	0.008	0.000	0.068	1.000
t	1612	a	0.310	0.049	1.000	0.000	0.143	0.371
c	1638	c	0.311	0.089	0.230	1.000	0.174	0.000
a	1664	c	0.311	0.091	0.004	1.000	0.000	0.016
c	1690	c	0.311	0.105	0.116	1.000	0.000	0.008

c	**1716**	**a**	0.310	0.049	1.000	0.000	0.129	0.195
c	**1742**	**c**	0.311	0.091	0.000	1.000	0.015	0.031
c	**1768**	**c**	0.311	0.105	0.111	1.000	0.059	0.000
a	**1794**	**g**	0.131	0.038	0.022	0.000	1.000	0.102
t	**1820**	**t**	0.248	0.073	0.000	0.238	0.326	1.000
a	**1846**	**a**	0.310	0.089	1.000	0.000	0.284	0.093
a	**1872**	**a**	0.310	0.074	1.000	0.000	0.182	0.218
a	**1898**	**g**	0.131	0.025	0.065	0.000	1.000	0.047
c	**1924**	**g**	0.131	0.025	0.000	0.037	1.000	0.167
a	**1950**	**t**	0.248	0.073	0.000	0.016	0.190	1.000
a	**1976**	**c**	0.311	0.105	0.316	1.000	0.000	0.194
a	**2002**	**g**	0.131	0.025	0.047	0.006	1.000	0.000
t	**2028**	**a**	0.310	0.098	1.000	0.000	0.098	0.128
a	**2054**	**t**	0.248	0.030	0.000	0.022	0.122	1.000
g	**2080**	**g**	0.131	0.043	0.025	0.000	1.000	0.023
g	**2106**	**a**	0.310	0.074	1.000	0.000	0.175	0.206
t	**2132**	**a**	0.310	0.074	1.000	0.000	0.103	0.177
t	**2158**	**a**	0.310	0.074	1.000	0.000	0.082	0.129
t	**2184**	**c**	0.311	0.091	0.017	1.000	0.000	0.006
g	**2210**	**a**	0.310	0.098	1.000	0.000	0.130	0.219
g	**2236**	**t**	0.248	0.073	0.003	0.000	0.043	1.000
t	**2262**	**c**	0.311	0.105	0.247	1.000	0.109	0.000
c	**2288**	**g**	0.131	0.038	0.062	0.000	1.000	0.032
c	**2314**	**a**	0.310	0.074	1.000	0.000	0.162	0.174
t	**2340**	**c**	0.311	0.091	0.098	1.000	0.000	0.071
a	**2366**	**a**	0.310	0.089	1.000	0.003	0.000	0.022
g	**2392**	**t**	0.248	0.084	0.027	0.000	0.192	1.000
c	**2418**	**a**	0.310	0.074	1.000	0.000	0.164	0.161
c	**2444**	**g**	0.131	0.043	0.000	0.062	1.000	0.068
t	**2470**	**t**	0.248	0.073	0.000	0.099	0.252	1.000
t	**2496**	**g**	0.131	0.038	0.020	0.000	1.000	0.092
t	**2522**	**g**	0.131	0.025	0.000	0.015	1.000	0.010
c	**2548**	**c**	0.311	0.105	0.161	1.000	0.075	0.000
t	**2574**	**g**	0.131	0.025	0.024	0.000	1.000	0.052
a	**2600**	**c**	0.311	0.026	0.054	1.000	0.000	0.042

t	2626	a	0.310	0.074	1.000	0.000	0.219	0.207
t	2652	g	0.131	0.025	0.000	0.008	1.000	0.017
a	2678	t	0.248	0.084	0.070	0.000	0.222	1.000
g	2704	a	0.310	0.049	1.000	0.000	0.174	0.222
c	2730	a	0.310	0.089	1.000	0.000	0.199	0.029
t	2756	c	0.311	0.089	0.131	1.000	0.130	0.000
c	2782	a	0.310	0.098	1.000	0.000	0.172	0.355
t	2808	g	0.131	0.038	0.010	0.000	1.000	0.005
t	2834	c	0.311	0.026	0.000	1.000	0.005	0.020
a	2860	c	0.311	0.089	0.065	1.000	0.000	0.050
g	2886	a	0.310	0.089	1.000	0.000	0.001	0.047
t	2912	c	0.311	0.105	0.134	1.000	0.041	0.000
a	2938	c	0.311	0.091	0.054	1.000	0.000	0.055
a	2964	t	0.248	0.084	0.162	0.000	0.073	1.000
g	2990	t	0.248	0.030	0.017	0.000	0.159	1.000
a	3016	c	0.311	0.105	0.127	1.000	0.000	0.003
t	3042	t	0.248	0.084	0.000	0.002	0.225	1.000
t	3068	g	0.131	0.043	0.014	0.022	1.000	0.000
a	3094	g	0.131	0.025	0.000	0.009	1.000	0.043
c	3120	t	0.248	0.030	0.040	0.000	0.168	1.000
a	3146	c	0.311	0.091	0.126	1.000	0.000	0.055
c	3172	a	0.310	0.049	1.000	0.000	0.151	0.245
a	3198	t	0.248	0.084	0.073	0.000	0.183	1.000
t	3224	a	0.310	0.089	1.000	0.000	0.126	0.024
g	3250	c	0.311	0.089	0.130	1.000	0.000	0.061
c	3276	t	0.248	0.030	0.000	0.027	0.062	1.000
a	3302	c	0.311	0.089	0.041	1.000	0.009	0.000
a	3328	c	0.311	0.026	0.112	1.000	0.000	0.169
g	3354	t	0.248	0.084	0.076	0.000	0.093	1.000
c	3380	g	0.131	0.043	0.015	0.128	1.000	0.000
a	3406	c	0.311	0.091	0.055	1.000	0.000	0.136
t	3432	t	0.248	0.084	0.006	0.000	0.252	1.000
c	3458	t	0.248	0.061	0.006	0.000	0.036	1.000
c	3484	c	0.311	0.089	0.139	1.000	0.000	0.081
c	3510	t	0.248	0.061	0.000	0.011	0.015	1.000

c	3536	t	0.248	0.073	0.000	0.042	0.167	1.000
g	3562	a	0.310	0.074	1.000	0.000	0.090	0.180
t	3588	t	0.248	0.073	0.000	0.019	0.106	1.000
t	3614	c	0.311	0.089	0.046	1.000	0.098	0.000
c	3640	a	0.310	0.074	1.000	0.000	0.136	0.141
c	3666	t	0.248	0.073	0.039	0.000	0.279	1.000
a	3692	g	0.131	0.038	0.019	0.000	1.000	0.009
g	3718	t	0.248	0.061	0.008	0.000	0.058	1.000
t	3744	a	0.310	0.098	1.000	0.000	0.039	0.048
g	3770	a	0.310	0.098	1.000	0.000	0.087	0.052
a	3796	a	0.310	0.098	1.000	0.048	0.000	0.115
g	3822	c	0.311	0.091	0.102	1.000	0.000	0.059
t	3848	a	0.310	0.049	1.000	0.000	0.053	0.149
t	3874	g	0.131	0.025	0.000	0.013	1.000	0.079
c	3900	t	0.248	0.084	0.042	0.000	0.065	1.000
a	3926	a	0.310	0.089	1.000	0.000	0.190	0.079
c	3952	a	0.310	0.074	1.000	0.000	0.087	0.151
c	3978	t	0.248	0.061	0.033	0.000	0.093	1.000
c	4004	a	0.310	0.089	1.000	0.022	0.000	0.020
t	4030	a	0.310	0.089	1.000	0.000	0.267	0.041
c	4056	a	0.310	0.049	1.000	0.000	0.120	0.398
t	4082	c	0.311	0.026	0.079	1.000	0.013	0.000
a	4108	c	0.311	0.089	0.031	1.000	0.000	0.013
a	4134	a	0.310	0.089	1.000	0.000	0.113	0.034
a	4160	a	0.310	0.074	1.000	0.000	0.046	0.093
t	4186	a	0.310	0.074	1.000	0.016	0.000	0.019
c	4212	a	0.310	0.089	1.000	0.006	0.019	0.000
a	4238	g	0.131	0.038	0.000	0.011	1.000	0.056
c	4264	c	0.311	0.105	0.175	1.000	0.000	0.008
c	4290	c	0.311	0.091	0.082	1.000	0.000	0.077
a	4316	c	0.311	0.091	0.114	1.000	0.000	0.057
c	4342	c	0.311	0.089	0.117	1.000	0.057	0.000
g	4368	c	0.311	0.089	0.089	1.000	0.012	0.000
a	4394	g	0.131	0.025	0.005	0.000	1.000	0.006
t	4420	c	0.311	0.089	0.159	1.000	0.167	0.000

c	4446	a	0.310	0.098	1.000	0.000	0.189	0.154
a	4472	c	0.311	0.091	0.007	1.000	0.016	0.000
a	4498	a	0.310	0.089	1.000	0.019	0.092	0.000
a	4524	t	0.248	0.084	0.000	0.032	0.072	1.000
a	4550	t	0.248	0.084	0.124	0.000	0.282	1.000
g	4576	a	0.310	0.089	1.000	0.010	0.132	0.000
g	4602	c	0.311	0.091	0.035	1.000	0.000	0.072
a	4628	t	0.248	0.061	0.121	0.000	0.272	1.000
a	4654	g	0.131	0.025	0.000	0.005	1.000	0.078
c	4680	g	0.131	0.043	0.000	0.050	1.000	0.096
a	4706	t	0.248	0.073	0.000	0.035	0.227	1.000
a	4732	c	0.311	0.105	0.135	1.000	0.108	0.000
g	4758	c	0.311	0.105	0.132	1.000	0.000	0.032
c	4784	t	0.248	0.084	0.059	0.000	0.279	1.000
a	4810	c	0.311	0.089	0.080	1.000	0.078	0.000
t	4836	a	0.310	0.098	1.000	0.000	0.181	0.149
c	4862	c	0.311	0.105	0.092	1.000	0.089	0.000
a	4888	g	0.131	0.043	0.021	0.074	1.000	0.000
a	4914	c	0.311	0.089	0.045	1.000	0.000	0.016
g	4940	a	0.310	0.049	1.000	0.000	0.175	0.158
c	4966	g	0.131	0.038	0.075	0.000	1.000	0.008
a	4992	c	0.311	0.091	0.001	1.000	0.010	0.000
c	5018	g	0.131	0.025	0.042	0.000	1.000	0.039
g	5044	a	0.310	0.089	1.000	0.031	0.000	0.101
c	5070	t	0.248	0.084	0.045	0.000	0.154	1.000
a	5096	g	0.131	0.025	0.000	0.079	1.000	0.133
g	5122	a	0.310	0.098	1.000	0.000	0.106	0.158
c	5148	c	0.311	0.089	0.102	1.000	0.084	0.000
a	5174	g	0.131	0.038	0.271	0.000	1.000	0.146
a	5200	g	0.131	0.025	0.029	0.000	1.000	0.061
t	5226	c	0.311	0.091	0.000	1.000	0.076	0.061
g	5252	g	0.131	0.043	0.000	0.004	1.000	0.022
c	5278	t	0.248	0.073	0.060	0.000	0.464	1.000
a	5304	a	0.310	0.089	1.000	0.000	0.096	0.064
g	5330	t	0.248	0.030	0.000	0.000	0.027	1.000

c	5356	t	0.248	0.073	0.000	0.009	0.224	1.000
t	5382	a	0.310	0.089	1.000	0.000	0.190	0.048
c	5408	c	0.311	0.089	0.143	1.000	0.000	0.070
a	5434	c	0.311	0.089	0.151	1.000	0.046	0.000
a	5460	g	0.131	0.043	0.041	0.002	1.000	0.000
a	5486	a	0.310	0.089	1.000	0.000	0.068	0.049
a	5512	c	0.311	0.091	0.096	1.000	0.000	0.088
c	5538	a	0.310	0.049	1.000	0.000	0.141	0.200
g	5564	t	0.248	0.061	0.062	0.000	0.065	1.000
c	5590	a	0.310	0.074	1.000	0.000	0.160	0.193
t	5616	c	0.311	0.105	0.074	1.000	0.000	0.101
t	5642	t	0.248	0.073	0.000	0.077	0.116	1.000
a	5668	c	0.311	0.089	0.181	1.000	0.136	0.000
g	5694	a	0.310	0.049	1.000	0.106	0.000	0.331
c	5720	c	0.311	0.089	0.106	1.000	0.047	0.000
c	5746	c	0.311	0.105	0.201	1.000	0.000	0.017
t	5772	c	0.311	0.089	0.098	1.000	0.000	0.016
a	5798	c	0.311	0.105	0.222	1.000	0.000	0.095
g	5824	a	0.310	0.074	1.000	0.000	0.239	0.160
c	5850	c	0.311	0.105	0.070	1.000	0.111	0.000
c	5876	c	0.311	0.105	0.166	1.000	0.000	0.019
a	5902	a	0.310	0.089	1.000	0.000	0.097	0.085
c	5928	c	0.311	0.105	0.107	1.000	0.000	0.040
a	5954	t	0.248	0.084	0.035	0.000	0.152	1.000
c	5980	a	0.310	0.089	1.000	0.032	0.131	0.000
c	6006	a	0.310	0.049	1.000	0.000	0.092	0.372
c	6032	t	0.248	0.084	0.000	0.003	0.203	1.000
c	6058	g	0.131	0.025	0.028	0.000	1.000	0.044
c	6084	c	0.311	0.026	0.078	1.000	0.000	0.124
a	6110	a	0.310	0.089	1.000	0.000	0.087	0.038
c	6136	g	0.131	0.025	0.000	0.012	1.000	0.063
g	6162	g	0.131	0.038	0.005	0.000	1.000	0.051
g	6188	t	0.248	0.073	0.027	0.000	0.239	1.000
g	6214	g	0.131	0.025	0.000	0.012	1.000	0.015
a	6240	a	0.310	0.074	1.000	0.000	0.118	0.240

a	6266	t	0.248	0.073	0.040	0.000	0.209	1.000
a	6292	t	0.248	0.061	0.139	0.000	0.249	1.000
c	6318	a	0.310	0.098	1.000	0.000	0.075	0.113
a	6344	t	0.248	0.030	0.000	0.044	0.211	1.000
g	6370	c	0.311	0.089	0.011	1.000	0.000	0.040
c	6396	c	0.311	0.091	0.000	1.000	0.042	0.083
a	6422	a	0.310	0.089	1.000	0.012	0.152	0.000
g	6448	g	0.131	0.025	0.000	0.000	1.000	0.018
t	6474	a	0.310	0.074	1.000	0.000	0.159	0.161
g	6500	t	0.248	0.084	0.089	0.000	0.215	1.000
a	6526	c	0.311	0.089	0.065	1.000	0.060	0.000
t	6552	a	0.310	0.098	1.000	0.000	0.041	0.182
t	6578	t	0.248	0.073	0.016	0.000	0.151	1.000
a	6604	t	0.248	0.061	0.019	0.000	0.083	1.000
a	6630	a	0.310	0.074	1.000	0.000	0.105	0.113
c	6656	a	0.310	0.089	1.000	0.058	0.178	0.000
c	6682	a	0.310	0.098	1.000	0.000	0.034	0.039
t	6708	c	0.311	0.091	0.000	1.000	0.020	0.116
t	6734	t	0.248	0.084	0.000	0.004	0.159	1.000
t	6760	g	0.131	0.043	0.067	0.135	1.000	0.000
a	6786	a	0.310	0.049	1.000	0.000	0.110	0.176
g	6812	g	0.131	0.038	0.000	0.000	1.000	0.018
c	6838	c	0.311	0.091	0.096	1.000	0.000	0.093
a	6864	a	0.310	0.098	1.000	0.000	0.055	0.063
a	6890	c	0.311	0.026	0.013	1.000	0.000	0.041
t	6916	c	0.311	0.091	0.030	1.000	0.000	0.030
a	6942	c	0.311	0.091	0.054	1.000	0.000	0.049
a	6968	a	0.310	0.074	1.000	0.000	0.257	0.191
a	6994	g	0.131	0.043	0.000	0.011	1.000	0.025
c	7020	c	0.311	0.091	0.046	1.000	0.000	0.007
g	7046	c	0.311	0.026	0.000	1.000	0.073	0.097
a	7072	t	0.248	0.061	0.003	0.000	0.098	1.000
a	7098	a	0.310	0.074	1.000	0.000	0.139	0.225
a	7124	a	0.310	0.049	1.000	0.000	0.248	0.283
g	7150	t	0.248	0.073	0.000	0.117	0.310	1.000

t	7176	a	0.310	0.074	1.000	0.000	0.030	0.181
t	7202	c	0.311	0.026	0.000	1.000	0.017	0.048
t	7228	t	0.248	0.073	0.028	0.000	0.273	1.000
a	7254	c	0.311	0.089	0.099	1.000	0.000	0.037
a	7280	g	0.131	0.025	0.001	0.000	1.000	0.022
c	7306	a	0.310	0.098	1.000	0.000	0.089	0.109
t	7332	a	0.310	0.089	1.000	0.012	0.038	0.000
a	7358	a	0.310	0.089	1.000	0.000	0.043	0.013
a	7384	c	0.311	0.105	0.154	1.000	0.075	0.000
g	7410	c	0.311	0.089	0.173	1.000	0.000	0.063
c	7436	a	0.310	0.049	1.000	0.000	0.080	0.170
t	7462	t	0.248	0.061	0.005	0.000	0.085	1.000
a	7488	g	0.131	0.025	0.000	0.001	1.000	0.008
t	7514	a	0.310	0.074	1.000	0.000	0.212	0.256
a	7540	c	0.311	0.105	0.184	1.000	0.000	0.021
c	7566	a	0.310	0.089	1.000	0.000	0.181	0.114
t	7592	c	0.311	0.026	0.041	1.000	0.000	0.050
a	7618	c	0.311	0.091	0.023	1.000	0.005	0.000
a	7644	c	0.311	0.089	0.027	1.000	0.032	0.000
c	7670	t	0.248	0.030	0.010	0.000	0.103	1.000
c	7696	a	0.310	0.049	1.000	0.000	0.147	0.288
c	7722	c	0.311	0.105	0.098	1.000	0.000	0.000
c	7748	c	0.311	0.105	0.006	1.000	0.034	0.000
a	7774	c	0.311	0.091	0.045	1.000	0.000	0.025
g	7800	c	0.311	0.091	0.123	1.000	0.000	0.032
g	7826	c	0.311	0.091	0.080	1.000	0.000	0.092
g	7852	t	0.248	0.084	0.054	0.000	0.161	1.000
t	7878	t	0.248	0.061	0.030	0.000	0.135	1.000
t	7904	c	0.311	0.089	0.077	1.000	0.026	0.000
g	7930	a	0.310	0.074	1.000	0.000	0.044	0.060
g	7956	t	0.248	0.030	0.133	0.000	0.297	1.000
t	7982	a	0.310	0.074	1.000	0.000	0.258	0.269
c	8008	c	0.311	0.026	0.086	1.000	0.000	0.093
a	8034	c	0.311	0.105	0.122	1.000	0.063	0.000
a	8060	t	0.248	0.084	0.091	0.000	0.266	1.000

t	8086	c	0.311	0.091	0.043	1.000	0.000	0.094
t	8112	c	0.311	0.105	0.078	1.000	0.214	0.000
t	8138	a	0.310	0.098	1.000	0.000	0.123	0.098
c	8164	t	0.248	0.084	0.076	0.000	0.112	1.000
g	8190	t	0.248	0.061	0.001	0.000	0.323	1.000
t	8216	c	0.311	0.089	0.151	1.000	0.003	0.000
g	8242	c	0.311	0.105	0.138	1.000	0.024	0.000
c	8268	c	0.311	0.105	0.329	1.000	0.000	0.028
c	8294	a	0.310	0.074	1.000	0.000	0.063	0.093
a	8320	a	0.310	0.098	1.000	0.000	0.077	0.083
g	8346	a	0.310	0.098	1.000	0.000	0.131	0.210
c	8372	a	0.310	0.074	1.000	0.000	0.095	0.080
c	8398	c	0.311	0.105	0.269	1.000	0.089	0.000
a	8424	c	0.311	0.105	0.385	1.000	0.000	0.224
c	8450	g	0.131	0.043	0.000	0.039	1.000	0.042
c	8476	t	0.248	0.030	0.000	0.035	0.081	1.000
g	8502	c	0.311	0.105	0.081	1.000	0.059	0.000
c	8528	a	0.310	0.074	1.000	0.000	0.142	0.124
g	8554	t	0.248	0.084	0.032	0.000	0.135	1.000
g	8580	t	0.248	0.073	0.089	0.000	0.313	1.000
t	8606	g	0.131	0.025	0.033	0.000	1.000	0.037
c	8632	a	0.310	0.074	1.000	0.000	0.245	0.378
a	8658	t	0.248	0.073	0.021	0.000	0.028	1.000
c	8684	c	0.311	0.105	0.000	1.000	0.013	0.015
a	8710	c	0.311	0.089	0.045	1.000	0.030	0.000
c	8736	a	0.310	0.074	1.000	0.000	0.126	0.133
g	8762	t	0.248	0.073	0.000	0.035	0.104	1.000
a	8788	c	0.311	0.089	0.057	1.000	0.017	0.000
t	8814	t	0.248	0.030	0.065	0.000	0.170	1.000
t	8840	g	0.131	0.043	0.048	0.127	1.000	0.000
a	8866	g	0.131	0.025	0.071	0.000	1.000	0.037
a	8892	t	0.248	0.084	0.004	0.000	0.034	1.000
c	8918	c	0.311	0.091	0.000	1.000	0.068	0.094
c	8944	t	0.248	0.073	0.000	0.151	0.100	1.000
c	8970	a	0.310	0.049	1.000	0.064	0.000	0.232

a	8996	t	0.248	0.073	0.009	0.000	0.231	1.000
a	9022	c	0.311	0.105	0.070	1.000	0.000	0.016
g	9048	c	0.311	0.026	0.064	1.000	0.000	0.075
t	9074	c	0.311	0.089	0.165	1.000	0.060	0.000
c	9100	c	0.311	0.026	0.000	1.000	0.003	0.035
a	9126	t	0.248	0.084	0.022	0.000	0.132	1.000
a	9152	t	0.248	0.061	0.038	0.000	0.088	1.000
t	9178	t	0.248	0.061	0.062	0.000	0.104	1.000
a	9204	c	0.311	0.026	0.050	1.000	0.000	0.005
g	9230	t	0.248	0.084	0.027	0.000	0.430	1.000
a	9256	g	0.131	0.025	0.001	0.000	1.000	0.037
a	9282	t	0.248	0.084	0.117	0.000	0.217	1.000
g	9308	t	0.248	0.061	0.000	0.009	0.091	1.000
c	9334	g	0.131	0.025	0.000	0.028	1.000	0.145
c	9360	g	0.131	0.043	0.000	0.047	1.000	0.043
g	9386	t	0.248	0.061	0.012	0.000	0.097	1.000
g	9412	g	0.131	0.025	0.008	0.000	1.000	0.018
c	9438	t	0.248	0.084	0.017	0.000	0.214	1.000
g	9464	t	0.248	0.061	0.000	0.013	0.040	1.000
t	9490	a	0.310	0.098	1.000	0.000	0.040	0.031
a	9516	c	0.311	0.091	0.000	1.000	0.041	0.050
a	9542	t	0.248	0.061	0.043	0.000	0.079	1.000
a	9568	a	0.310	0.098	1.000	0.000	0.262	0.272
g	9594	t	0.248	0.061	0.000	0.093	0.055	1.000
a	9620	c	0.311	0.105	0.173	1.000	0.000	0.146
g	9646	a	0.310	0.098	1.000	0.000	0.006	0.059
t	9672	c	0.311	0.089	0.018	1.000	0.000	0.050
g	9698	t	0.248	0.073	0.000	0.097	0.243	1.000
t	9724	c	0.311	0.091	0.046	1.000	0.000	0.019
t	9750	t	0.248	0.084	0.055	0.000	0.047	1.000
t	9776	t	0.248	0.073	0.033	0.000	0.291	1.000
t	9802	t	0.248	0.030	0.000	0.000	0.091	1.000
a	9828	c	0.311	0.026	0.015	1.000	0.013	0.000
g	9854	c	0.311	0.026	0.000	1.000	0.069	0.055
a	9880	g	0.131	0.038	0.000	0.003	1.000	0.043

t	9906	c	0.311	0.091	0.023	1.000	0.000	0.090
c	9932	t	0.248	0.061	0.028	0.000	0.068	1.000
a	9958	g	0.131	0.025	0.018	0.000	1.000	0.030
c	9984	c	0.311	0.105	0.306	1.000	0.052	0.000
c	10010	t	0.248	0.084	0.051	0.000	0.144	1.000
c	10036	a	0.310	0.074	1.000	0.000	0.230	0.234
c	10062	a	0.310	0.049	1.000	0.000	0.072	0.256
c	10088	t	0.248	0.084	0.000	0.016	0.075	1.000
t	10114	g	0.131	0.025	0.064	0.000	1.000	0.111
c	10140	c	0.311	0.089	0.094	1.000	0.000	0.031
c	10166	t	0.248	0.084	0.028	0.000	0.282	1.000
c	10192	t	0.248	0.084	0.058	0.000	0.152	1.000
c	10218	t	0.248	0.061	0.021	0.000	0.108	1.000
a	10244	a	0.310	0.098	1.000	0.000	0.075	0.074
a	10270	c	0.311	0.105	0.140	1.000	0.076	0.000
t	10296	a	0.310	0.074	1.000	0.000	0.033	0.033
a	10322	c	0.311	0.091	0.036	1.000	0.000	0.045
a	10348	c	0.311	0.091	0.079	1.000	0.000	0.048
a	10374	a	0.310	0.089	1.000	0.192	0.038	0.000
g	10400	a	0.310	0.089	1.000	0.031	0.188	0.000
c	10426	g	0.131	0.043	0.091	0.128	1.000	0.000
t	10452	c	0.311	0.089	0.044	1.000	0.014	0.000
a	10478	t	0.248	0.084	0.051	0.000	0.196	1.000
a	10504	t	0.248	0.084	0.044	0.000	0.159	1.000
a	10530	c	0.311	0.089	0.044	1.000	0.053	0.000
a	10556	a	0.310	0.098	1.000	0.000	0.098	0.099
c	10582	t	0.248	0.073	0.030	0.000	0.245	1.000
t	10608	g	0.131	0.038	0.071	0.000	1.000	0.060
c	10634	c	0.311	0.105	0.178	1.000	0.000	0.073
a	10660	t	0.248	0.061	0.000	0.035	0.041	1.000
c	10686	a	0.310	0.089	1.000	0.000	0.293	0.065
c	10712	c	0.311	0.091	0.000	1.000	0.008	0.059
t	10738	a	0.310	0.098	1.000	0.000	0.015	0.078
g	10764	a	0.310	0.098	1.000	0.000	0.089	0.108
a	10790	a	0.310	0.049	1.000	0.000	0.200	0.141

g	10816	a	0.310	0.089	1.000	0.000	0.273	0.195
t	10842	g	0.131	0.043	0.094	0.059	1.000	0.000
t	10868	g	0.131	0.025	0.000	0.018	1.000	0.058
g	10894	t	0.248	0.084	0.168	0.000	0.131	1.000
t	10920	c	0.311	0.026	0.305	1.000	0.000	0.091
a	10946	g	0.131	0.038	0.000	0.014	1.000	0.062
a	10972	t	0.248	0.061	0.000	0.065	0.185	1.000
a	10998	a	0.310	0.074	1.000	0.000	0.121	0.205
a	11024	a	0.310	0.098	1.000	0.003	0.000	0.179
a	11050	a	0.310	0.089	1.000	0.000	0.088	0.070
a	11076	g	0.131	0.025	0.000	0.003	1.000	0.008
c	11102	c	0.311	0.105	0.115	1.000	0.034	0.000
t	11128	t	0.248	0.061	0.076	0.000	0.268	1.000
c	11154	t	0.248	0.073	0.019	0.000	0.046	1.000
c	11180	c	0.311	0.089	0.088	1.000	0.076	0.000
a	11206	g	0.131	0.043	0.043	0.083	1.000	0.000
g	11232	c	0.311	0.091	0.133	1.000	0.000	0.129
t	11258	t	0.248	0.061	0.016	0.000	0.079	1.000
t	11284	a	0.310	0.089	1.000	0.045	0.000	0.012
g	11310	c	0.311	0.089	0.043	1.000	0.026	0.000
a	11336	a	0.310	0.098	1.000	0.000	0.177	0.260
c	11362	c	0.311	0.089	0.050	1.000	0.055	0.000
a	11388	c	0.311	0.089	0.149	1.000	0.024	0.000
c	11414	c	0.311	0.091	0.000	1.000	0.059	0.121
a	11440	c	0.311	0.091	0.082	1.000	0.000	0.082
a	11466	a	0.310	0.098	1.000	0.000	0.066	0.101
a	11492	c	0.311	0.105	0.141	1.000	0.064	0.000
a	11518	t	0.248	0.061	0.117	0.000	0.274	1.000
t	11544	a	0.310	0.074	1.000	0.000	0.032	0.154
a	11570	a	0.310	0.074	1.000	0.000	0.093	0.120
g	11596	g	0.131	0.025	0.010	0.000	1.000	0.067
a	11622	c	0.311	0.105	0.135	1.000	0.000	0.066
c	11648	a	0.310	0.049	1.000	0.000	0.116	0.249
t	11674	c	0.311	0.089	0.084	1.000	0.007	0.000
a	11700	a	0.310	0.098	1.000	0.000	0.129	0.215

c	11726	t	0.248	0.084	0.034	0.000	0.187	1.000
g	11752	c	0.311	0.105	0.084	1.000	0.000	0.055
a	11778	c	0.311	0.089	0.061	1.000	0.013	0.000
a	11804	t	0.248	0.073	0.000	0.082	0.115	1.000
a	11830	c	0.311	0.105	0.188	1.000	0.005	0.000
g	11856	c	0.311	0.091	0.056	1.000	0.000	0.096
t	11882	g	0.131	0.025	0.034	0.000	1.000	0.053
g	11908	c	0.311	0.089	0.074	1.000	0.000	0.067
g	11934	t	0.248	0.084	0.020	0.000	0.205	1.000
c	11960	t	0.248	0.030	0.000	0.059	0.215	1.000
t	11986	c	0.311	0.105	0.025	1.000	0.000	0.030
t	12012	a	0.310	0.089	1.000	0.000	0.139	0.038
t	12038	g	0.131	0.025	0.000	0.013	1.000	0.048
a	12064	c	0.311	0.091	0.049	1.000	0.000	0.046
a	12090	a	0.310	0.089	1.000	0.026	0.217	0.000
c	12116	t	0.248	0.061	0.000	0.018	0.061	1.000
a	12142	t	0.248	0.084	0.027	1.000	0.000	0.321
t	12168	a	0.310	0.089	1.000	0.000	0.060	0.068
a	12194	t	0.248	0.030	0.097	0.000	0.132	1.000
t	12220	c	0.311	0.089	0.061	1.000	0.000	0.034
c	12246	g	0.131	0.038	0.037	0.095	1.000	0.000
t	12272	c	0.311	0.091	0.135	1.000	0.000	0.108
g	12298	c	0.311	0.026	0.000	1.000	0.048	0.071
a	12324	c	0.311	0.091	0.063	1.000	0.000	0.079
a	12350	t	0.248	0.030	0.048	0.000	0.171	1.000
c	12376	a	0.310	0.098	1.000	0.000	0.030	0.043
a	12402	t	0.248	0.084	0.074	0.000	0.138	1.000
c	12428	g	0.131	0.025	0.014	0.023	1.000	0.000
a	12454	c	0.311	0.105	0.053	1.000	0.064	0.000
c	12480	g	0.131	0.043	0.019	0.118	1.000	0.000
a	12506	c	0.311	0.091	0.114	1.000	0.000	0.070
a	12532	a	0.310	0.098	1.000	0.000	0.111	0.060
t	12558	c	0.311	0.105	0.179	1.000	0.000	0.020
a	12584	c	0.311	0.091	0.086	1.000	0.000	0.041
g	12610	t	0.248	0.084	0.162	0.000	0.253	1.000

c	12636	g	0.131	0.043	0.000	0.064	1.000	0.087
t	12662	a	0.310	0.098	1.000	0.000	0.127	0.135
a	12688	t	0.248	0.073	0.000	0.184	0.461	1.000
a	12714	t	0.248	0.073	0.090	0.000	0.431	1.000
g	12740	t	0.248	0.061	0.145	0.000	0.196	1.000
a	12766	c	0.311	0.089	0.158	1.000	0.000	0.039
c	12792	c	0.311	0.105	0.235	1.000	0.000	0.144
c	12818	a	0.310	0.098	1.000	0.000	0.176	0.156
c	12844	a	0.310	0.049	1.000	0.071	0.000	0.281
a	12870	c	0.311	0.089	0.191	1.000	0.082	0.000
a	12896	c	0.311	0.089	0.099	1.000	0.000	0.011
a	12922	c	0.311	0.091	0.047	1.000	0.000	0.081
c	12948	t	0.248	0.073	0.059	0.000	0.420	1.000
t	12974	a	0.310	0.089	1.000	0.000	0.127	0.068
g	13000	t	0.248	0.073	0.022	0.000	0.102	1.000
g	13026	c	0.311	0.105	0.224	1.000	0.001	0.000
g	13052	a	0.310	0.089	1.000	0.000	0.204	0.071
a	13078	c	0.311	0.091	0.114	1.000	0.000	0.052
t	13104	a	0.310	0.089	1.000	0.104	0.127	0.000
t	13130	c	0.311	0.026	0.020	1.000	0.000	0.018
a	13156	a	0.310	0.049	1.000	0.000	0.061	0.305
g	13182	c	0.311	0.105	0.243	1.000	0.000	0.156
a	13208	a	0.310	0.074	1.000	0.000	0.084	0.180
t	13234	c	0.311	0.105	0.167	1.000	0.000	0.024
a	13260	t	0.248	0.084	0.085	0.000	0.148	1.000
c	13286	a	0.310	0.089	1.000	0.000	0.113	0.017
c	13312	a	0.310	0.089	1.000	0.000	0.120	0.013
c	13338	t	0.248	0.084	0.118	0.000	0.269	1.000
c	13364	c	0.311	0.091	0.121	1.000	0.000	0.137
a	13390	t	0.248	0.073	0.008	0.000	0.116	1.000
c	13416	t	0.248	0.073	0.000	0.085	0.176	1.000
t	13442	a	0.310	0.098	1.000	0.000	0.052	0.001
a	13468	g	0.131	0.043	0.000	0.012	1.000	0.030
t	13494	c	0.311	0.105	0.124	1.000	0.146	0.000
g	13520	c	0.311	0.089	0.010	1.000	0.025	0.000

c	13546	a	0.310	0.089	1.000	0.000	0.312	0.026
t	13572	c	0.311	0.091	0.046	1.000	0.000	0.041
t	13598	a	0.310	0.098	1.000	0.000	0.155	0.095
a	13624	a	0.310	0.074	1.000	0.000	0.142	0.186
g	13650	c	0.311	0.091	0.095	1.000	0.000	0.107
c	13676	c	0.311	0.026	0.045	1.000	0.000	0.051
c	13702	a	0.310	0.089	1.000	0.068	0.094	0.000
c	13728	a	0.310	0.074	1.000	0.000	0.172	0.109
t	13754	t	0.248	0.084	0.030	0.000	0.118	1.000
a	13780	c	0.311	0.089	0.086	1.000	0.000	0.061
a	13806	t	0.248	0.073	0.000	0.101	0.141	1.000
a	13832	c	0.311	0.091	0.127	1.000	0.000	0.100
c	13858	a	0.310	0.074	1.000	0.016	0.000	0.141
c	13884	a	0.310	0.074	1.000	0.000	0.106	0.136
t	13910	c	0.311	0.091	0.018	1.000	0.000	0.118
c	13936	a	0.310	0.098	1.000	0.000	0.056	0.039
a	13962	c	0.311	0.105	0.116	1.000	0.078	0.000
a	13988	g	0.131	0.038	0.000	0.099	1.000	0.035
c	14014	a	0.310	0.098	1.000	0.000	0.077	0.055
a	14040	g	0.131	0.038	0.000	0.013	1.000	0.032
g	14066	t	0.248	0.084	0.283	0.000	0.327	1.000
t	14092	c	0.311	0.091	0.000	1.000	0.042	0.029
t	14118	a	0.310	0.089	1.000	0.000	0.152	0.153
a	14144	c	0.311	0.105	0.099	1.000	0.115	0.000
a	14170	c	0.311	0.091	0.000	1.000	0.019	0.046
a	14196	a	0.310	0.098	1.000	0.000	0.021	0.236
t	14222	c	0.311	0.089	0.060	1.000	0.089	0.000
c	14248	c	0.311	0.105	0.275	1.000	0.000	0.121
a	14274	g	0.131	0.025	0.000	0.043	1.000	0.007
a	14300	c	0.311	0.105	0.310	1.000	0.000	0.066
c	14326	t	0.248	0.084	0.064	0.000	0.175	1.000
a	14352	t	0.248	0.073	0.012	0.000	0.114	1.000
a	14378	c	0.311	0.091	0.072	1.000	0.000	0.083
a	14404	g	0.131	0.038	0.016	0.000	1.000	0.034
a	14430	g	0.131	0.025	0.035	0.000	1.000	0.020

c	14456	c	0.311	0.089	0.071	1.000	0.000	0.060
t	14482	a	0.310	0.074	1.000	0.000	0.096	0.166
g	14508	a	0.310	0.089	1.000	0.024	0.049	0.000
c	14534	g	0.131	0.043	0.000	0.010	1.000	0.006
t	14560	t	0.248	0.084	0.001	0.000	0.157	1.000
c	14586	a	0.310	0.074	1.000	0.000	0.070	0.091
g	14612	t	0.248	0.084	0.073	0.000	0.108	1.000
c	14638	a	0.310	0.089	1.000	0.000	0.025	0.096
c	14664	a	0.310	0.074	1.000	0.000	0.077	0.106
a	14690	t	0.248	0.073	0.000	0.066	0.295	1.000
g	14716	t	0.248	0.061	0.000	0.003	0.079	1.000
a	14742	g	0.131	0.043	0.000	0.007	1.000	0.065
a	14768	a	0.310	0.074	1.000	0.000	0.179	0.222
c	14794	a	0.310	0.089	1.000	0.218	0.137	0.000
a	14820	g	0.131	0.038	0.000	0.047	1.000	0.017
c	14846	c	0.311	0.091	0.000	1.000	0.041	0.080
t	14872	t	0.248	0.061	0.000	0.057	0.137	1.000
a	14898	t	0.248	0.061	0.011	0.000	0.102	1.000
c	14924	a	0.310	0.049	1.000	0.048	0.000	0.306
g	14950	c	0.311	0.105	0.196	1.000	0.026	0.000
a	14976	a	0.310	0.074	1.000	0.000	0.145	0.225
g	15002	c	0.311	0.026	0.024	1.000	0.000	0.088
c	15028	c	0.311	0.089	0.000	1.000	0.099	0.065
c	15054	c	0.311	0.089	0.094	1.000	0.009	0.000
a	15080	a	0.310	0.098	1.000	0.000	0.127	0.209
c	15106	a	0.310	0.098	1.000	0.000	0.028	0.020
a	15132	c	0.311	0.089	0.060	1.000	0.070	0.000
g	15158	a	0.310	0.049	1.000	0.000	0.180	0.363
c	15184	g	0.131	0.038	0.025	0.000	1.000	0.046
t	15210	c	0.311	0.105	0.227	1.000	0.000	0.066
t	15236	a	0.310	0.074	1.000	0.000	0.017	0.047
a	15262	t	0.248	0.084	0.000	0.006	0.093	1.000
a	15288	a	0.310	0.074	1.000	0.000	0.022	0.027
a	15314	c	0.311	0.089	0.017	1.000	0.020	0.000
a	15340	c	0.311	0.105	0.011	1.000	0.029	0.000

c	**15366**	**c**	0.311	0.105	0.083	1.000	0.047	0.000
t	**15392**	**a**	0.310	0.098	1.000	0.000	0.066	0.064
c	**15418**	**c**	0.311	0.105	0.188	1.000	0.000	0.001
a	**15444**	**a**	0.310	0.074	1.000	0.000	0.129	0.100

Appendix B

Predicted normalized intensities with 20 neurons in the hidden layer using 1-grams for the respective nucleotides

CRS ref	Position	CRS ref (26)	1-gram values	A	C	G	T
c	26	a	0.310	1.000	0.013	0.105	0.127
a	52	c	0.311	0.098	1.000	0.023	0.041
c	78	t	0.248	0.033	0.019	0.162	1.000
a	104	g	0.131	0.027	0.023	1.000	0.039
g	130	c	0.311	0.098	1.000	0.023	0.041
t	156	c	0.311	0.098	1.000	0.023	0.041
t	182	a	0.310	1.000	0.013	0.105	0.127
t	208	c	0.311	0.098	1.000	0.023	0.041
a	234	c	0.311	0.098	1.000	0.023	0.041
t	260	t	0.248	0.033	0.019	0.162	1.000
g	286	c	0.311	0.098	1.000	0.023	0.041
t	312	t	0.248	0.033	0.019	0.162	1.000
a	338	a	0.310	1.000	0.013	0.105	0.127
g	364	g	0.131	0.027	0.023	1.000	0.039
c	390	c	0.311	0.098	1.000	0.023	0.041
t	416	g	0.131	0.027	0.023	1.000	0.039
t	442	a	0.310	1.000	0.013	0.105	0.127
a	468	t	0.248	0.033	0.019	0.162	1.000
c	494	c	0.311	0.098	1.000	0.023	0.041
c	520	g	0.131	0.027	0.023	1.000	0.039
t	546	a	0.310	1.000	0.013	0.105	0.127
c	572	t	0.248	0.033	0.019	0.162	1.000
c	598	g	0.131	0.027	0.023	1.000	0.039
t	624	a	0.310	1.000	0.013	0.105	0.127
c	650	a	0.310	1.000	0.013	0.105	0.127
a	676	c	0.311	0.098	1.000	0.023	0.041
a	702	c	0.311	0.098	1.000	0.023	0.041
a	728	a	0.310	1.000	0.013	0.105	0.127
g	754	a	0.310	1.000	0.013	0.105	0.127
c	780	c	0.311	0.098	1.000	0.023	0.041
a	806	c	0.311	0.098	1.000	0.023	0.041

a	832	a	0.310	1.000	0.013	0.105	0.127
t	858	a	0.310	1.000	0.013	0.105	0.127
a	884	t	0.248	0.033	0.019	0.162	1.000
c	910	a	0.310	1.000	0.013	0.105	0.127
a	936	c	0.311	0.098	1.000	0.023	0.041
c	962	c	0.311	0.098	1.000	0.023	0.041
t	988	t	0.248	0.033	0.019	0.162	1.000
g	1014	g	0.131	0.027	0.023	1.000	0.039
a	1040	c	0.311	0.098	1.000	0.023	0.041
a	1066	t	0.248	0.033	0.019	0.162	1.000
a	1092	g	0.131	0.027	0.023	1.000	0.039
a	1118	c	0.311	0.098	1.000	0.023	0.041
t	1144	t	0.248	0.033	0.019	0.162	1.000
g	1170	g	0.131	0.027	0.023	1.000	0.039
t	1196	g	0.131	0.027	0.023	1.000	0.039
t	1222	a	0.310	1.000	0.013	0.105	0.127
t	1248	a	0.310	1.000	0.013	0.105	0.127
a	1274	c	0.311	0.098	1.000	0.023	0.041
g	1300	t	0.248	0.033	0.019	0.162	1.000
a	1326	a	0.310	1.000	0.013	0.105	0.127
c	1352	t	0.248	0.033	0.019	0.162	1.000
g	1378	t	0.248	0.033	0.019	0.162	1.000
g	1404	t	0.248	0.033	0.019	0.162	1.000
g	1430	g	0.131	0.027	0.023	1.000	0.039
c	1456	g	0.131	0.027	0.023	1.000	0.039
t	1482	t	0.248	0.033	0.019	0.162	1.000
c	1508	t	0.248	0.033	0.019	0.162	1.000
a	1534	a	0.310	1.000	0.013	0.105	0.127
c	1560	a	0.310	1.000	0.013	0.105	0.127
a	1586	t	0.248	0.033	0.019	0.162	1.000
t	1612	a	0.310	1.000	0.013	0.105	0.127
c	1638	c	0.311	0.098	1.000	0.023	0.041
a	1664	c	0.311	0.098	1.000	0.023	0.041
c	1690	c	0.311	0.098	1.000	0.023	0.041
c	1716	a	0.310	1.000	0.013	0.105	0.127

c	1742	c	0.311	0.098	1.000	0.023	0.041
c	1768	c	0.311	0.098	1.000	0.023	0.041
a	1794	g	0.131	0.027	0.023	1.000	0.039
t	1820	t	0.248	0.033	0.019	0.162	1.000
a	1846	a	0.310	1.000	0.013	0.105	0.127
a	1872	a	0.310	1.000	0.013	0.105	0.127
a	1898	g	0.131	0.027	0.023	1.000	0.039
c	1924	g	0.131	0.027	0.023	1.000	0.039
a	1950	t	0.248	0.033	0.019	0.162	1.000
a	1976	c	0.311	0.098	1.000	0.023	0.041
a	2002	g	0.131	0.027	0.023	1.000	0.039
t	2028	a	0.310	1.000	0.013	0.105	0.127
a	2054	t	0.248	0.033	0.019	0.162	1.000
g	2080	g	0.131	0.027	0.023	1.000	0.039
g	2106	a	0.310	1.000	0.013	0.105	0.127
t	2132	a	0.310	1.000	0.013	0.105	0.127
t	2158	a	0.310	1.000	0.013	0.105	0.127
t	2184	c	0.311	0.098	1.000	0.023	0.041
g	2210	a	0.310	1.000	0.013	0.105	0.127
g	2236	t	0.248	0.033	0.019	0.162	1.000
t	2262	c	0.311	0.098	1.000	0.023	0.041
c	2288	g	0.131	0.027	0.023	1.000	0.039
c	2314	a	0.310	1.000	0.013	0.105	0.127
t	2340	c	0.311	0.098	1.000	0.023	0.041
a	2366	a	0.310	1.000	0.013	0.105	0.127
g	2392	t	0.248	0.033	0.019	0.162	1.000
c	2418	a	0.310	1.000	0.013	0.105	0.127
c	2444	g	0.131	0.027	0.023	1.000	0.039
t	2470	t	0.248	0.033	0.019	0.162	1.000
t	2496	g	0.131	0.027	0.023	1.000	0.039
t	2522	g	0.131	0.027	0.023	1.000	0.039
c	2548	c	0.311	0.098	1.000	0.023	0.041
t	2574	g	0.131	0.027	0.023	1.000	0.039
a	2600	c	0.311	0.098	1.000	0.023	0.041
t	2626	a	0.310	1.000	0.013	0.105	0.127

t	2652	g	0.131	0.027	0.023	1.000	0.039
a	2678	t	0.248	0.033	0.019	0.162	1.000
g	2704	a	0.310	1.000	0.013	0.105	0.127
c	2730	a	0.310	1.000	0.013	0.105	0.127
t	2756	c	0.311	0.098	1.000	0.023	0.041
c	2782	a	0.310	1.000	0.013	0.105	0.127
t	2808	g	0.131	0.027	0.023	1.000	0.039
t	2834	c	0.311	0.098	1.000	0.023	0.041
a	2860	c	0.311	0.098	1.000	0.023	0.041
g	2886	a	0.310	1.000	0.013	0.105	0.127
t	2912	c	0.311	0.098	1.000	0.023	0.041
a	2938	c	0.311	0.098	1.000	0.023	0.041
a	2964	t	0.248	0.033	0.019	0.162	1.000
g	2990	t	0.248	0.033	0.019	0.162	1.000
a	3016	c	0.311	0.098	1.000	0.023	0.041
t	3042	t	0.248	0.033	0.019	0.162	1.000
t	3068	g	0.131	0.027	0.023	1.000	0.039
a	3094	g	0.131	0.027	0.023	1.000	0.039
c	3120	t	0.248	0.033	0.019	0.162	1.000
a	3146	c	0.311	0.098	1.000	0.023	0.041
c	3172	a	0.310	1.000	0.013	0.105	0.127
a	3198	t	0.248	0.033	0.019	0.162	1.000
t	3224	a	0.310	1.000	0.013	0.105	0.127
g	3250	c	0.311	0.098	1.000	0.023	0.041
c	3276	t	0.248	0.033	0.019	0.162	1.000
a	3302	c	0.311	0.098	1.000	0.023	0.041
a	3328	c	0.311	0.098	1.000	0.023	0.041
g	3354	t	0.248	0.033	0.019	0.162	1.000
c	3380	g	0.131	0.027	0.023	1.000	0.039
a	3406	c	0.311	0.098	1.000	0.023	0.041
t	3432	t	0.248	0.033	0.019	0.162	1.000
c	3458	t	0.248	0.033	0.019	0.162	1.000
c	3484	c	0.311	0.098	1.000	0.023	0.041
c	3510	t	0.248	0.033	0.019	0.162	1.000
c	3536	t	0.248	0.033	0.019	0.162	1.000

g	3562	a	0.310	1.000	0.013	0.105	0.127
t	3588	t	0.248	0.033	0.019	0.162	1.000
t	3614	c	0.311	0.098	1.000	0.023	0.041
c	3640	a	0.310	1.000	0.013	0.105	0.127
c	3666	t	0.248	0.033	0.019	0.162	1.000
a	3692	g	0.131	0.027	0.023	1.000	0.039
g	3718	t	0.248	0.033	0.019	0.162	1.000
t	3744	a	0.310	1.000	0.013	0.105	0.127
g	3770	a	0.310	1.000	0.013	0.105	0.127
a	3796	a	0.310	1.000	0.013	0.105	0.127
g	3822	c	0.311	0.098	1.000	0.023	0.041
t	3848	a	0.310	1.000	0.013	0.105	0.127
t	3874	g	0.131	0.027	0.023	1.000	0.039
c	3900	t	0.248	0.033	0.019	0.162	1.000
a	3926	a	0.310	1.000	0.013	0.105	0.127
c	3952	a	0.310	1.000	0.013	0.105	0.127
c	3978	t	0.248	0.033	0.019	0.162	1.000
c	4004	a	0.310	1.000	0.013	0.105	0.127
t	4030	a	0.310	1.000	0.013	0.105	0.127
c	4056	a	0.310	1.000	0.013	0.105	0.127
t	4082	c	0.311	0.098	1.000	0.023	0.041
a	4108	c	0.311	0.098	1.000	0.023	0.041
a	4134	a	0.310	1.000	0.013	0.105	0.127
a	4160	a	0.310	1.000	0.013	0.105	0.127
t	4186	a	0.310	1.000	0.013	0.105	0.127
c	4212	a	0.310	1.000	0.013	0.105	0.127
a	4238	g	0.131	0.027	0.023	1.000	0.039
c	4264	c	0.311	0.098	1.000	0.023	0.041
c	4290	c	0.311	0.098	1.000	0.023	0.041
a	4316	c	0.311	0.098	1.000	0.023	0.041
c	4342	c	0.311	0.098	1.000	0.023	0.041
g	4368	c	0.311	0.098	1.000	0.023	0.041
a	4394	g	0.131	0.027	0.023	1.000	0.039
t	4420	c	0.311	0.098	1.000	0.023	0.041
c	4446	a	0.310	1.000	0.013	0.105	0.127

a	4472	c	0.311	0.098	1.000	0.023	0.041
a	4498	a	0.310	1.000	0.013	0.105	0.127
a	4524	t	0.248	0.033	0.019	0.162	1.000
a	4550	t	0.248	0.033	0.019	0.162	1.000
g	4576	a	0.310	1.000	0.013	0.105	0.127
g	4602	c	0.311	0.098	1.000	0.023	0.041
a	4628	t	0.248	0.033	0.019	0.162	1.000
a	4654	g	0.131	0.027	0.023	1.000	0.039
c	4680	g	0.131	0.027	0.023	1.000	0.039
a	4706	t	0.248	0.033	0.019	0.162	1.000
a	4732	c	0.311	0.098	1.000	0.023	0.041
g	4758	c	0.311	0.098	1.000	0.023	0.041
c	4784	t	0.248	0.033	0.019	0.162	1.000
a	4810	c	0.311	0.098	1.000	0.023	0.041
t	4836	a	0.310	1.000	0.013	0.105	0.127
c	4862	c	0.311	0.098	1.000	0.023	0.041
a	4888	g	0.131	0.027	0.023	1.000	0.039
a	4914	c	0.311	0.098	1.000	0.023	0.041
g	4940	a	0.310	1.000	0.013	0.105	0.127
c	4966	g	0.131	0.027	0.023	1.000	0.039
a	4992	c	0.311	0.098	1.000	0.023	0.041
c	5018	g	0.131	0.027	0.023	1.000	0.039
g	5044	a	0.310	1.000	0.013	0.105	0.127
c	5070	t	0.248	0.033	0.019	0.162	1.000
a	5096	g	0.131	0.027	0.023	1.000	0.039
g	5122	a	0.310	1.000	0.013	0.105	0.127
c	5148	c	0.311	0.098	1.000	0.023	0.041
a	5174	g	0.131	0.027	0.023	1.000	0.039
a	5200	g	0.131	0.027	0.023	1.000	0.039
t	5226	c	0.311	0.098	1.000	0.023	0.041
g	5252	g	0.131	0.027	0.023	1.000	0.039
c	5278	t	0.248	0.033	0.019	0.162	1.000
a	5304	a	0.310	1.000	0.013	0.105	0.127
g	5330	t	0.248	0.033	0.019	0.162	1.000
c	5356	t	0.248	0.033	0.019	0.162	1.000

t	5382	a	0.310	1.000	0.013	0.105	0.127
c	5408	c	0.311	0.098	1.000	0.023	0.041
a	5434	c	0.311	0.098	1.000	0.023	0.041
a	5460	g	0.131	0.027	0.023	1.000	0.039
a	5486	a	0.310	1.000	0.013	0.105	0.127
a	5512	c	0.311	0.098	1.000	0.023	0.041
c	5538	a	0.310	1.000	0.013	0.105	0.127
g	5564	t	0.248	0.033	0.019	0.162	1.000
c	5590	a	0.310	1.000	0.013	0.105	0.127
t	5616	c	0.311	0.098	1.000	0.023	0.041
t	5642	t	0.248	0.033	0.019	0.162	1.000
a	5668	c	0.311	0.098	1.000	0.023	0.041
g	5694	a	0.310	1.000	0.013	0.105	0.127
c	5720	c	0.311	0.098	1.000	0.023	0.041
c	5746	c	0.311	0.098	1.000	0.023	0.041
t	5772	c	0.311	0.098	1.000	0.023	0.041
a	5798	c	0.311	0.098	1.000	0.023	0.041
g	5824	a	0.310	1.000	0.013	0.105	0.127
c	5850	c	0.311	0.098	1.000	0.023	0.041
c	5876	c	0.311	0.098	1.000	0.023	0.041
a	5902	a	0.310	1.000	0.013	0.105	0.127
c	5928	c	0.311	0.098	1.000	0.023	0.041
a	5954	t	0.248	0.033	0.019	0.162	1.000
c	5980	a	0.310	1.000	0.013	0.105	0.127
c	6006	a	0.310	1.000	0.013	0.105	0.127
c	6032	t	0.248	0.033	0.019	0.162	1.000
c	6058	g	0.131	0.027	0.023	1.000	0.039
c	6084	c	0.311	0.098	1.000	0.023	0.041
a	6110	a	0.310	1.000	0.013	0.105	0.127
c	6136	g	0.131	0.027	0.023	1.000	0.039
g	6162	g	0.131	0.027	0.023	1.000	0.039
g	6188	t	0.248	0.033	0.019	0.162	1.000
g	6214	g	0.131	0.027	0.023	1.000	0.039
a	6240	a	0.310	1.000	0.013	0.105	0.127
a	6266	t	0.248	0.033	0.019	0.162	1.000

a	6292	t	0.248	0.033	0.019	0.162	1.000
c	6318	a	0.310	1.000	0.013	0.105	0.127
a	6344	t	0.248	0.033	0.019	0.162	1.000
g	6370	c	0.311	0.098	1.000	0.023	0.041
c	6396	c	0.311	0.098	1.000	0.023	0.041
a	6422	a	0.310	1.000	0.013	0.105	0.127
g	6448	g	0.131	0.027	0.023	1.000	0.039
t	6474	a	0.310	1.000	0.013	0.105	0.127
g	6500	t	0.248	0.033	0.019	0.162	1.000
a	6526	c	0.311	0.098	1.000	0.023	0.041
t	6552	a	0.310	1.000	0.013	0.105	0.127
t	6578	t	0.248	0.033	0.019	0.162	1.000
a	6604	t	0.248	0.033	0.019	0.162	1.000
a	6630	a	0.310	1.000	0.013	0.105	0.127
c	6656	a	0.310	1.000	0.013	0.105	0.127
c	6682	a	0.310	1.000	0.013	0.105	0.127
t	6708	c	0.311	0.098	1.000	0.023	0.041
t	6734	t	0.248	0.033	0.019	0.162	1.000
t	6760	g	0.131	0.027	0.023	1.000	0.039
a	6786	a	0.310	1.000	0.013	0.105	0.127
g	6812	g	0.131	0.027	0.023	1.000	0.039
c	6838	c	0.311	0.098	1.000	0.023	0.041
a	6864	a	0.310	1.000	0.013	0.105	0.127
a	6890	c	0.311	0.098	1.000	0.023	0.041
t	6916	c	0.311	0.098	1.000	0.023	0.041
a	6942	c	0.311	0.098	1.000	0.023	0.041
a	6968	a	0.310	1.000	0.013	0.105	0.127
a	6994	g	0.131	0.027	0.023	1.000	0.039
c	7020	c	0.311	0.098	1.000	0.023	0.041
g	7046	c	0.311	0.098	1.000	0.023	0.041
a	7072	t	0.248	0.033	0.019	0.162	1.000
a	7098	a	0.310	1.000	0.013	0.105	0.127
a	7124	a	0.310	1.000	0.013	0.105	0.127
g	7150	t	0.248	0.033	0.019	0.162	1.000
t	7176	a	0.310	1.000	0.013	0.105	0.127

t	7202	c	0.311	0.098	1.000	0.023	0.041
t	7228	t	0.248	0.033	0.019	0.162	1.000
a	7254	c	0.311	0.098	1.000	0.023	0.041
a	7280	g	0.131	0.027	0.023	1.000	0.039
c	7306	a	0.310	1.000	0.013	0.105	0.127
t	7332	a	0.310	1.000	0.013	0.105	0.127
a	7358	a	0.310	1.000	0.013	0.105	0.127
a	7384	c	0.311	0.098	1.000	0.023	0.041
g	7410	c	0.311	0.098	1.000	0.023	0.041
c	7436	a	0.310	1.000	0.013	0.105	0.127
t	7462	t	0.248	0.033	0.019	0.162	1.000
a	7488	g	0.131	0.027	0.023	1.000	0.039
t	7514	a	0.310	1.000	0.013	0.105	0.127
a	7540	c	0.311	0.098	1.000	0.023	0.041
c	7566	a	0.310	1.000	0.013	0.105	0.127
t	7592	c	0.311	0.098	1.000	0.023	0.041
a	7618	c	0.311	0.098	1.000	0.023	0.041
a	7644	c	0.311	0.098	1.000	0.023	0.041
c	7670	t	0.248	0.033	0.019	0.162	1.000
c	7696	a	0.310	1.000	0.013	0.105	0.127
c	7722	c	0.311	0.098	1.000	0.023	0.041
c	7748	c	0.311	0.098	1.000	0.023	0.041
a	7774	c	0.311	0.098	1.000	0.023	0.041
g	7800	c	0.311	0.098	1.000	0.023	0.041
g	7826	c	0.311	0.098	1.000	0.023	0.041
g	7852	t	0.248	0.033	0.019	0.162	1.000
t	7878	t	0.248	0.033	0.019	0.162	1.000
t	7904	c	0.311	0.098	1.000	0.023	0.041
g	7930	a	0.310	1.000	0.013	0.105	0.127
g	7956	t	0.248	0.033	0.019	0.162	1.000
t	7982	a	0.310	1.000	0.013	0.105	0.127
c	8008	c	0.311	0.098	1.000	0.023	0.041
a	8034	c	0.311	0.098	1.000	0.023	0.041
a	8060	t	0.248	0.033	0.019	0.162	1.000
t	8086	c	0.311	0.098	1.000	0.023	0.041

t	8112	c	0.311	0.098	1.000	0.023	0.041
t	8138	a	0.310	1.000	0.013	0.105	0.127
c	8164	t	0.248	0.033	0.019	0.162	1.000
g	8190	t	0.248	0.033	0.019	0.162	1.000
t	8216	c	0.311	0.098	1.000	0.023	0.041
g	8242	c	0.311	0.098	1.000	0.023	0.041
c	8268	c	0.311	0.098	1.000	0.023	0.041
c	8294	a	0.310	1.000	0.013	0.105	0.127
a	8320	a	0.310	1.000	0.013	0.105	0.127
g	8346	a	0.310	1.000	0.013	0.105	0.127
c	8372	a	0.310	1.000	0.013	0.105	0.127
c	8398	c	0.311	0.098	1.000	0.023	0.041
a	8424	c	0.311	0.098	1.000	0.023	0.041
c	8450	g	0.131	0.027	0.023	1.000	0.039
c	8476	t	0.248	0.033	0.019	0.162	1.000
g	8502	c	0.311	0.098	1.000	0.023	0.041
c	8528	a	0.310	1.000	0.013	0.105	0.127
g	8554	t	0.248	0.033	0.019	0.162	1.000
g	8580	t	0.248	0.033	0.019	0.162	1.000
t	8606	g	0.131	0.027	0.023	1.000	0.039
c	8632	a	0.310	1.000	0.013	0.105	0.127
a	8658	t	0.248	0.033	0.019	0.162	1.000
c	8684	c	0.311	0.098	1.000	0.023	0.041
a	8710	c	0.311	0.098	1.000	0.023	0.041
c	8736	a	0.310	1.000	0.013	0.105	0.127
g	8762	t	0.248	0.033	0.019	0.162	1.000
a	8788	c	0.311	0.098	1.000	0.023	0.041
t	8814	t	0.248	0.033	0.019	0.162	1.000
t	8840	g	0.131	0.027	0.023	1.000	0.039
a	8866	g	0.131	0.027	0.023	1.000	0.039
a	8892	t	0.248	0.033	0.019	0.162	1.000
c	8918	c	0.311	0.098	1.000	0.023	0.041
c	8944	t	0.248	0.033	0.019	0.162	1.000
c	8970	a	0.310	1.000	0.013	0.105	0.127
a	8996	t	0.248	0.033	0.019	0.162	1.000

a	9022	c	0.311	0.098	1.000	0.023	0.041
g	9048	c	0.311	0.098	1.000	0.023	0.041
t	9074	c	0.311	0.098	1.000	0.023	0.041
c	9100	c	0.311	0.098	1.000	0.023	0.041
a	9126	t	0.248	0.033	0.019	0.162	1.000
a	9152	t	0.248	0.033	0.019	0.162	1.000
t	9178	t	0.248	0.033	0.019	0.162	1.000
a	9204	c	0.311	0.098	1.000	0.023	0.041
g	9230	t	0.248	0.033	0.019	0.162	1.000
a	9256	g	0.131	0.027	0.023	1.000	0.039
a	9282	t	0.248	0.033	0.019	0.162	1.000
g	9308	t	0.248	0.033	0.019	0.162	1.000
c	9334	g	0.131	0.027	0.023	1.000	0.039
c	9360	g	0.131	0.027	0.023	1.000	0.039
g	9386	t	0.248	0.033	0.019	0.162	1.000
g	9412	g	0.131	0.027	0.023	1.000	0.039
c	9438	t	0.248	0.033	0.019	0.162	1.000
g	9464	t	0.248	0.033	0.019	0.162	1.000
t	9490	a	0.310	1.000	0.013	0.105	0.127
a	9516	c	0.311	0.098	1.000	0.023	0.041
a	9542	t	0.248	0.033	0.019	0.162	1.000
a	9568	a	0.310	1.000	0.013	0.105	0.127
g	9594	t	0.248	0.033	0.019	0.162	1.000
a	9620	c	0.311	0.098	1.000	0.023	0.041
g	9646	a	0.310	1.000	0.013	0.105	0.127
t	9672	c	0.311	0.098	1.000	0.023	0.041
g	9698	t	0.248	0.033	0.019	0.162	1.000
t	9724	c	0.311	0.098	1.000	0.023	0.041
t	9750	t	0.248	0.033	0.019	0.162	1.000
t	9776	t	0.248	0.033	0.019	0.162	1.000
t	9802	t	0.248	0.033	0.019	0.162	1.000
a	9828	c	0.311	0.098	1.000	0.023	0.041
g	9854	c	0.311	0.098	1.000	0.023	0.041
a	9880	g	0.131	0.027	0.023	1.000	0.039
t	9906	c	0.311	0.098	1.000	0.023	0.041

c	9932	t	0.248	0.033	0.019	0.162	1.000
a	9958	g	0.131	0.027	0.023	1.000	0.039
c	9984	c	0.311	0.098	1.000	0.023	0.041
c	10010	t	0.248	0.033	0.019	0.162	1.000
c	10036	a	0.310	1.000	0.013	0.105	0.127
c	10062	a	0.310	1.000	0.013	0.105	0.127
c	10088	t	0.248	0.033	0.019	0.162	1.000
t	10114	g	0.131	0.027	0.023	1.000	0.039
c	10140	c	0.311	0.098	1.000	0.023	0.041
c	10166	t	0.248	0.033	0.019	0.162	1.000
c	10192	t	0.248	0.033	0.019	0.162	1.000
c	10218	t	0.248	0.033	0.019	0.162	1.000
a	10244	a	0.310	1.000	0.013	0.105	0.127
a	10270	c	0.311	0.098	1.000	0.023	0.041
t	10296	a	0.310	1.000	0.013	0.105	0.127
a	10322	c	0.311	0.098	1.000	0.023	0.041
a	10348	c	0.311	0.098	1.000	0.023	0.041
a	10374	a	0.310	1.000	0.013	0.105	0.127
g	10400	a	0.310	1.000	0.013	0.105	0.127
c	10426	g	0.131	0.027	0.023	1.000	0.039
t	10452	c	0.311	0.098	1.000	0.023	0.041
a	10478	t	0.248	0.033	0.019	0.162	1.000
a	10504	t	0.248	0.033	0.019	0.162	1.000
a	10530	c	0.311	0.098	1.000	0.023	0.041
a	10556	a	0.310	1.000	0.013	0.105	0.127
c	10582	t	0.248	0.033	0.019	0.162	1.000
t	10608	g	0.131	0.027	0.023	1.000	0.039
c	10634	c	0.311	0.098	1.000	0.023	0.041
a	10660	t	0.248	0.033	0.019	0.162	1.000
c	10686	a	0.310	1.000	0.013	0.105	0.127
c	10712	c	0.311	0.098	1.000	0.023	0.041
t	10738	a	0.310	1.000	0.013	0.105	0.127
g	10764	a	0.310	1.000	0.013	0.105	0.127
a	10790	a	0.310	1.000	0.013	0.105	0.127
g	10816	a	0.310	1.000	0.013	0.105	0.127

t	10842	g	0.131	0.027	0.023	1.000	0.039
t	10868	g	0.131	0.027	0.023	1.000	0.039
g	10894	t	0.248	0.033	0.019	0.162	1.000
t	10920	c	0.311	0.098	1.000	0.023	0.041
a	10946	g	0.131	0.027	0.023	1.000	0.039
a	10972	t	0.248	0.033	0.019	0.162	1.000
a	10998	a	0.310	1.000	0.013	0.105	0.127
a	11024	a	0.310	1.000	0.013	0.105	0.127
a	11050	a	0.310	1.000	0.013	0.105	0.127
a	11076	g	0.131	0.027	0.023	1.000	0.039
c	11102	c	0.311	0.098	1.000	0.023	0.041
t	11128	t	0.248	0.033	0.019	0.162	1.000
c	11154	t	0.248	0.033	0.019	0.162	1.000
c	11180	c	0.311	0.098	1.000	0.023	0.041
a	11206	g	0.131	0.027	0.023	1.000	0.039
g	11232	c	0.311	0.098	1.000	0.023	0.041
t	11258	t	0.248	0.033	0.019	0.162	1.000
t	11284	a	0.310	1.000	0.013	0.105	0.127
g	11310	c	0.311	0.098	1.000	0.023	0.041
a	11336	a	0.310	1.000	0.013	0.105	0.127
c	11362	c	0.311	0.098	1.000	0.023	0.041
a	11388	c	0.311	0.098	1.000	0.023	0.041
c	11414	c	0.311	0.098	1.000	0.023	0.041
a	11440	c	0.311	0.098	1.000	0.023	0.041
a	11466	a	0.310	1.000	0.013	0.105	0.127
a	11492	c	0.311	0.098	1.000	0.023	0.041
a	11518	t	0.248	0.033	0.019	0.162	1.000
t	11544	a	0.310	1.000	0.013	0.105	0.127
a	11570	a	0.310	1.000	0.013	0.105	0.127
g	11596	g	0.131	0.027	0.023	1.000	0.039
a	11622	c	0.311	0.098	1.000	0.023	0.041
c	11648	a	0.310	1.000	0.013	0.105	0.127
t	11674	c	0.311	0.098	1.000	0.023	0.041
a	11700	a	0.310	1.000	0.013	0.105	0.127
c	11726	t	0.248	0.033	0.019	0.162	1.000

g	11752	c	0.311	0.098	1.000	0.023	0.041
a	11778	c	0.311	0.098	1.000	0.023	0.041
a	11804	t	0.248	0.033	0.019	0.162	1.000
a	11830	c	0.311	0.098	1.000	0.023	0.041
g	11856	c	0.311	0.098	1.000	0.023	0.041
t	11882	g	0.131	0.027	0.023	1.000	0.039
g	11908	c	0.311	0.098	1.000	0.023	0.041
g	11934	t	0.248	0.033	0.019	0.162	1.000
c	11960	t	0.248	0.033	0.019	0.162	1.000
t	11986	c	0.311	0.098	1.000	0.023	0.041
t	12012	a	0.310	1.000	0.013	0.105	0.127
t	12038	g	0.131	0.027	0.023	1.000	0.039
a	12064	c	0.311	0.098	1.000	0.023	0.041
a	12090	a	0.310	1.000	0.013	0.105	0.127
c	12116	t	0.248	0.033	0.019	0.162	1.000
a	12142	t	0.248	0.033	0.019	0.162	1.000
t	12168	a	0.310	1.000	0.013	0.105	0.127
a	12194	t	0.248	0.033	0.019	0.162	1.000
t	12220	c	0.311	0.098	1.000	0.023	0.041
c	12246	g	0.131	0.027	0.023	1.000	0.039
t	12272	c	0.311	0.098	1.000	0.023	0.041
g	12298	c	0.311	0.098	1.000	0.023	0.041
a	12324	c	0.311	0.098	1.000	0.023	0.041
a	12350	t	0.248	0.033	0.019	0.162	1.000
c	12376	a	0.310	1.000	0.013	0.105	0.127
a	12402	t	0.248	0.033	0.019	0.162	1.000
c	12428	g	0.131	0.027	0.023	1.000	0.039
a	12454	c	0.311	0.098	1.000	0.023	0.041
c	12480	g	0.131	0.027	0.023	1.000	0.039
a	12506	c	0.311	0.098	1.000	0.023	0.041
a	12532	a	0.310	1.000	0.013	0.105	0.127
t	12558	c	0.311	0.098	1.000	0.023	0.041
a	12584	c	0.311	0.098	1.000	0.023	0.041
g	12610	t	0.248	0.033	0.019	0.162	1.000
c	12636	g	0.131	0.027	0.023	1.000	0.039

t	12662	a	0.310	1.000	0.013	0.105	0.127
a	12688	t	0.248	0.033	0.019	0.162	1.000
a	12714	t	0.248	0.033	0.019	0.162	1.000
g	12740	t	0.248	0.033	0.019	0.162	1.000
a	12766	c	0.311	0.098	1.000	0.023	0.041
c	12792	c	0.311	0.098	1.000	0.023	0.041
c	12818	a	0.310	1.000	0.013	0.105	0.127
c	12844	a	0.310	1.000	0.013	0.105	0.127
a	12870	c	0.311	0.098	1.000	0.023	0.041
a	12896	c	0.311	0.098	1.000	0.023	0.041
a	12922	c	0.311	0.098	1.000	0.023	0.041
c	12948	t	0.248	0.033	0.019	0.162	1.000
t	12974	a	0.310	1.000	0.013	0.105	0.127
g	13000	t	0.248	0.033	0.019	0.162	1.000
g	13026	c	0.311	0.098	1.000	0.023	0.041
g	13052	a	0.310	1.000	0.013	0.105	0.127
a	13078	c	0.311	0.098	1.000	0.023	0.041
t	13104	a	0.310	1.000	0.013	0.105	0.127
t	13130	c	0.311	0.098	1.000	0.023	0.041
a	13156	a	0.310	1.000	0.013	0.105	0.127
g	13182	c	0.311	0.098	1.000	0.023	0.041
a	13208	a	0.310	1.000	0.013	0.105	0.127
t	13234	c	0.311	0.098	1.000	0.023	0.041
a	13260	t	0.248	0.033	0.019	0.162	1.000
c	13286	a	0.310	1.000	0.013	0.105	0.127
c	13312	a	0.310	1.000	0.013	0.105	0.127
c	13338	t	0.248	0.033	0.019	0.162	1.000
c	13364	c	0.311	0.098	1.000	0.023	0.041
a	13390	t	0.248	0.033	0.019	0.162	1.000
c	13416	t	0.248	0.033	0.019	0.162	1.000
t	13442	a	0.310	1.000	0.013	0.105	0.127
a	13468	g	0.131	0.027	0.023	1.000	0.039
t	13494	c	0.311	0.098	1.000	0.023	0.041
g	13520	c	0.311	0.098	1.000	0.023	0.041
c	13546	a	0.310	1.000	0.013	0.105	0.127

t	13572	c	0.311	0.098	1.000	0.023	0.041
t	13598	a	0.310	1.000	0.013	0.105	0.127
a	13624	a	0.310	1.000	0.013	0.105	0.127
g	13650	c	0.311	0.098	1.000	0.023	0.041
c	13676	c	0.311	0.098	1.000	0.023	0.041
c	13702	a	0.310	1.000	0.013	0.105	0.127
c	13728	a	0.310	1.000	0.013	0.105	0.127
t	13754	t	0.248	0.033	0.019	0.162	1.000
a	13780	c	0.311	0.098	1.000	0.023	0.041
a	13806	t	0.248	0.033	0.019	0.162	1.000
a	13832	c	0.311	0.098	1.000	0.023	0.041
c	13858	a	0.310	1.000	0.013	0.105	0.127
c	13884	a	0.310	1.000	0.013	0.105	0.127
t	13910	c	0.311	0.098	1.000	0.023	0.041
c	13936	a	0.310	1.000	0.013	0.105	0.127
a	13962	c	0.311	0.098	1.000	0.023	0.041
a	13988	g	0.131	0.027	0.023	1.000	0.039
c	14014	a	0.310	1.000	0.013	0.105	0.127
a	14040	g	0.131	0.027	0.023	1.000	0.039
g	14066	t	0.248	0.033	0.019	0.162	1.000
t	14092	c	0.311	0.098	1.000	0.023	0.041
t	14118	a	0.310	1.000	0.013	0.105	0.127
a	14144	c	0.311	0.098	1.000	0.023	0.041
a	14170	c	0.311	0.098	1.000	0.023	0.041
a	14196	a	0.310	1.000	0.013	0.105	0.127
t	14222	c	0.311	0.098	1.000	0.023	0.041
c	14248	c	0.311	0.098	1.000	0.023	0.041
a	14274	g	0.131	0.027	0.023	1.000	0.039
a	14300	c	0.311	0.098	1.000	0.023	0.041
c	14326	t	0.248	0.033	0.019	0.162	1.000
a	14352	t	0.248	0.033	0.019	0.162	1.000
a	14378	c	0.311	0.098	1.000	0.023	0.041
a	14404	g	0.131	0.027	0.023	1.000	0.039
a	14430	g	0.131	0.027	0.023	1.000	0.039
c	14456	c	0.311	0.098	1.000	0.023	0.041

t	14482	a	0.310	1.000	0.013	0.105	0.127
g	14508	a	0.310	1.000	0.013	0.105	0.127
c	14534	g	0.131	0.027	0.023	1.000	0.039
t	14560	t	0.248	0.033	0.019	0.162	1.000
c	14586	a	0.310	1.000	0.013	0.105	0.127
g	14612	t	0.248	0.033	0.019	0.162	1.000
c	14638	a	0.310	1.000	0.013	0.105	0.127
c	14664	a	0.310	1.000	0.013	0.105	0.127
a	14690	t	0.248	0.033	0.019	0.162	1.000
g	14716	t	0.248	0.033	0.019	0.162	1.000
a	14742	g	0.131	0.027	0.023	1.000	0.039
a	14768	a	0.310	1.000	0.013	0.105	0.127
c	14794	a	0.310	1.000	0.013	0.105	0.127
a	14820	g	0.131	0.027	0.023	1.000	0.039
c	14846	c	0.311	0.098	1.000	0.023	0.041
t	14872	t	0.248	0.033	0.019	0.162	1.000
a	14898	t	0.248	0.033	0.019	0.162	1.000
c	14924	a	0.310	1.000	0.013	0.105	0.127
g	14950	c	0.311	0.098	1.000	0.023	0.041
a	14976	a	0.310	1.000	0.013	0.105	0.127
g	15002	c	0.311	0.098	1.000	0.023	0.041
c	15028	c	0.311	0.098	1.000	0.023	0.041
c	15054	c	0.311	0.098	1.000	0.023	0.041
a	15080	a	0.310	1.000	0.013	0.105	0.127
c	15106	a	0.310	1.000	0.013	0.105	0.127
a	15132	c	0.311	0.098	1.000	0.023	0.041
g	15158	a	0.310	1.000	0.013	0.105	0.127
c	15184	g	0.131	0.027	0.023	1.000	0.039
t	15210	c	0.311	0.098	1.000	0.023	0.041
t	15236	a	0.310	1.000	0.013	0.105	0.127
a	15262	t	0.248	0.033	0.019	0.162	1.000
a	15288	a	0.310	1.000	0.013	0.105	0.127
a	15314	c	0.311	0.098	1.000	0.023	0.041
a	15340	c	0.311	0.098	1.000	0.023	0.041
c	15366	c	0.311	0.098	1.000	0.023	0.041

t	**15392**	**a**	0.310	1.000	0.013	0.105	0.127
c	**15418**	**c**	0.311	0.098	1.000	0.023	0.041
a	**15444**	**a**	0.310	1.000	0.013	0.105	0.127

Appendix C

Predicted normalized intensities with 30 neurons in the hidden layer using 2-grams for the respective nucleotides

CRS ref	Position	CRS ref (26)	2-gram letters	2-gram values	A	C	G	T
c	26	a	aa	0.098	1.000	0.002	0.089	0.118
a	52	c	cg	0.026	0.053	1.000	0.009	0.064
c	78	t	ta	0.084	0.052	0.006	0.170	1.000
a	104	g	gc	0.043	0.028	0.058	1.000	0.024
g	130	c	ca	0.091	0.054	1.000	0.010	0.067
t	156	c	ct	0.089	0.482	0.579	0.072	0.022
t	182	a	ag	0.049	1.000	0.005	0.118	0.254
t	208	c	ca	0.091	0.054	1.000	0.010	0.067
a	234	c	ca	0.091	0.054	1.000	0.010	0.067
t	260	t	ta	0.084	0.052	0.006	0.170	1.000
g	286	c	ct	0.089	0.482	0.579	0.072	0.022
t	312	t	tt	0.061	0.031	0.014	0.111	1.000
a	338	a	at	0.074	1.000	0.001	0.118	0.155
g	364	g	gt	0.025	0.015	0.011	1.000	0.044
c	390	c	cc	0.105	0.148	1.000	0.036	0.030
t	416	g	gt	0.025	0.015	0.011	1.000	0.044
t	442	a	aa	0.098	1.000	0.002	0.089	0.118
a	468	t	ta	0.084	0.052	0.006	0.170	1.000
c	494	c	ca	0.091	0.054	1.000	0.010	0.067
c	520	g	gc	0.043	0.028	0.058	1.000	0.024
t	546	a	at	0.074	1.000	0.001	0.118	0.155
c	572	t	ta	0.084	0.052	0.006	0.170	1.000
c	598	g	ga	0.038	0.036	0.009	1.000	0.044
t	624	a	ag	0.049	1.000	0.005	0.118	0.254
c	650	a	ac	0.089	0.482	0.579	0.072	0.022
a	676	c	ct	0.089	0.482	0.579	0.072	0.022
a	702	c	ca	0.091	0.054	1.000	0.010	0.067
a	728	a	ag	0.049	1.000	0.005	0.118	0.254
g	754	a	ag	0.049	1.000	0.005	0.118	0.254
c	780	c	ca	0.091	0.054	1.000	0.010	0.067

a	806	c	ca	0.091	0.054	1.000	0.010	0.067
a	832	a	ac	0.089	0.482	0.579	0.072	0.022
t	858	a	ag	0.049	1.000	0.005	0.118	0.254
a	884	t	tg	0.030	0.033	0.011	0.156	1.000
c	910	a	ac	0.089	0.482	0.579	0.072	0.022
a	936	c	ac	0.089	0.482	0.579	0.072	0.022
c	962	c	ac	0.089	0.482	0.579	0.072	0.022
t	988	t	tc	0.073	0.018	0.045	0.216	1.000
g	1014	g	gt	0.025	0.015	0.011	1.000	0.044
a	1040	c	ct	0.089	0.482	0.579	0.072	0.022
a	1066	t	tt	0.061	0.031	0.014	0.111	1.000
a	1092	g	gc	0.043	0.028	0.058	1.000	0.024
a	1118	c	cc	0.105	0.148	1.000	0.036	0.030
t	1144	t	tt	0.061	0.031	0.014	0.111	1.000
g	1170	g	gt	0.025	0.015	0.011	1.000	0.044
t	1196	g	gc	0.043	0.028	0.058	1.000	0.024
t	1222	a	aa	0.098	1.000	0.002	0.089	0.118
t	1248	a	aa	0.098	1.000	0.002	0.089	0.118
a	1274	c	ct	0.089	0.482	0.579	0.072	0.022
g	1300	t	ta	0.084	0.052	0.006	0.170	1.000
a	1326	a	ag	0.049	1.000	0.005	0.118	0.254
c	1352	t	ta	0.084	0.052	0.006	0.170	1.000
g	1378	t	tc	0.073	0.018	0.045	0.216	1.000
g	1404	t	tt	0.061	0.031	0.014	0.111	1.000
g	1430	g	ga	0.038	0.036	0.009	1.000	0.044
c	1456	g	ga	0.038	0.036	0.009	1.000	0.044
t	1482	t	tt	0.061	0.031	0.014	0.111	1.000
c	1508	t	tt	0.061	0.031	0.014	0.111	1.000
a	1534	a	ag	0.049	1.000	0.005	0.118	0.254
c	1560	a	aa	0.098	1.000	0.002	0.089	0.118
a	1586	t	tt	0.061	0.031	0.014	0.111	1.000
t	1612	a	ag	0.049	1.000	0.005	0.118	0.254
c	1638	c	ct	0.089	0.482	0.579	0.072	0.022
a	1664	c	ca	0.091	0.054	1.000	0.010	0.067
c	1690	c	cc	0.105	0.148	1.000	0.036	0.030

c	1716	a	ag	0.049	1.000	0.005	0.118	0.254
c	1742	c	ca	0.091	0.054	1.000	0.010	0.067
c	1768	c	cc	0.105	0.148	1.000	0.036	0.030
a	1794	g	ga	0.038	0.036	0.009	1.000	0.044
t	1820	t	tc	0.073	0.018	0.045	0.216	1.000
a	1846	a	ac	0.089	0.482	0.579	0.072	0.022
a	1872	a	at	0.074	1.000	0.001	0.118	0.155
a	1898	g	gg	0.025	0.015	0.011	1.000	0.044
c	1924	g	gt	0.025	0.015	0.011	1.000	0.044
a	1950	t	tc	0.073	0.018	0.045	0.216	1.000
a	1976	c	cc	0.105	0.148	1.000	0.036	0.030
a	2002	g	gt	0.025	0.015	0.011	1.000	0.044
t	2028	a	aa	0.098	1.000	0.002	0.089	0.118
a	2054	t	tg	0.030	0.033	0.011	0.156	1.000
g	2080	g	gc	0.043	0.028	0.058	1.000	0.024
g	2106	a	at	0.074	1.000	0.001	0.118	0.155
t	2132	a	at	0.074	1.000	0.001	0.118	0.155
t	2158	a	at	0.074	1.000	0.001	0.118	0.155
t	2184	c	ca	0.091	0.054	1.000	0.010	0.067
g	2210	a	aa	0.098	1.000	0.002	0.089	0.118
g	2236	t	tc	0.073	0.018	0.045	0.216	1.000
t	2262	c	cc	0.105	0.148	1.000	0.036	0.030
c	2288	g	ga	0.038	0.036	0.009	1.000	0.044
c	2314	a	at	0.074	1.000	0.001	0.118	0.155
t	2340	c	ca	0.091	0.054	1.000	0.010	0.067
a	2366	a	ac	0.089	0.482	0.579	0.072	0.022
g	2392	t	ta	0.084	0.052	0.006	0.170	1.000
c	2418	a	at	0.074	1.000	0.001	0.118	0.155
c	2444	g	gc	0.043	0.028	0.058	1.000	0.024
t	2470	t	tc	0.073	0.018	0.045	0.216	1.000
t	2496	g	ga	0.038	0.036	0.009	1.000	0.044
t	2522	g	gg	0.025	0.015	0.011	1.000	0.044
c	2548	c	cc	0.105	0.148	1.000	0.036	0.030
t	2574	g	gg	0.025	0.015	0.011	1.000	0.044
a	2600	c	cg	0.026	0.053	1.000	0.009	0.064

t	2626	a	at	0.074	1.000	0.001	0.118	0.155
t	2652	g	gg	0.025	0.015	0.011	1.000	0.044
a	2678	t	ta	0.084	0.052	0.006	0.170	1.000
g	2704	a	ag	0.049	1.000	0.005	0.118	0.254
c	2730	a	ac	0.089	0.482	0.579	0.072	0.022
t	2756	c	ct	0.089	0.482	0.579	0.072	0.022
c	2782	a	aa	0.098	1.000	0.002	0.089	0.118
t	2808	g	ga	0.038	0.036	0.009	1.000	0.044
t	2834	c	cg	0.026	0.053	1.000	0.009	0.064
a	2860	c	ct	0.089	0.482	0.579	0.072	0.022
g	2886	a	ac	0.089	0.482	0.579	0.072	0.022
t	2912	c	cc	0.105	0.148	1.000	0.036	0.030
a	2938	c	ca	0.091	0.054	1.000	0.010	0.067
a	2964	t	ta	0.084	0.052	0.006	0.170	1.000
g	2990	t	tg	0.030	0.033	0.011	0.156	1.000
a	3016	c	cc	0.105	0.148	1.000	0.036	0.030
t	3042	t	ta	0.084	0.052	0.006	0.170	1.000
t	3068	g	gc	0.043	0.028	0.058	1.000	0.024
a	3094	g	gt	0.025	0.015	0.011	1.000	0.044
c	3120	t	tg	0.030	0.033	0.011	0.156	1.000
a	3146	c	ca	0.091	0.054	1.000	0.010	0.067
c	3172	a	ag	0.049	1.000	0.005	0.118	0.254
a	3198	t	ta	0.084	0.052	0.006	0.170	1.000
t	3224	a	ac	0.089	0.482	0.579	0.072	0.022
g	3250	c	ct	0.089	0.482	0.579	0.072	0.022
c	3276	t	tg	0.030	0.033	0.011	0.156	1.000
a	3302	c	ct	0.089	0.482	0.579	0.072	0.022
a	3328	c	cg	0.026	0.053	1.000	0.009	0.064
g	3354	t	ta	0.084	0.052	0.006	0.170	1.000
c	3380	g	gc	0.043	0.028	0.058	1.000	0.024
a	3406	c	ca	0.091	0.054	1.000	0.010	0.067
t	3432	t	ta	0.084	0.052	0.006	0.170	1.000
c	3458	t	tt	0.061	0.031	0.014	0.111	1.000
c	3484	c	ct	0.089	0.482	0.579	0.072	0.022
c	3510	t	tt	0.061	0.031	0.014	0.111	1.000

c	3536	t	tc	0.073	0.018	0.045	0.216	1.000
g	3562	a	at	0.074	1.000	0.001	0.118	0.155
t	3588	t	tc	0.073	0.018	0.045	0.216	1.000
t	3614	c	ct	0.089	0.482	0.579	0.072	0.022
c	3640	a	at	0.074	1.000	0.001	0.118	0.155
c	3666	t	tc	0.073	0.018	0.045	0.216	1.000
a	3692	g	ga	0.038	0.036	0.009	1.000	0.044
g	3718	t	tt	0.061	0.031	0.014	0.111	1.000
t	3744	a	aa	0.098	1.000	0.002	0.089	0.118
g	3770	a	aa	0.098	1.000	0.002	0.089	0.118
a	3796	a	aa	0.098	1.000	0.002	0.089	0.118
g	3822	c	ca	0.091	0.054	1.000	0.010	0.067
t	3848	a	ag	0.049	1.000	0.005	0.118	0.254
t	3874	g	gt	0.025	0.015	0.011	1.000	0.044
c	3900	t	ta	0.084	0.052	0.006	0.170	1.000
a	3926	a	ac	0.089	0.482	0.579	0.072	0.022
c	3952	a	at	0.074	1.000	0.001	0.118	0.155
c	3978	t	tt	0.061	0.031	0.014	0.111	1.000
c	4004	a	ac	0.089	0.482	0.579	0.072	0.022
t	4030	a	ac	0.089	0.482	0.579	0.072	0.022
c	4056	a	ag	0.049	1.000	0.005	0.118	0.254
t	4082	c	cg	0.026	0.053	1.000	0.009	0.064
a	4108	c	ct	0.089	0.482	0.579	0.072	0.022
a	4134	a	ac	0.089	0.482	0.579	0.072	0.022
a	4160	a	at	0.074	1.000	0.001	0.118	0.155
t	4186	a	at	0.074	1.000	0.001	0.118	0.155
c	4212	a	ac	0.089	0.482	0.579	0.072	0.022
a	4238	g	ga	0.038	0.036	0.009	1.000	0.044
c	4264	c	cc	0.105	0.148	1.000	0.036	0.030
c	4290	c	ca	0.091	0.054	1.000	0.010	0.067
a	4316	c	ca	0.091	0.054	1.000	0.010	0.067
c	4342	c	ct	0.089	0.482	0.579	0.072	0.022
g	4368	c	ct	0.089	0.482	0.579	0.072	0.022
a	4394	g	gt	0.025	0.015	0.011	1.000	0.044
t	4420	c	ct	0.089	0.482	0.579	0.072	0.022

c	4446	a	aa	0.098	1.000	0.002	0.089	0.118
a	4472	c	ca	0.091	0.054	1.000	0.010	0.067
a	4498	a	ac	0.089	0.482	0.579	0.072	0.022
a	4524	t	ta	0.084	0.052	0.006	0.170	1.000
a	4550	t	ta	0.084	0.052	0.006	0.170	1.000
g	4576	a	ac	0.089	0.482	0.579	0.072	0.022
g	4602	c	ca	0.091	0.054	1.000	0.010	0.067
a	4628	t	tt	0.061	0.031	0.014	0.111	1.000
a	4654	g	gg	0.025	0.015	0.011	1.000	0.044
c	4680	g	gc	0.043	0.028	0.058	1.000	0.024
a	4706	t	tc	0.073	0.018	0.045	0.216	1.000
a	4732	c	cc	0.105	0.148	1.000	0.036	0.030
g	4758	c	cc	0.105	0.148	1.000	0.036	0.030
c	4784	t	ta	0.084	0.052	0.006	0.170	1.000
a	4810	c	ct	0.089	0.482	0.579	0.072	0.022
t	4836	a	aa	0.098	1.000	0.002	0.089	0.118
c	4862	c	cc	0.105	0.148	1.000	0.036	0.030
a	4888	g	gc	0.043	0.028	0.058	1.000	0.024
a	4914	c	ct	0.089	0.482	0.579	0.072	0.022
g	4940	a	ag	0.049	1.000	0.005	0.118	0.254
c	4966	g	ga	0.038	0.036	0.009	1.000	0.044
a	4992	c	ca	0.091	0.054	1.000	0.010	0.067
c	5018	g	gg	0.025	0.015	0.011	1.000	0.044
g	5044	a	ac	0.089	0.482	0.579	0.072	0.022
c	5070	t	ta	0.084	0.052	0.006	0.170	1.000
a	5096	g	gg	0.025	0.015	0.011	1.000	0.044
g	5122	a	aa	0.098	1.000	0.002	0.089	0.118
c	5148	c	ct	0.089	0.482	0.579	0.072	0.022
a	5174	g	ga	0.038	0.036	0.009	1.000	0.044
a	5200	g	gg	0.025	0.015	0.011	1.000	0.044
t	5226	c	ca	0.091	0.054	1.000	0.010	0.067
g	5252	g	gc	0.043	0.028	0.058	1.000	0.024
c	5278	t	tc	0.073	0.018	0.045	0.216	1.000
a	5304	a	ac	0.089	0.482	0.579	0.072	0.022
g	5330	t	tg	0.030	0.033	0.011	0.156	1.000

c	5356	t	tc	0.073	0.018	0.045	0.216	1.000
t	5382	a	ac	0.089	0.482	0.579	0.072	0.022
c	5408	c	ct	0.089	0.482	0.579	0.072	0.022
a	5434	c	ct	0.089	0.482	0.579	0.072	0.022
a	5460	g	gc	0.043	0.028	0.058	1.000	0.024
a	5486	a	ac	0.089	0.482	0.579	0.072	0.022
a	5512	c	ca	0.091	0.054	1.000	0.010	0.067
c	5538	a	ag	0.049	1.000	0.005	0.118	0.254
g	5564	t	tt	0.061	0.031	0.014	0.111	1.000
c	5590	a	at	0.074	1.000	0.001	0.118	0.155
t	5616	c	cc	0.105	0.148	1.000	0.036	0.030
t	5642	t	tc	0.073	0.018	0.045	0.216	1.000
a	5668	c	ct	0.089	0.482	0.579	0.072	0.022
g	5694	a	ag	0.049	1.000	0.005	0.118	0.254
c	5720	c	ct	0.089	0.482	0.579	0.072	0.022
c	5746	c	cc	0.105	0.148	1.000	0.036	0.030
t	5772	c	ct	0.089	0.482	0.579	0.072	0.022
a	5798	c	cc	0.105	0.148	1.000	0.036	0.030
g	5824	a	at	0.074	1.000	0.001	0.118	0.155
c	5850	c	cc	0.105	0.148	1.000	0.036	0.030
c	5876	c	cc	0.105	0.148	1.000	0.036	0.030
a	5902	a	ac	0.089	0.482	0.579	0.072	0.022
c	5928	c	cc	0.105	0.148	1.000	0.036	0.030
a	5954	t	ta	0.084	0.052	0.006	0.170	1.000
c	5980	a	ac	0.089	0.482	0.579	0.072	0.022
c	6006	a	ag	0.049	1.000	0.005	0.118	0.254
c	6032	t	ta	0.084	0.052	0.006	0.170	1.000
c	6058	g	gt	0.025	0.015	0.011	1.000	0.044
c	6084	c	cg	0.026	0.053	1.000	0.009	0.064
a	6110	a	ac	0.089	0.482	0.579	0.072	0.022
c	6136	g	gg	0.025	0.015	0.011	1.000	0.044
g	6162	g	ga	0.038	0.036	0.009	1.000	0.044
g	6188	t	tc	0.073	0.018	0.045	0.216	1.000
g	6214	g	gt	0.025	0.015	0.011	1.000	0.044
a	6240	a	at	0.074	1.000	0.001	0.118	0.155

a	6266	t	tc	0.073	0.018	0.045	0.216	1.000
a	6292	t	tt	0.061	0.031	0.014	0.111	1.000
c	6318	a	aa	0.098	1.000	0.002	0.089	0.118
a	6344	t	tg	0.030	0.033	0.011	0.156	1.000
g	6370	c	ct	0.089	0.482	0.579	0.072	0.022
c	6396	c	ca	0.091	0.054	1.000	0.010	0.067
a	6422	a	ac	0.089	0.482	0.579	0.072	0.022
g	6448	g	gt	0.025	0.015	0.011	1.000	0.044
t	6474	a	at	0.074	1.000	0.001	0.118	0.155
g	6500	t	ta	0.084	0.052	0.006	0.170	1.000
a	6526	c	ct	0.089	0.482	0.579	0.072	0.022
t	6552	a	aa	0.098	1.000	0.002	0.089	0.118
t	6578	t	tc	0.073	0.018	0.045	0.216	1.000
a	6604	t	tt	0.061	0.031	0.014	0.111	1.000
a	6630	a	at	0.074	1.000	0.001	0.118	0.155
c	6656	a	ac	0.089	0.482	0.579	0.072	0.022
c	6682	a	aa	0.098	1.000	0.002	0.089	0.118
t	6708	c	ca	0.091	0.054	1.000	0.010	0.067
t	6734	t	ta	0.084	0.052	0.006	0.170	1.000
t	6760	g	gc	0.043	0.028	0.058	1.000	0.024
a	6786	a	ag	0.049	1.000	0.005	0.118	0.254
g	6812	g	ga	0.038	0.036	0.009	1.000	0.044
c	6838	c	ca	0.091	0.054	1.000	0.010	0.067
a	6864	a	aa	0.098	1.000	0.002	0.089	0.118
a	6890	c	cg	0.026	0.053	1.000	0.009	0.064
t	6916	c	ca	0.091	0.054	1.000	0.010	0.067
a	6942	c	ca	0.091	0.054	1.000	0.010	0.067
a	6968	a	at	0.074	1.000	0.001	0.118	0.155
a	6994	g	gc	0.043	0.028	0.058	1.000	0.024
c	7020	c	ca	0.091	0.054	1.000	0.010	0.067
g	7046	c	cg	0.026	0.053	1.000	0.009	0.064
a	7072	t	tt	0.061	0.031	0.014	0.111	1.000
a	7098	a	at	0.074	1.000	0.001	0.118	0.155
a	7124	a	ag	0.049	1.000	0.005	0.118	0.254
g	7150	t	tc	0.073	0.018	0.045	0.216	1.000

t	7176	a	at	0.074	1.000	0.001	0.118	0.155
t	7202	c	cg	0.026	0.053	1.000	0.009	0.064
t	7228	t	tc	0.073	0.018	0.045	0.216	1.000
a	7254	c	ct	0.089	0.482	0.579	0.072	0.022
a	7280	g	gg	0.025	0.015	0.011	1.000	0.044
c	7306	a	aa	0.098	1.000	0.002	0.089	0.118
t	7332	a	ac	0.089	0.482	0.579	0.072	0.022
a	7358	a	ac	0.089	0.482	0.579	0.072	0.022
a	7384	c	cc	0.105	0.148	1.000	0.036	0.030
g	7410	c	ct	0.089	0.482	0.579	0.072	0.022
c	7436	a	ag	0.049	1.000	0.005	0.118	0.254
t	7462	t	tt	0.061	0.031	0.014	0.111	1.000
a	7488	g	gt	0.025	0.015	0.011	1.000	0.044
t	7514	a	at	0.074	1.000	0.001	0.118	0.155
a	7540	c	cc	0.105	0.148	1.000	0.036	0.030
c	7566	a	ac	0.089	0.482	0.579	0.072	0.022
t	7592	c	cg	0.026	0.053	1.000	0.009	0.064
a	7618	c	ca	0.091	0.054	1.000	0.010	0.067
a	7644	c	ct	0.089	0.482	0.579	0.072	0.022
c	7670	t	tg	0.030	0.033	0.011	0.156	1.000
c	7696	a	ag	0.049	1.000	0.005	0.118	0.254
c	7722	c	cc	0.105	0.148	1.000	0.036	0.030
c	7748	c	cc	0.105	0.148	1.000	0.036	0.030
a	7774	c	ca	0.091	0.054	1.000	0.010	0.067
g	7800	c	ca	0.091	0.054	1.000	0.010	0.067
g	7826	c	ca	0.091	0.054	1.000	0.010	0.067
g	7852	t	ta	0.084	0.052	0.006	0.170	1.000
t	7878	t	tt	0.061	0.031	0.014	0.111	1.000
t	7904	c	ct	0.089	0.482	0.579	0.072	0.022
g	7930	a	at	0.074	1.000	0.001	0.118	0.155
g	7956	t	tg	0.030	0.033	0.011	0.156	1.000
t	7982	a	at	0.074	1.000	0.001	0.118	0.155
c	8008	c	cg	0.026	0.053	1.000	0.009	0.064
a	8034	c	cc	0.105	0.148	1.000	0.036	0.030
a	8060	t	ta	0.084	0.052	0.006	0.170	1.000

t	8086	c	ca	0.091	0.054	1.000	0.010	0.067
t	8112	c	cc	0.105	0.148	1.000	0.036	0.030
t	8138	a	aa	0.098	1.000	0.002	0.089	0.118
c	8164	t	ta	0.084	0.052	0.006	0.170	1.000
g	8190	t	tt	0.061	0.031	0.014	0.111	1.000
t	8216	c	ac	0.089	0.482	0.579	0.072	0.022
g	8242	c	cc	0.105	0.148	1.000	0.036	0.030
c	8268	c	cc	0.105	0.148	1.000	0.036	0.030
c	8294	a	at	0.074	1.000	0.001	0.118	0.155
a	8320	a	aa	0.098	1.000	0.002	0.089	0.118
g	8346	a	aa	0.098	1.000	0.002	0.089	0.118
c	8372	a	at	0.074	1.000	0.001	0.118	0.155
c	8398	c	cc	0.105	0.148	1.000	0.036	0.030
a	8424	c	cc	0.105	0.148	1.000	0.036	0.030
c	8450	g	gc	0.043	0.028	0.058	1.000	0.024
c	8476	t	tg	0.030	0.033	0.011	0.156	1.000
g	8502	c	cc	0.105	0.148	1.000	0.036	0.030
c	8528	a	at	0.074	1.000	0.001	0.118	0.155
g	8554	t	ta	0.084	0.052	0.006	0.170	1.000
g	8580	t	tc	0.073	0.018	0.045	0.216	1.000
t	8606	g	gt	0.025	0.015	0.011	1.000	0.044
c	8632	a	at	0.074	1.000	0.001	0.118	0.155
a	8658	t	tc	0.073	0.018	0.045	0.216	1.000
c	8684	c	cc	0.105	0.148	1.000	0.036	0.030
a	8710	c	ct	0.089	0.482	0.579	0.072	0.022
c	8736	a	at	0.074	1.000	0.001	0.118	0.155
g	8762	t	tc	0.073	0.018	0.045	0.216	1.000
a	8788	c	ct	0.089	0.482	0.579	0.072	0.022
t	8814	t	tg	0.030	0.033	0.011	0.156	1.000
t	8840	g	gc	0.043	0.028	0.058	1.000	0.024
a	8866	g		0.025	0.015	0.011	1.000	0.044
a	8892	t	ta	0.084	0.052	0.006	0.170	1.000
c	8918	c	ca	0.091	0.054	1.000	0.010	0.067
c	8944	t	tc	0.073	0.018	0.045	0.216	1.000
c	8970	a	ag	0.049	1.000	0.005	0.118	0.254

a	8996	t	tc	0.073	0.018	0.045	0.216	1.000
a	9022	c	cc	0.105	0.148	1.000	0.036	0.030
g	9048	c	cg	0.026	0.053	1.000	0.009	0.064
t	9074	c	ct	0.089	0.482	0.579	0.072	0.022
c	9100	c	cg	0.026	0.053	1.000	0.009	0.064
a	9126	t	ta	0.084	0.052	0.006	0.170	1.000
a	9152	t	tt	0.061	0.031	0.014	0.111	1.000
t	9178	t	tt	0.061	0.031	0.014	0.111	1.000
a	9204	c	cg	0.026	0.053	1.000	0.009	0.064
g	9230	t	ta	0.084	0.052	0.006	0.170	1.000
a	9256	g	gt	0.025	0.015	0.011	1.000	0.044
a	9282	t	ta	0.084	0.052	0.006	0.170	1.000
g	9308	t	tt	0.061	0.031	0.014	0.111	1.000
c	9334	g	gg	0.025	0.015	0.011	1.000	0.044
c	9360	g	gc	0.043	0.028	0.058	1.000	0.024
g	9386	t	tt	0.061	0.031	0.014	0.111	1.000
g	9412	g	gg	0.025	0.015	0.011	1.000	0.044
c	9438	t	ta	0.084	0.052	0.006	0.170	1.000
g	9464	t	tt	0.061	0.031	0.014	0.111	1.000
t	9490	a	aa	0.098	1.000	0.002	0.089	0.118
a	9516	c	ca	0.091	0.054	1.000	0.010	0.067
a	9542	t	tt	0.061	0.031	0.014	0.111	1.000
a	9568	a	aa	0.098	1.000	0.002	0.089	0.118
g	9594	t	tt	0.061	0.031	0.014	0.111	1.000
a	9620	c	cc	0.105	0.148	1.000	0.036	0.030
g	9646	a	aa	0.098	1.000	0.002	0.089	0.118
t	9672	c	ct	0.089	0.482	0.579	0.072	0.022
g	9698	t	tc	0.073	0.018	0.045	0.216	1.000
t	9724	c	ca	0.091	0.054	1.000	0.010	0.067
t	9750	t	ta	0.084	0.052	0.006	0.170	1.000
t	9776	t	tc	0.073	0.018	0.045	0.216	1.000
t	9802	t	tg	0.030	0.033	0.011	0.156	1.000
a	9828	c	cg	0.026	0.053	1.000	0.009	0.064
g	9854	c	cg	0.026	0.053	1.000	0.009	0.064
a	9880	g	ga	0.038	0.036	0.009	1.000	0.044

t	9906	c	ca	0.091	0.054	1.000	0.010	0.067
c	9932	t	tt	0.061	0.031	0.014	0.111	1.000
a	9958	g	gt	0.025	0.015	0.011	1.000	0.044
c	9984	c	cc	0.105	0.148	1.000	0.036	0.030
c	10010	t	ta	0.084	0.052	0.006	0.170	1.000
c	10036	a	at	0.074	1.000	0.001	0.118	0.155
c	10062	a	ag	0.049	1.000	0.005	0.118	0.254
c	10088	t	ta	0.084	0.052	0.006	0.170	1.000
t	10114	g	gt	0.025	0.015	0.011	1.000	0.044
c	10140	c	ct	0.089	0.482	0.579	0.072	0.022
c	10166	t	ta	0.084	0.052	0.006	0.170	1.000
c	10192	t	ta	0.084	0.052	0.006	0.170	1.000
c	10218	t	tt	0.061	0.031	0.014	0.111	1.000
a	10244	a	aa	0.098	1.000	0.002	0.089	0.118
a	10270	c	cc	0.105	0.148	1.000	0.036	0.030
t	10296	a	at	0.074	1.000	0.001	0.118	0.155
a	10322	c	ca	0.091	0.054	1.000	0.010	0.067
a	10348	c	ca	0.091	0.054	1.000	0.010	0.067
a	10374	a	ac	0.089	0.482	0.579	0.072	0.022
g	10400	a	ac	0.089	0.482	0.579	0.072	0.022
c	10426	g	gc	0.043	0.028	0.058	1.000	0.024
t	10452	c	ct	0.089	0.482	0.579	0.072	0.022
a	10478	t	ta	0.084	0.052	0.006	0.170	1.000
a	10504	t	ta	0.084	0.052	0.006	0.170	1.000
a	10530	c	ct	0.089	0.482	0.579	0.072	0.022
a	10556	a	aa	0.098	1.000	0.002	0.089	0.118
c	10582	t	tc	0.073	0.018	0.045	0.216	1.000
t	10608	g	ga	0.038	0.036	0.009	1.000	0.044
c	10634	c	cc	0.105	0.148	1.000	0.036	0.030
a	10660	t	tt	0.061	0.031	0.014	0.111	1.000
c	10686	a	ac	0.089	0.482	0.579	0.072	0.022
c	10712	c	ca	0.091	0.054	1.000	0.010	0.067
t	10738	a	aa	0.098	1.000	0.002	0.089	0.118
g	10764	a	aa	0.098	1.000	0.002	0.089	0.118
a	10790	a	ag	0.049	1.000	0.005	0.118	0.254

g	10816	a	ac	0.089	0.482	0.579	0.072	0.022
t	10842	g	gc	0.043	0.028	0.058	1.000	0.024
t	10868	g	gg	0.025	0.015	0.011	1.000	0.044
g	10894	t	ta	0.084	0.052	0.006	0.170	1.000
t	10920	c	cg	0.026	0.053	1.000	0.009	0.064
a	10946	g	ga	0.038	0.036	0.009	1.000	0.044
a	10972	t	tt	0.061	0.031	0.014	0.111	1.000
a	10998	a	at	0.074	1.000	0.001	0.118	0.155
a	11024	a	aa	0.098	1.000	0.002	0.089	0.118
a	11050	a	ac	0.089	0.482	0.579	0.072	0.022
a	11076	g	gt	0.025	0.015	0.011	1.000	0.044
c	11102	c	cc	0.105	0.148	1.000	0.036	0.030
t	11128	t	tt	0.061	0.031	0.014	0.111	1.000
c	11154	t	tc	0.073	0.018	0.045	0.216	1.000
c	11180	c	ct	0.089	0.482	0.579	0.072	0.022
a	11206	g	gc	0.043	0.028	0.058	1.000	0.024
g	11232	c	ca	0.091	0.054	1.000	0.010	0.067
t	11258	t	tt	0.061	0.031	0.014	0.111	1.000
t	11284	a	ac	0.089	0.482	0.579	0.072	0.022
g	11310	c	ac	0.089	0.482	0.579	0.072	0.022
a	11336	a	aa	0.098	1.000	0.002	0.089	0.118
c	11362	c	ct	0.089	0.482	0.579	0.072	0.022
a	11388	c	ct	0.089	0.482	0.579	0.072	0.022
c	11414	c	ca	0.091	0.054	1.000	0.010	0.067
a	11440	c	ca	0.091	0.054	1.000	0.010	0.067
a	11466	a	aa	0.098	1.000	0.002	0.089	0.118
a	11492	c	cc	0.105	0.148	1.000	0.036	0.030
a	11518	t	tt	0.061	0.031	0.014	0.111	1.000
t	11544	a	at	0.074	1.000	0.001	0.118	0.155
a	11570	a	at	0.074	1.000	0.001	0.118	0.155
g	11596	g	gt	0.025	0.015	0.011	1.000	0.044
a	11622	c	cc	0.105	0.148	1.000	0.036	0.030
c	11648	a	ag	0.049	1.000	0.005	0.118	0.254
t	11674	c	ct	0.089	0.482	0.579	0.072	0.022
a	11700	a	aa	0.098	1.000	0.002	0.089	0.118

c	11726	t	ta	0.084	0.052	0.006	0.170	1.000
g	11752	c	cc	0.105	0.148	1.000	0.036	0.030
a	11778	c	ct	0.089	0.482	0.579	0.072	0.022
a	11804	t	tc	0.073	0.018	0.045	0.216	1.000
a	11830	c	cc	0.105	0.148	1.000	0.036	0.030
g	11856	c	ca	0.091	0.054	1.000	0.010	0.067
t	11882	g	gt	0.025	0.015	0.011	1.000	0.044
g	11908	c	ct	0.089	0.482	0.579	0.072	0.022
g	11934	t	ta	0.084	0.052	0.006	0.170	1.000
c	11960	t	tg	0.030	0.033	0.011	0.156	1.000
t	11986	c	cc	0.105	0.148	1.000	0.036	0.030
t	12012	a	ac	0.089	0.482	0.579	0.072	0.022
t	12038	g	gt	0.025	0.015	0.011	1.000	0.044
a	12064	c	ca	0.091	0.054	1.000	0.010	0.067
a	12090	a	ac	0.089	0.482	0.579	0.072	0.022
c	12116	t	tt	0.061	0.031	0.014	0.111	1.000
a	12142	t	ta	0.084	0.052	0.006	0.170	1.000
t	12168	a	ac	0.089	0.482	0.579	0.072	0.022
a	12194	t	tg	0.030	0.033	0.011	0.156	1.000
t	12220	c	ct	0.089	0.482	0.579	0.072	0.022
c	12246	g	ga	0.038	0.036	0.009	1.000	0.044
t	12272	c	ca	0.091	0.054	1.000	0.010	0.067
g	12298	c	cg	0.026	0.053	1.000	0.009	0.064
a	12324	c	ca	0.091	0.054	1.000	0.010	0.067
a	12350	t	tg	0.030	0.033	0.011	0.156	1.000
c	12376	a	aa	0.098	1.000	0.002	0.089	0.118
a	12402	t	ta	0.084	0.052	0.006	0.170	1.000
c	12428	g	gg	0.025	0.015	0.011	1.000	0.044
a	12454	c	cc	0.105	0.148	1.000	0.036	0.030
c	12480	g	gc	0.043	0.028	0.058	1.000	0.024
a	12506	c	ca	0.091	0.054	1.000	0.010	0.067
a	12532	a	aa	0.098	1.000	0.002	0.089	0.118
t	12558	c	cc	0.105	0.148	1.000	0.036	0.030
a	12584	c	ca	0.091	0.054	1.000	0.010	0.067
g	12610	t	ta	0.084	0.052	0.006	0.170	1.000

c	12636	g	gc	0.043	0.028	0.058	1.000	0.024
t	12662	a	aa	0.098	1.000	0.002	0.089	0.118
a	12688	t	tc	0.073	0.018	0.045	0.216	1.000
a	12714	t	tc	0.073	0.018	0.045	0.216	1.000
g	12740	t	tt	0.061	0.031	0.014	0.111	1.000
a	12766	c	ct	0.089	0.482	0.579	0.072	0.022
c	12792	c	cc	0.105	0.148	1.000	0.036	0.030
c	12818	a	aa	0.098	1.000	0.002	0.089	0.118
c	12844	a	ag	0.049	1.000	0.005	0.118	0.254
a	12870	c	ct	0.089	0.482	0.579	0.072	0.022
a	12896	c	ct	0.089	0.482	0.579	0.072	0.022
a	12922	c	ca	0.091	0.054	1.000	0.010	0.067
c	12948	t	tc	0.073	0.018	0.045	0.216	1.000
t	12974	a	ac	0.089	0.482	0.579	0.072	0.022
g	13000	t	tc	0.073	0.018	0.045	0.216	1.000
g	13026	c	cc	0.105	0.148	1.000	0.036	0.030
g	13052	a	ac	0.089	0.482	0.579	0.072	0.022
a	13078	c	ca	0.091	0.054	1.000	0.010	0.067
t	13104	a	ac	0.089	0.482	0.579	0.072	0.022
t	13130	c	cg	0.026	0.053	1.000	0.009	0.064
a	13156	a	ag	0.049	1.000	0.005	0.118	0.254
g	13182	c	cc	0.105	0.148	1.000	0.036	0.030
a	13208	a	at	0.074	1.000	0.001	0.118	0.155
t	13234	c	cc	0.105	0.148	1.000	0.036	0.030
a	13260	t	ta	0.084	0.052	0.006	0.170	1.000
c	13286	a	ac	0.089	0.482	0.579	0.072	0.022
c	13312	a	ac	0.089	0.482	0.579	0.072	0.022
c	13338	t	ta	0.084	0.052	0.006	0.170	1.000
c	13364	c	ca	0.091	0.054	1.000	0.010	0.067
a	13390	t	tc	0.073	0.018	0.045	0.216	1.000
c	13416	t	tc	0.073	0.018	0.045	0.216	1.000
t	13442	a	aa	0.098	1.000	0.002	0.089	0.118
a	13468	g	gc	0.043	0.028	0.058	1.000	0.024
t	13494	c	cc	0.105	0.148	1.000	0.036	0.030
g	13520	c	ct	0.089	0.482	0.579	0.072	0.022

c	13546	a	ac	0.089	0.482	0.579	0.072	0.022
t	13572	c	ca	0.091	0.054	1.000	0.010	0.067
t	13598	a	aa	0.098	1.000	0.002	0.089	0.118
a	13624	a	at	0.074	1.000	0.001	0.118	0.155
g	13650	c	ca	0.091	0.054	1.000	0.010	0.067
c	13676	c	cg	0.026	0.053	1.000	0.009	0.064
c	13702	a	ac	0.089	0.482	0.579	0.072	0.022
c	13728	a	at	0.074	1.000	0.001	0.118	0.155
t	13754	t	ta	0.084	0.052	0.006	0.170	1.000
a	13780	c	ct	0.089	0.482	0.579	0.072	0.022
a	13806	t	tc	0.073	0.018	0.045	0.216	1.000
a	13832	c	ca	0.091	0.054	1.000	0.010	0.067
c	13858	a	at	0.074	1.000	0.001	0.118	0.155
c	13884	a	at	0.074	1.000	0.001	0.118	0.155
t	13910	c	ca	0.091	0.054	1.000	0.010	0.067
c	13936	a	aa	0.098	1.000	0.002	0.089	0.118
a	13962	c	cc	0.105	0.148	1.000	0.036	0.030
a	13988	g	ga	0.038	0.036	0.009	1.000	0.044
c	14014	a	aa	0.098	1.000	0.002	0.089	0.118
a	14040	g	ga	0.038	0.036	0.009	1.000	0.044
g	14066	t	ta	0.084	0.052	0.006	0.170	1.000
t	14092	c	ca	0.091	0.054	1.000	0.010	0.067
t	14118	a	ac	0.089	0.482	0.579	0.072	0.022
a	14144	c	cc	0.105	0.148	1.000	0.036	0.030
a	14170	c	ca	0.091	0.054	1.000	0.010	0.067
a	14196	a	aa	0.098	1.000	0.002	0.089	0.118
t	14222	c	ct	0.089	0.482	0.579	0.072	0.022
c	14248	c	cc	0.105	0.148	1.000	0.036	0.030
a	14274	g	gg	0.025	0.015	0.011	1.000	0.044
a	14300	c	cc	0.105	0.148	1.000	0.036	0.030
c	14326	t	ta	0.084	0.052	0.006	0.170	1.000
a	14352	t	tc	0.073	0.018	0.045	0.216	1.000
a	14378	c	ca	0.091	0.054	1.000	0.010	0.067
a	14404	g	ga	0.038	0.036	0.009	1.000	0.044
a	14430	g	gg	0.025	0.015	0.011	1.000	0.044

c	14456	c	ct	0.089	0.482	0.579	0.072	0.022
t	14482	a	at	0.074	1.000	0.001	0.118	0.155
g	14508	a	ac	0.089	0.482	0.579	0.072	0.022
c	14534	g	gc	0.043	0.028	0.058	1.000	0.024
t	14560	t	ta	0.084	0.052	0.006	0.170	1.000
c	14586	a	at	0.074	1.000	0.001	0.118	0.155
g	14612	t	ta	0.084	0.052	0.006	0.170	1.000
c	14638	a	ac	0.089	0.482	0.579	0.072	0.022
c	14664	a	at	0.074	1.000	0.001	0.118	0.155
a	14690	t	tc	0.073	0.018	0.045	0.216	1.000
g	14716	t	tt	0.061	0.031	0.014	0.111	1.000
a	14742	g	gc	0.043	0.028	0.058	1.000	0.024
a	14768	a	at	0.074	1.000	0.001	0.118	0.155
c	14794	a	ac	0.089	0.482	0.579	0.072	0.022
a	14820	g	ga	0.038	0.036	0.009	1.000	0.044
c	14846	c	ca	0.091	0.054	1.000	0.010	0.067
t	14872	t	tt	0.061	0.031	0.014	0.111	1.000
a	14898	t	tt	0.061	0.031	0.014	0.111	1.000
c	14924	a	ag	0.049	1.000	0.005	0.118	0.254
g	14950	c	cc	0.105	0.148	1.000	0.036	0.030
a	14976	a	at	0.074	1.000	0.001	0.118	0.155
g	15002	c	cg	0.026	0.053	1.000	0.009	0.064
c	15028	c	ct	0.089	0.482	0.579	0.072	0.022
c	15054	c	ct	0.089	0.482	0.579	0.072	0.022
a	15080	a	aa	0.098	1.000	0.002	0.089	0.118
c	15106	a	aa	0.098	1.000	0.002	0.089	0.118
a	15132	c	ct	0.089	0.482	0.579	0.072	0.022
g	15158	a	ag	0.049	1.000	0.005	0.118	0.254
c	15184	g	ga	0.038	0.036	0.009	1.000	0.044
t	15210	c	cc	0.105	0.148	1.000	0.036	0.030
t	15236	a	at	0.074	1.000	0.001	0.118	0.155
a	15262	t	ta	0.084	0.052	0.006	0.170	1.000
a	15288	a	at	0.074	1.000	0.001	0.118	0.155
a	15314	c	ct	0.089	0.482	0.579	0.072	0.022
a	15340	c	cc	0.105	0.148	1.000	0.036	0.030

c	15366	c	cc	0.105	0.148	1.000	0.036	0.030
t	15392	a	aa	0.098	1.000	0.002	0.089	0.118
c	15418	c	cc	0.105	0.148	1.000	0.036	0.030
a	15444	a	at	0.074	1.000	0.001	0.118	0.155

Appendix D

Predicted normalized intensities with 40 neurons in the hidden layer using 1-2-gram composition for the nucleotides

CRS ref	Position	CRS ref (26)	1-gram values	2-gram letters	2-gram values	A	C	G	T
c	26	a	0.310	aa	0.098	1.000	0.002	0.089	0.129
a	52	c	0.311	cg	0.026	0.046	1.000	0.011	0.048
c	78	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
a	104	g	0.131	gc	0.043	0.023	0.044	1.000	0.023
g	130	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
t	156	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
t	182	a	0.310	ag	0.049	1.000	0.014	0.106	0.267
t	208	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
a	234	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
t	260	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
g	286	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
t	312	t	0.248	tt	0.061	0.032	0.014	0.103	1.000
a	338	a	0.310	at	0.074	1.000	0.000	0.124	0.168
g	364	g	0.131	gt	0.025	0.021	0.011	1.000	0.048
c	390	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
t	416	g	0.131	gt	0.025	0.021	0.011	1.000	0.048
t	442	a	0.310	aa	0.098	1.000	0.002	0.089	0.129
a	468	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
c	494	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
c	520	g	0.131	gc	0.043	0.023	0.044	1.000	0.023
t	546	a	0.310	at	0.074	1.000	0.000	0.124	0.168
c	572	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
c	598	g	0.131	ga	0.038	0.037	0.016	1.000	0.035
t	624	a	0.310	ag	0.049	1.000	0.014	0.106	0.267
c	650	a	0.310	ac	0.089	1.000	0.022	0.121	0.047
a	676	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
a	702	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
a	728	a	0.310	ag	0.049	1.000	0.014	0.106	0.267
g	754	a	0.310	ag	0.049	1.000	0.014	0.106	0.267
c	780	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
a	806	c	0.311	ca	0.091	0.061	1.000	0.006	0.068

a	832	a	0.310	ac	0.089	1.000	0.022	0.121	0.047
t	858	a	0.310	ag	0.049	1.000	0.014	0.106	0.267
a	884	t	0.248	tg	0.030	0.024	0.008	0.136	1.000
c	910	a	0.310	ac	0.089	1.000	0.022	0.121	0.047
a	936	c	0.311	ac	0.089	0.094	1.000	0.034	0.015
c	962	c	0.311	ac	0.089	0.094	1.000	0.034	0.015
t	988	t	0.248	tc	0.073	0.015	0.047	0.213	1.000
g	1014	g	0.131	gt	0.025	0.021	0.011	1.000	0.048
a	1040	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
a	1066	t	0.248	tt	0.061	0.032	0.014	0.103	1.000
a	1092	g	0.131	gc	0.043	0.023	0.044	1.000	0.023
a	1118	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
t	1144	t	0.248	tt	0.061	0.032	0.014	0.103	1.000
g	1170	g	0.131	gt	0.025	0.021	0.011	1.000	0.048
t	1196	g	0.131	gc	0.043	0.023	0.044	1.000	0.023
t	1222	a	0.310	aa	0.098	1.000	0.002	0.089	0.129
t	1248	a	0.310	aa	0.098	1.000	0.002	0.089	0.129
a	1274	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
g	1300	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
a	1326	a	0.310	ag	0.049	1.000	0.014	0.106	0.267
c	1352	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
g	1378	t	0.248	tc	0.073	0.015	0.047	0.213	1.000
g	1404	t	0.248	tt	0.061	0.032	0.014	0.103	1.000
g	1430	g	0.131	ga	0.038	0.037	0.016	1.000	0.035
c	1456	g	0.131	ga	0.038	0.037	0.016	1.000	0.035
t	1482	t	0.248	tt	0.061	0.032	0.014	0.103	1.000
c	1508	t	0.248	tt	0.061	0.032	0.014	0.103	1.000
a	1534	a	0.310	ag	0.049	1.000	0.014	0.106	0.267
c	1560	a	0.310	aa	0.098	1.000	0.002	0.089	0.129
a	1586	t	0.248	tt	0.061	0.032	0.014	0.103	1.000
t	1612	a	0.310	ag	0.049	1.000	0.014	0.106	0.267
c	1638	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
a	1664	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
c	1690	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
c	1716	a	0.310	ag	0.049	1.000	0.014	0.106	0.267

c	1742	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
c	1768	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
a	1794	g	0.131	ga	0.038	0.037	0.016	1.000	0.035
t	1820	t	0.248	tc	0.073	0.015	0.047	0.213	1.000
a	1846	a	0.310	ac	0.089	1.000	0.022	0.121	0.047
a	1872	a	0.310	at	0.074	1.000	0.000	0.124	0.168
a	1898	g	0.131	gg	0.025	0.021	0.011	1.000	0.048
c	1924	g	0.131	gt	0.025	0.021	0.011	1.000	0.048
a	1950	t	0.248	tc	0.073	0.015	0.047	0.213	1.000
a	1976	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
a	2002	g	0.131	gt	0.025	0.021	0.011	1.000	0.048
t	2028	a	0.310	aa	0.098	1.000	0.002	0.089	0.129
a	2054	t	0.248	tg	0.030	0.024	0.008	0.136	1.000
g	2080	g	0.131	gc	0.043	0.023	0.044	1.000	0.023
g	2106	a	0.310	at	0.074	1.000	0.000	0.124	0.168
t	2132	a	0.310	at	0.074	1.000	0.000	0.124	0.168
t	2158	a	0.310	at	0.074	1.000	0.000	0.124	0.168
t	2184	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
g	2210	a	0.310	aa	0.098	1.000	0.002	0.089	0.129
g	2236	t	0.248	tc	0.073	0.015	0.047	0.213	1.000
t	2262	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
c	2288	g	0.131	ga	0.038	0.037	0.016	1.000	0.035
c	2314	a	0.310	at	0.074	1.000	0.000	0.124	0.168
t	2340	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
a	2366	a	0.310	ac	0.089	1.000	0.022	0.121	0.047
g	2392	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
c	2418	a	0.310	at	0.074	1.000	0.000	0.124	0.168
c	2444	g	0.131	gc	0.043	0.023	0.044	1.000	0.023
t	2470	t	0.248	tc	0.073	0.015	0.047	0.213	1.000
t	2496	g	0.131	ga	0.038	0.037	0.016	1.000	0.035
t	2522	g	0.131	gg	0.025	0.021	0.011	1.000	0.048
c	2548	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
t	2574	g	0.131	gg	0.025	0.021	0.011	1.000	0.048
a	2600	c	0.311	cg	0.026	0.046	1.000	0.011	0.048
t	2626	a	0.310	at	0.074	1.000	0.000	0.124	0.168

t	2652	g	0.131	gg	0.025	0.021	0.011	1.000	0.048
a	2678	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
g	2704	a	0.310	ag	0.049	1.000	0.014	0.106	0.267
c	2730	a	0.310	ac	0.089	1.000	0.022	0.121	0.047
t	2756	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
c	2782	a	0.310	aa	0.098	1.000	0.002	0.089	0.129
t	2808	g	0.131	ga	0.038	0.037	0.016	1.000	0.035
t	2834	c	0.311	cg	0.026	0.046	1.000	0.011	0.048
a	2860	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
g	2886	a	0.310	ac	0.089	1.000	0.022	0.121	0.047
t	2912	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
a	2938	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
a	2964	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
g	2990	t	0.248	tg	0.030	0.024	0.008	0.136	1.000
a	3016	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
t	3042	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
t	3068	g	0.131	gc	0.043	0.023	0.044	1.000	0.023
a	3094	g	0.131	gt	0.025	0.021	0.011	1.000	0.048
c	3120	t	0.248	tg	0.030	0.024	0.008	0.136	1.000
a	3146	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
c	3172	a	0.310	ag	0.049	1.000	0.014	0.106	0.267
a	3198	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
t	3224	a	0.310	ac	0.089	1.000	0.022	0.121	0.047
g	3250	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
c	3276	t	0.248	tg	0.030	0.024	0.008	0.136	1.000
a	3302	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
a	3328	c	0.311	cg	0.026	0.046	1.000	0.011	0.048
g	3354	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
c	3380	g	0.131	gc	0.043	0.023	0.044	1.000	0.023
a	3406	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
t	3432	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
c	3458	t	0.248	tt	0.061	0.032	0.014	0.103	1.000
c	3484	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
c	3510	t	0.248	tt	0.061	0.032	0.014	0.103	1.000
c	3536	t	0.248	tc	0.073	0.015	0.047	0.213	1.000

g	3562	a	0.310	at	0.074	1.000	0.000	0.124	0.168
t	3588	t	0.248	tc	0.073	0.015	0.047	0.213	1.000
t	3614	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
c	3640	a	0.310	at	0.074	1.000	0.000	0.124	0.168
c	3666	t	0.248	tc	0.073	0.015	0.047	0.213	1.000
a	3692	g	0.131	ga	0.038	0.037	0.016	1.000	0.035
g	3718	t	0.248	tt	0.061	0.032	0.014	0.103	1.000
t	3744	a	0.310	aa	0.098	1.000	0.002	0.089	0.129
g	3770	a	0.310	aa	0.098	1.000	0.002	0.089	0.129
a	3796	a	0.310	aa	0.098	1.000	0.002	0.089	0.129
g	3822	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
t	3848	a	0.310	ag	0.049	1.000	0.014	0.106	0.267
t	3874	g	0.131	gt	0.025	0.021	0.011	1.000	0.048
c	3900	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
a	3926	a	0.310	ac	0.089	1.000	0.022	0.121	0.047
c	3952	a	0.310	at	0.074	1.000	0.000	0.124	0.168
c	3978	t	0.248	tt	0.061	0.032	0.014	0.103	1.000
c	4004	a	0.310	ac	0.089	1.000	0.022	0.121	0.047
t	4030	a	0.310	ac	0.089	1.000	0.022	0.121	0.047
c	4056	a	0.310	ag	0.049	1.000	0.014	0.106	0.267
t	4082	c	0.311	cg	0.026	0.046	1.000	0.011	0.048
a	4108	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
a	4134	a	0.310	ac	0.089	1.000	0.022	0.121	0.047
a	4160	a	0.310	at	0.074	1.000	0.000	0.124	0.168
t	4186	a	0.310	at	0.074	1.000	0.000	0.124	0.168
c	4212	a	0.310	ac	0.089	1.000	0.022	0.121	0.047
a	4238	g	0.131	ga	0.038	0.037	0.016	1.000	0.035
c	4264	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
c	4290	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
a	4316	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
c	4342	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
g	4368	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
a	4394	g	0.131	gt	0.025	0.021	0.011	1.000	0.048
t	4420	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
c	4446	a	0.310	aa	0.098	1.000	0.002	0.089	0.129

a	4472	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
a	4498	a	0.310	ac	0.089	1.000	0.022	0.121	0.047
a	4524	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
a	4550	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
g	4576	a	0.310	ac	0.089	1.000	0.022	0.121	0.047
g	4602	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
a	4628	t	0.248	tt	0.061	0.032	0.014	0.103	1.000
a	4654	g	0.131	gg	0.025	0.021	0.011	1.000	0.048
c	4680	g	0.131	gc	0.043	0.023	0.044	1.000	0.023
a	4706	t	0.248	tc	0.073	0.015	0.047	0.213	1.000
a	4732	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
g	4758	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
c	4784	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
a	4810	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
t	4836	a	0.310	aa	0.098	1.000	0.002	0.089	0.129
c	4862	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
a	4888	g	0.131	gc	0.043	0.023	0.044	1.000	0.023
a	4914	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
g	4940	a	0.310	ag	0.049	1.000	0.014	0.106	0.267
c	4966	g	0.131	ga	0.038	0.037	0.016	1.000	0.035
a	4992	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
c	5018	g	0.131	gg	0.025	0.021	0.011	1.000	0.048
g	5044	a	0.310	ac	0.089	1.000	0.022	0.121	0.047
c	5070	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
a	5096	g	0.131	gg	0.025	0.021	0.011	1.000	0.048
g	5122	a	0.310	aa	0.098	1.000	0.002	0.089	0.129
c	5148	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
a	5174	g	0.131	ga	0.038	0.037	0.016	1.000	0.035
a	5200	g	0.131	gg	0.025	0.021	0.011	1.000	0.048
t	5226	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
g	5252	g	0.131	gc	0.043	0.023	0.044	1.000	0.023
c	5278	t	0.248	tc	0.073	0.015	0.047	0.213	1.000
a	5304	a	0.310	ac	0.089	1.000	0.022	0.121	0.047
g	5330	t	0.248	tg	0.030	0.024	0.008	0.136	1.000
c	5356	t	0.248	tc	0.073	0.015	0.047	0.213	1.000

t	5382	a	0.310	ac	0.089	1.000	0.022	0.121	0.047
c	5408	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
a	5434	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
a	5460	g	0.131	gc	0.043	0.023	0.044	1.000	0.023
a	5486	a	0.310	ac	0.089	1.000	0.022	0.121	0.047
a	5512	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
c	5538	a	0.310	ag	0.049	1.000	0.014	0.106	0.267
g	5564	t	0.248	tt	0.061	0.032	0.014	0.103	1.000
c	5590	a	0.310	at	0.074	1.000	0.000	0.124	0.168
t	5616	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
t	5642	t	0.248	tc	0.073	0.015	0.047	0.213	1.000
a	5668	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
g	5694	a	0.310	ag	0.049	1.000	0.014	0.106	0.267
c	5720	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
c	5746	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
t	5772	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
a	5798	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
g	5824	a	0.310	at	0.074	1.000	0.000	0.124	0.168
c	5850	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
c	5876	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
a	5902	a	0.310	ac	0.089	1.000	0.022	0.121	0.047
c	5928	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
a	5954	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
c	5980	a	0.310	ac	0.089	1.000	0.022	0.121	0.047
c	6006	a	0.310	ag	0.049	1.000	0.014	0.106	0.267
c	6032	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
c	6058	g	0.131	gt	0.025	0.021	0.011	1.000	0.048
c	6084	c	0.311	cg	0.026	0.046	1.000	0.011	0.048
a	6110	a	0.310	ac	0.089	1.000	0.022	0.121	0.047
c	6136	g	0.131	gg	0.025	0.021	0.011	1.000	0.048
g	6162	g	0.131	ga	0.038	0.037	0.016	1.000	0.035
g	6188	t	0.248	tc	0.073	0.015	0.047	0.213	1.000
g	6214	g	0.131	gt	0.025	0.021	0.011	1.000	0.048
a	6240	a	0.310	at	0.074	1.000	0.000	0.124	0.168
a	6266	t	0.248	tc	0.073	0.015	0.047	0.213	1.000

a	6292	t	0.248	tt	0.061	0.032	0.014	0.103	1.000
c	6318	a	0.310	aa	0.098	1.000	0.002	0.089	0.129
a	6344	t	0.248	tg	0.030	0.024	0.008	0.136	1.000
g	6370	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
c	6396	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
a	6422	a	0.310	ac	0.089	1.000	0.022	0.121	0.047
g	6448	g	0.131	gt	0.025	0.021	0.011	1.000	0.048
t	6474	a	0.310	at	0.074	1.000	0.000	0.124	0.168
g	6500	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
a	6526	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
t	6552	a	0.310	aa	0.098	1.000	0.002	0.089	0.129
t	6578	t	0.248	tc	0.073	0.015	0.047	0.213	1.000
a	6604	t	0.248	tt	0.061	0.032	0.014	0.103	1.000
a	6630	a	0.310	at	0.074	1.000	0.000	0.124	0.168
c	6656	a	0.310	ac	0.089	1.000	0.022	0.121	0.047
c	6682	a	0.310	aa	0.098	1.000	0.002	0.089	0.129
t	6708	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
t	6734	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
t	6760	g	0.131	gc	0.043	0.023	0.044	1.000	0.023
a	6786	a	0.310	ag	0.049	1.000	0.014	0.106	0.267
g	6812	g	0.131	ga	0.038	0.037	0.016	1.000	0.035
c	6838	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
a	6864	a	0.310	aa	0.098	1.000	0.002	0.089	0.129
a	6890	c	0.311	cg	0.026	0.046	1.000	0.011	0.048
t	6916	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
a	6942	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
a	6968	a	0.310	at	0.074	1.000	0.000	0.124	0.168
a	6994	g	0.131	gc	0.043	0.023	0.044	1.000	0.023
c	7020	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
g	7046	c	0.311	cg	0.026	0.046	1.000	0.011	0.048
a	7072	t	0.248	tt	0.061	0.032	0.014	0.103	1.000
a	7098	a	0.310	at	0.074	1.000	0.000	0.124	0.168
a	7124	a	0.310	ag	0.049	1.000	0.014	0.106	0.267
g	7150	t	0.248	tc	0.073	0.015	0.047	0.213	1.000
t	7176	a	0.310	at	0.074	1.000	0.000	0.124	0.168

t	7202	c	0.311	cg	0.026	0.046	1.000	0.011	0.048
t	7228	t	0.248	tc	0.073	0.015	0.047	0.213	1.000
a	7254	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
a	7280	g	0.131	gg	0.025	0.021	0.011	1.000	0.048
c	7306	a	0.310	aa	0.098	1.000	0.002	0.089	0.129
t	7332	a	0.310	ac	0.089	1.000	0.022	0.121	0.047
a	7358	a	0.310	ac	0.089	1.000	0.022	0.121	0.047
a	7384	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
g	7410	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
c	7436	a	0.310	ag	0.049	1.000	0.014	0.106	0.267
t	7462	t	0.248	tt	0.061	0.032	0.014	0.103	1.000
a	7488	g	0.131	gt	0.025	0.021	0.011	1.000	0.048
t	7514	a	0.310	at	0.074	1.000	0.000	0.124	0.168
a	7540	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
c	7566	a	0.310	ac	0.089	1.000	0.022	0.121	0.047
t	7592	c	0.311	cg	0.026	0.046	1.000	0.011	0.048
a	7618	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
a	7644	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
c	7670	t	0.248	tg	0.030	0.024	0.008	0.136	1.000
c	7696	a	0.310	ag	0.049	1.000	0.014	0.106	0.267
c	7722	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
c	7748	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
a	7774	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
g	7800	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
g	7826	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
g	7852	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
t	7878	t	0.248	tt	0.061	0.032	0.014	0.103	1.000
t	7904	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
g	7930	a	0.310	at	0.074	1.000	0.000	0.124	0.168
g	7956	t	0.248	tg	0.030	0.024	0.008	0.136	1.000
t	7982	a	0.310	at	0.074	1.000	0.000	0.124	0.168
c	8008	c	0.311	cg	0.026	0.046	1.000	0.011	0.048
a	8034	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
a	8060	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
t	8086	c	0.311	ca	0.091	0.061	1.000	0.006	0.068

t	8112	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
t	8138	a	0.310	aa	0.098	1.000	0.002	0.089	0.129
c	8164	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
g	8190	t	0.248	tt	0.061	0.032	0.014	0.103	1.000
t	8216	c	0.311	ac	0.089	0.094	1.000	0.034	0.015
g	8242	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
c	8268	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
c	8294	a	0.310	at	0.074	1.000	0.000	0.124	0.168
a	8320	a	0.310	aa	0.098	1.000	0.002	0.089	0.129
g	8346	a	0.310	aa	0.098	1.000	0.002	0.089	0.129
c	8372	a	0.310	at	0.074	1.000	0.000	0.124	0.168
c	8398	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
a	8424	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
c	8450	g	0.131	gc	0.043	0.023	0.044	1.000	0.023
c	8476	t	0.248	tg	0.030	0.024	0.008	0.136	1.000
g	8502	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
c	8528	a	0.310	at	0.074	1.000	0.000	0.124	0.168
g	8554	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
g	8580	t	0.248	tc	0.073	0.015	0.047	0.213	1.000
t	8606	g	0.131	gt	0.025	0.021	0.011	1.000	0.048
c	8632	a	0.310	at	0.074	1.000	0.000	0.124	0.168
a	8658	t	0.248	tc	0.073	0.015	0.047	0.213	1.000
c	8684	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
a	8710	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
c	8736	a	0.310	at	0.074	1.000	0.000	0.124	0.168
g	8762	t	0.248	tc	0.073	0.015	0.047	0.213	1.000
a	8788	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
t	8814	t	0.248	tg	0.030	0.024	0.008	0.136	1.000
t	8840	g	0.131	gc	0.043	0.023	0.044	1.000	0.023
a	8866	g	0.131		0.025	0.021	0.011	1.000	0.048
a	8892	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
c	8918	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
c	8944	t	0.248	tc	0.073	0.015	0.047	0.213	1.000
c	8970	a	0.310	ag	0.049	1.000	0.014	0.106	0.267
a	8996	t	0.248	tc	0.073	0.015	0.047	0.213	1.000

a	9022	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
g	9048	c	0.311	cg	0.026	0.046	1.000	0.011	0.048
t	9074	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
c	9100	c	0.311	cg	0.026	0.046	1.000	0.011	0.048
a	9126	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
a	9152	t	0.248	tt	0.061	0.032	0.014	0.103	1.000
t	9178	t	0.248	tt	0.061	0.032	0.014	0.103	1.000
a	9204	c	0.311	cg	0.026	0.046	1.000	0.011	0.048
g	9230	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
a	9256	g	0.131	gt	0.025	0.021	0.011	1.000	0.048
a	9282	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
g	9308	t	0.248	tt	0.061	0.032	0.014	0.103	1.000
c	9334	g	0.131	gg	0.025	0.021	0.011	1.000	0.048
c	9360	g	0.131	gc	0.043	0.023	0.044	1.000	0.023
g	9386	t	0.248	tt	0.061	0.032	0.014	0.103	1.000
g	9412	g	0.131	gg	0.025	0.021	0.011	1.000	0.048
c	9438	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
g	9464	t	0.248	tt	0.061	0.032	0.014	0.103	1.000
t	9490	a	0.310	aa	0.098	1.000	0.002	0.089	0.129
a	9516	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
a	9542	t	0.248	tt	0.061	0.032	0.014	0.103	1.000
a	9568	a	0.310	aa	0.098	1.000	0.002	0.089	0.129
g	9594	t	0.248	tt	0.061	0.032	0.014	0.103	1.000
a	9620	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
g	9646	a	0.310	aa	0.098	1.000	0.002	0.089	0.129
t	9672	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
g	9698	t	0.248	tc	0.073	0.015	0.047	0.213	1.000
t	9724	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
t	9750	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
t	9776	t	0.248	tc	0.073	0.015	0.047	0.213	1.000
t	9802	t	0.248	tg	0.030	0.024	0.008	0.136	1.000
a	9828	c	0.311	cg	0.026	0.046	1.000	0.011	0.048
g	9854	c	0.311	cg	0.026	0.046	1.000	0.011	0.048
a	9880	g	0.131	ga	0.038	0.037	0.016	1.000	0.035
t	9906	c	0.311	ca	0.091	0.061	1.000	0.006	0.068

c	9932	t	0.248	tt	0.061	0.032	0.014	0.103	1.000
a	9958	g	0.131	gt	0.025	0.021	0.011	1.000	0.048
c	9984	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
c	10010	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
c	10036	a	0.310	at	0.074	1.000	0.000	0.124	0.168
c	10062	a	0.310	ag	0.049	1.000	0.014	0.106	0.267
c	10088	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
t	10114	g	0.131	gt	0.025	0.021	0.011	1.000	0.048
c	10140	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
c	10166	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
c	10192	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
c	10218	t	0.248	tt	0.061	0.032	0.014	0.103	1.000
a	10244	a	0.310	aa	0.098	1.000	0.002	0.089	0.129
a	10270	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
t	10296	a	0.310	at	0.074	1.000	0.000	0.124	0.168
a	10322	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
a	10348	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
a	10374	a	0.310	ac	0.089	1.000	0.022	0.121	0.047
g	10400	a	0.310	ac	0.089	1.000	0.022	0.121	0.047
c	10426	g	0.131	gc	0.043	0.023	0.044	1.000	0.023
t	10452	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
a	10478	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
a	10504	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
a	10530	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
a	10556	a	0.310	aa	0.098	1.000	0.002	0.089	0.129
c	10582	t	0.248	tc	0.073	0.015	0.047	0.213	1.000
t	10608	g	0.131	ga	0.038	0.037	0.016	1.000	0.035
c	10634	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
a	10660	t	0.248	tt	0.061	0.032	0.014	0.103	1.000
c	10686	a	0.310	ac	0.089	1.000	0.022	0.121	0.047
c	10712	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
t	10738	a	0.310	aa	0.098	1.000	0.002	0.089	0.129
g	10764	a	0.310	aa	0.098	1.000	0.002	0.089	0.129
a	10790	a	0.310	ag	0.049	1.000	0.014	0.106	0.267
g	10816	a	0.310	ac	0.089	1.000	0.022	0.121	0.047

t	10842	g	0.131	gc	0.043	0.023	0.044	1.000	0.023
t	10868	g	0.131	gg	0.025	0.021	0.011	1.000	0.048
g	10894	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
t	10920	c	0.311	cg	0.026	0.046	1.000	0.011	0.048
a	10946	g	0.131	ga	0.038	0.037	0.016	1.000	0.035
a	10972	t	0.248	tt	0.061	0.032	0.014	0.103	1.000
a	10998	a	0.310	at	0.074	1.000	0.000	0.124	0.168
a	11024	a	0.310	aa	0.098	1.000	0.002	0.089	0.129
a	11050	a	0.310	ac	0.089	1.000	0.022	0.121	0.047
a	11076	g	0.131	gt	0.025	0.021	0.011	1.000	0.048
c	11102	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
t	11128	t	0.248	tt	0.061	0.032	0.014	0.103	1.000
c	11154	t	0.248	tc	0.073	0.015	0.047	0.213	1.000
c	11180	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
a	11206	g	0.131	gc	0.043	0.023	0.044	1.000	0.023
g	11232	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
t	11258	t	0.248	tt	0.061	0.032	0.014	0.103	1.000
t	11284	a	0.310	ac	0.089	1.000	0.022	0.121	0.047
g	11310	c	0.311	ac	0.089	0.094	1.000	0.034	0.015
a	11336	a	0.310	aa	0.098	1.000	0.002	0.089	0.129
c	11362	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
a	11388	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
c	11414	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
a	11440	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
a	11466	a	0.310	aa	0.098	1.000	0.002	0.089	0.129
a	11492	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
a	11518	t	0.248	tt	0.061	0.032	0.014	0.103	1.000
t	11544	a	0.310	at	0.074	1.000	0.000	0.124	0.168
a	11570	a	0.310	at	0.074	1.000	0.000	0.124	0.168
g	11596	g	0.131	gt	0.025	0.021	0.011	1.000	0.048
a	11622	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
c	11648	a	0.310	ag	0.049	1.000	0.014	0.106	0.267
t	11674	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
a	11700	a	0.310	aa	0.098	1.000	0.002	0.089	0.129
c	11726	t	0.248	ta	0.084	0.064	0.001	0.179	1.000

g	11752	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
a	11778	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
a	11804	t	0.248	tc	0.073	0.015	0.047	0.213	1.000
a	11830	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
g	11856	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
t	11882	g	0.131	gt	0.025	0.021	0.011	1.000	0.048
g	11908	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
g	11934	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
c	11960	t	0.248	tg	0.030	0.024	0.008	0.136	1.000
t	11986	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
t	12012	a	0.310	ac	0.089	1.000	0.022	0.121	0.047
t	12038	g	0.131	gt	0.025	0.021	0.011	1.000	0.048
a	12064	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
a	12090	a	0.310	ac	0.089	1.000	0.022	0.121	0.047
c	12116	t	0.248	tt	0.061	0.032	0.014	0.103	1.000
a	12142	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
t	12168	a	0.310	ac	0.089	1.000	0.022	0.121	0.047
a	12194	t	0.248	tg	0.030	0.024	0.008	0.136	1.000
t	12220	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
c	12246	g	0.131	ga	0.038	0.037	0.016	1.000	0.035
t	12272	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
g	12298	c	0.311	cg	0.026	0.046	1.000	0.011	0.048
a	12324	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
a	12350	t	0.248	tg	0.030	0.024	0.008	0.136	1.000
c	12376	a	0.310	aa	0.098	1.000	0.002	0.089	0.129
a	12402	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
c	12428	g	0.131	gg	0.025	0.021	0.011	1.000	0.048
a	12454	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
c	12480	g	0.131	gc	0.043	0.023	0.044	1.000	0.023
a	12506	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
a	12532	a	0.310	aa	0.098	1.000	0.002	0.089	0.129
t	12558	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
a	12584	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
g	12610	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
c	12636	g	0.131	gc	0.043	0.023	0.044	1.000	0.023

t	12662	a	0.310	aa	0.098	1.000	0.002	0.089	0.129
a	12688	t	0.248	tc	0.073	0.015	0.047	0.213	1.000
a	12714	t	0.248	tc	0.073	0.015	0.047	0.213	1.000
g	12740	t	0.248	tt	0.061	0.032	0.014	0.103	1.000
a	12766	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
c	12792	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
c	12818	a	0.310	aa	0.098	1.000	0.002	0.089	0.129
c	12844	a	0.310	ag	0.049	1.000	0.014	0.106	0.267
a	12870	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
a	12896	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
a	12922	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
c	12948	t	0.248	tc	0.073	0.015	0.047	0.213	1.000
t	12974	a	0.310	ac	0.089	1.000	0.022	0.121	0.047
g	13000	t	0.248	tc	0.073	0.015	0.047	0.213	1.000
g	13026	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
g	13052	a	0.310	ac	0.089	1.000	0.022	0.121	0.047
a	13078	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
t	13104	a	0.310	ac	0.089	1.000	0.022	0.121	0.047
t	13130	c	0.311	cg	0.026	0.046	1.000	0.011	0.048
a	13156	a	0.310	ag	0.049	1.000	0.014	0.106	0.267
g	13182	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
a	13208	a	0.310	at	0.074	1.000	0.000	0.124	0.168
t	13234	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
a	13260	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
c	13286	a	0.310	ac	0.089	1.000	0.022	0.121	0.047
c	13312	a	0.310	ac	0.089	1.000	0.022	0.121	0.047
c	13338	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
c	13364	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
a	13390	t	0.248	tc	0.073	0.015	0.047	0.213	1.000
c	13416	t	0.248	tc	0.073	0.015	0.047	0.213	1.000
t	13442	a	0.310	aa	0.098	1.000	0.002	0.089	0.129
a	13468	g	0.131	gc	0.043	0.023	0.044	1.000	0.023
t	13494	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
g	13520	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
c	13546	a	0.310	ac	0.089	1.000	0.022	0.121	0.047

t	13572	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
t	13598	a	0.310	aa	0.098	1.000	0.002	0.089	0.129
a	13624	a	0.310	at	0.074	1.000	0.000	0.124	0.168
g	13650	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
c	13676	c	0.311	cg	0.026	0.046	1.000	0.011	0.048
c	13702	a	0.310	ac	0.089	1.000	0.022	0.121	0.047
c	13728	a	0.310	at	0.074	1.000	0.000	0.124	0.168
t	13754	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
a	13780	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
a	13806	t	0.248	tc	0.073	0.015	0.047	0.213	1.000
a	13832	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
c	13858	a	0.310	at	0.074	1.000	0.000	0.124	0.168
c	13884	a	0.310	at	0.074	1.000	0.000	0.124	0.168
t	13910	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
c	13936	a	0.310	aa	0.098	1.000	0.002	0.089	0.129
a	13962	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
a	13988	g	0.131	ga	0.038	0.037	0.016	1.000	0.035
c	14014	a	0.310	aa	0.098	1.000	0.002	0.089	0.129
a	14040	g	0.131	ga	0.038	0.037	0.016	1.000	0.035
g	14066	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
t	14092	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
t	14118	a	0.310	ac	0.089	1.000	0.022	0.121	0.047
a	14144	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
a	14170	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
a	14196	a	0.310	aa	0.098	1.000	0.002	0.089	0.129
t	14222	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
c	14248	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
a	14274	g	0.131	gg	0.025	0.021	0.011	1.000	0.048
a	14300	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
c	14326	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
a	14352	t	0.248	tc	0.073	0.015	0.047	0.213	1.000
a	14378	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
a	14404	g	0.131	ga	0.038	0.037	0.016	1.000	0.035
a	14430	g	0.131	gg	0.025	0.021	0.011	1.000	0.048
c	14456	c	0.311	ct	0.089	0.094	1.000	0.034	0.015

t	14482	a	0.310	at	0.074	1.000	0.000	0.124	0.168
g	14508	a	0.310	ac	0.089	1.000	0.022	0.121	0.047
c	14534	g	0.131	gc	0.043	0.023	0.044	1.000	0.023
t	14560	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
c	14586	a	0.310	at	0.074	1.000	0.000	0.124	0.168
g	14612	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
c	14638	a	0.310	ac	0.089	1.000	0.022	0.121	0.047
c	14664	a	0.310	at	0.074	1.000	0.000	0.124	0.168
a	14690	t	0.248	tc	0.073	0.015	0.047	0.213	1.000
g	14716	t	0.248	tt	0.061	0.032	0.014	0.103	1.000
a	14742	g	0.131	gc	0.043	0.023	0.044	1.000	0.023
a	14768	a	0.310	at	0.074	1.000	0.000	0.124	0.168
c	14794	a	0.310	ac	0.089	1.000	0.022	0.121	0.047
a	14820	g	0.131	ga	0.038	0.037	0.016	1.000	0.035
c	14846	c	0.311	ca	0.091	0.061	1.000	0.006	0.068
t	14872	t	0.248	tt	0.061	0.032	0.014	0.103	1.000
a	14898	t	0.248	tt	0.061	0.032	0.014	0.103	1.000
c	14924	a	0.310	ag	0.049	1.000	0.014	0.106	0.267
g	14950	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
a	14976	a	0.310	at	0.074	1.000	0.000	0.124	0.168
g	15002	c	0.311	cg	0.026	0.046	1.000	0.011	0.048
c	15028	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
c	15054	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
a	15080	a	0.310	aa	0.098	1.000	0.002	0.089	0.129
c	15106	a	0.310	aa	0.098	1.000	0.002	0.089	0.129
a	15132	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
g	15158	a	0.310	ag	0.049	1.000	0.014	0.106	0.267
c	15184	g	0.131	ga	0.038	0.037	0.016	1.000	0.035
t	15210	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
t	15236	a	0.310	at	0.074	1.000	0.000	0.124	0.168
a	15262	t	0.248	ta	0.084	0.064	0.001	0.179	1.000
a	15288	a	0.310	at	0.074	1.000	0.000	0.124	0.168
a	15314	c	0.311	ct	0.089	0.094	1.000	0.034	0.015
a	15340	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
c	15366	c	0.311	cc	0.105	0.156	1.000	0.032	0.033

t	15392	a	0.310	aa	0.098	1.000	0.002	0.089	0.129
c	15418	c	0.311	cc	0.105	0.156	1.000	0.032	0.033
a	15444	a	0.310	at	0.074	1.000	0.000	0.124	0.168