Department of Computer Science
George Mason University Technical Reports

4400 University Drive MS#4A5
Fairfax, VA 22030-4444 USA
http://cs.gmu.edu/ 703-993-1530

TAC-ELM: Metagenomic Taxonomic Classification with Extreme Learning Machines

Zeehasham Rasheed zrasheed@gmu.edu
Huzefa Rangwala rangwala@cs.gmu.edu

Technical Report GMU-CS-TR-2010-19

Abstract

Next-generation technologies have allowed researchers to determine the collective genomes of all organisms within specific environments or communities. Varying species abundance, length and complexities within different communities, coupled with discovery of new species makes the problem of taxonomic assignment to short DNA sequence reads extremely challenging.

We have developed a new sequence composition-based taxonomic classifier, TAC-ELM for metagenomic analysis. TAC-ELM uses the framework of extreme learning machines to quickly and accurately learn the weights for a neural network model, with input features consisting of GC content and oligonucleotides.

TAC-ELM is evaluated on two standard metagenomic benchmarks with sequence read lengths reflecting the traditional and current technologies. Our empirical results indicate the strength of the developed approach, which outperforms state-of-the-art taxonomic classifiers in terms of accuracy, training time and implementation complexity. We also perform experiments that evaluate the pervasive case within metagenome analysis, where a species may not have been previously sequenced or discovered and will not exist in the reference genome databases.

1 Introduction

The recent advances in sequencing technologies have allowed researchers to determine the genomes of organisms existing as communities across different environments ranging from sea [20], soil and human body [14]. This collective sequencing of organisms without culturing and cloning each organism individually is known as "metagenomics". As an example, the human body contains one of the most densely populated microbial environments known on earth where over 10^{14} microbial cell interact with human host cells. Sequencing these microbial communities, identifying the microbes and understanding their function is critical for an understanding of disease and normality conditions [18].

Metagenomic sequencing projects attempt at sequencing the genomes of all the microbes within a sample which leads to several challenges related to assembly and identification [7]. Specifically, the coverage needed for accurately assembling the different genomes increases several folds in comparison to a single whole genome sequencing [7]. Complex microbial communities hosted within the communities have species of varying length as well as abundance, and most of them do not have a previously cultured reference genome. The short read sequences (ranging from 35 base pairs (bp) to 500 bp) obtained from the next generation sequencing technologies [16] provides an additional level of challenge to the problem [22].

One of the first steps for analyzing metagenomic samples is to separate and identify the sequence reads in terms of phylogeny or taxonomy. This problem of assigning a label to metagenomic sequence reads is called as taxonomic or phylogenetic classification. Several supervised and unsupervised machine learning based approaches have been developed [3, 7] to classify short sequence reads or assembled contigs into different categories or classes across the taxonomy tree.

In this work, we propose a new taxonomy classification scheme that extracts composition-based features (oligonucleotides and GC content) from the the short sequence reads and develops a neural network-based model. To train the parameters of the model we use an analytical framework, called extreme learning machine (ELM) [8] to learn the parameters of the models. We refer to this approach as TAC-ELM (Taxonomic Classification with Extreme Learning Machines).

Traditionally, for a single-layer feed forward neural network, all the parameters (weights and biases) for the different layers are learned using the gradient descent

algorithm. However, this approach is slow, has a high probability of converging to a local minima, and involves several iterative steps. The ELM scheme [8] overcomes these problems by randomly assigning weights to the input layers and analytically computing the weights for the output layer using a simple generalized inverse operation.

We perform a set of experiments evaluating the classification performance of TAC-ELM on two different datasets. We present results evaluating taxonomic classification performance for short metagenomic reads varying from 100 bp to 1000 bp. We also evaluate the different composition-based features and the parameters associated with the ELM scheme. Our results show good classification accuracy for short (100 bp) as well as large metagenomic sequence reads (900-1000 bp). One of the main challenges of metagenomic analysis, is the presence of species within the sample that have never been sequenced before. As such, using the approach in Phymm [3] we present a set of results that exclude specific species (or clade-levels) while assessing the performance of our classification method.

The results reported in this paper show that TAC-ELM consistently outperforms previously developed state-of-the-art algorithms like Phymm [3], PhyloPythia [12] and a PCA-based classification algorithm [23]. We also show that TAC-ELM produces good classification results for the higher levels of taxonomy i.e., order, class and phylum level whereas the BLAST algorithm performs well for lower levels of taxonomy i.e., family and genus levels. This suggests that combining the results of TAC-ELM with BLAST will produce an accurate and efficient phylogenetic classifier.

The rest of this paper is organized as follows. Section 2 presents a review of existing taxonomic classifiers. Section 3 describes the TAC-ELM algorithm and provides an overview of the extreme machine learning theory. Section 4 provides a description of the experimental setup and Section 5 provides a discussion of the results. Finally, Section 6 concludes this paper along with plans for future work.

2 Related Work

Metagenomics, the sequencing of pooled diverse collection of genomes using the next generation short read technologies makes the process of taxonomic or phylogenetic identification challenging. Different organisms within a community have varying abundance and length which along with different sampling procedures will lead to varying coverage for the different genomes. Several computational approaches have been developed for handling the vast amount of metagenomic data to solve two related problems: (i) binning, and (ii) taxonomic classification.

The "binning" problem is the first step in a metagenomic assembly process and involves separation of short read metagenomic sequence data into subgroups that could be organism-specific. In comparison the "taxonomic classification" problem involves assigning a specific label (i.e., a phylogenetic group label) to sequence reads or assembled contigs [3]. The major distinction between the two problems lies in the fact that in case of the binning problem, the subgroups though distinct from one other remain unlabeled. Machine learning methods that use a unsupervised and supervised framework have been developed for the binning and classification problem.

A traditional approach for phylogenetic classification of microbial communities is to use marker genes (e.g., 16S rRNA sequences) for identification of source organism of a sequence read or fragment [21]. Marker genes are highly conserved and provide accurate identification of the taxonomical class [11]. Such approaches use a homology transfer methodology relying on a reference dataset like RDP [4]. These methods can provide valuable community estimates but are limited to specific reads or contigs (marker genes constitute a small fraction of a metagenomic sequence set) and have low sensitivity due to reliance on an incomplete and taxon-biased reference genome database.

Taxonomic classification methods for metagenomic reads fall into two main categories: (i) comparative approaches and (ii) composition based approaches. Comparative or sequence similarity based approaches rely on the principles of homology for the taxonomic assignment. Such methods align the reads or contigs using BLAST [1], and assign taxonomy based on the best hit to a reference database. MEGAN [9] is a metagenomic analysis and visualization software that make the assignment based on multiple BLAST hits and optimized parameters. Specifically, MEGAN assigns reads to a common ancestor of those BLAST matches that exceed a particular bit score threshold. The MG-RAST [13] web server provides taxonomical and functional annotation by comparative searches performed across multiple reference databases. GAAS [2] is a novel BLAST-based tool that includes genome length normalization along with a similarity weighting for multiple BLAST hits to provide improved classification results. Comparative approaches are accurate if the source organism's genome has been sequenced, and is present in the reference genome databases. Notably, in a recent coral reef study [6], only 12% of reads had matches in a comprehensive microbial BLAST database.

On the other hand composition-based methods have been developed, that extract key sequence features like GC composition and subsequence or *k*-mer frequencies and build supervised classification models using these features.

PhyloPythia [12] uses a support vector machine framework [19] to classify long reads into taxonomical groups using a *k*-mer based kernel function. TETRA [17] corre-

lates the k-mer pattern feature to different taxonomical groups. Phymm [3] trains an interpolated Markov model to characterize variable length subsequences specific to different taxonomical subgroups. In combination with BLAST, Phymm shows improved classification accuracy for short reads of 100 base pair (bp) length. Such Markovian models have been very successful in gene finding algorithms like Glimmer [5]. In [23], the authors use principle component analysis (PCA) to select the best k-mer type features and use a linear classify for the taxonomic assignment of reads averaging 900 base pairs in length.

The focus of our work is to develop a supervised classifier using the extreme learning machine framework, that is quick and works accurately for short DNA sequence reads, a characteristic of today's sequencing technologies. In this study, we compare our classifier referred to as TAC-ELM to the Phymm [3], BLAST [1], PhyloPythia [12] and PCA [23] classifiers.

3 Methods

We developed a metagenomic taxonomic classifier using a single feed forward neural network with parameters learned using a scheme [8], called "Extreme Learning Machines" (ELM).

3.1 Extreme Learning Machines

Extreme Learning Machines (ELM) is an efficient learning scheme that determines the output weights of a single-layer feedforward neural network (SLFN) using an analytical solution instead of the standard gradient descent algorithm [8]. Neural networks have been used to solve classification problems in several domains ranging from computer vision to bioinformatics. Traditionally, for a SLFN, all the parameters (weights and biases) for the different layers need to be tuned/learned and there exists dependency between the different layers. The gradient descent algorithm is slow, has a high chance of converging to a local minima, and to achieve good generalization performance demands several iterative steps. The ELM scheme proposed by Huang et. al. [8] overcomes these problems by randomly assigning weights to the input layers and analytically computing the weights for the output layer using a simple generalized inverse operation. The ELM framework has shown comparable classification performance, improved model representation (less complexity) and faster run times in comparison to support vector machines [19] for the microarray classification problem [24].

Given N distinct training examples (x_i, y_i), where x_i with n features is represented as $[x_{i1}, \ldots, x_{in}]^T \in \mathbb{R}^n$ and the output or target vector y_i is represented as $[y_{i1}, \ldots, y_{im}]^T \in \mathbb{R}^m$, then the SLFN with $\tilde{N}$ hidden neurons, and an activation function $g(x)$ is given by [8]:

$$\sum_{i=1}^{\tilde{N}} \beta_i g(w_i \cdot x_j + b_i) = o_j, j = 1 \ldots N. \quad (1)$$

The weight vector, $w_i = [w_{i1}, \ldots, w_{in}]^T$ connects the i^{th} hidden node and the input neuron, and the vector $\beta_i = [\beta_{i1}, \ldots, \beta_{im}]^T$ connects the i^{th} hidden node and the output node, and

We learn the parameters for a standard SLFN, i.e., β_i, w_i and b_i so that across the N samples we achieve close to zero error given by $\sum_{j=1}^{N} \|o_j - y_j\|$. Huang et. al. [8] expresses the N equations in Equation 1 as:

$$H\beta = Y \quad (2)$$

where,

$$H(w_1, \ldots, w_{\tilde{N}}, b_1, \ldots, b_{\tilde{N}}, x_1, \ldots, x_N) \quad (3)$$

$$= \begin{bmatrix} g(w_1 \cdot x_1 + b_1) & \ldots & g(w_{\tilde{N}} \cdot x_1 + b_{\tilde{N}}) \\ \vdots & \ddots & \vdots \\ g(w_1 \cdot x_N + b_1) & \ldots & g(w_{\tilde{N}} \cdot x_{\tilde{N}} + b_{\tilde{N}}) \end{bmatrix}_{N \times \tilde{N}}$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_{\tilde{N}}^T \end{bmatrix}_{\tilde{N} \times m} \quad Y = \begin{bmatrix} y_1^T \\ \vdots \\ t_N^T \end{bmatrix}_{N \times m} \quad (4)$$

The matrix H is the hidden layer output matrix of the neural network where the i^{th} column of H is the i^{th} hidden node's output vector with respect to the inputs x. The ELM learning algorithm proceeds by choosing an activation function $g(x)$ and the number of hidden nodes/neurons $\tilde{N}$. At the first step, arbitrary weights are assigned to the input weight vectors w_i and bias terms b_i. The matrix H is then computed using Equation 3. The output weights β are computed as $\beta = H^{\dagger}Y$ where $H^{\dagger}$ is the Moore-Penrose generalized inverse of the hidden layer output matrix H.

The Moore-Penrose based solution for β is shown to be one of the least-square solutions of the general linear system $H\beta = Y$, and thus can achieve the smallest training error (and not get stuck in local minima as the gradient descent algorithms). It was also shown that the solution is unique, has the smallest norm of weights and hence, good generalization performance [8].

3.2 Feature Extraction

The input to our TAC-ELM classification algorithm is metagenomic short reads or fragments obtained from the next generation sequencing technologies. From the DNA fragment sequences we extract composition-based features that allows the development of taxonomy classifier. Specifically, we use two different features: (i) GC content and (ii) oligonucleotide frequencies (also called as k-mers).

The GC content feature captures the percentage of nucleotides that are cytosine (C) and guanine (G) within the DNA fragment. The composition of GC content is an indicator of certain taxonomic classes, and is expressed as

$$\frac{G+C}{A+C+G+T} \times 100. \quad (5)$$

We capture composition-based statistics using frequencies of contiguous subsequences. Given a DNA sequence X of length n and a user-supplied parameter k, the k-mer at position i of X $(1 < i < n-k)$ is defined to be the k-length subsequence of X starting at position i. These k-mers are also referred to as oligonucleotides. Given a length k, we count the occurrences of all the k-mers in a DNA fragment. The possible number of k-mers is equal to 4^k and signifies the number of features extracted using this approach. For this study we specifically set the k value to be 3 and 4, referred to as tri-nucleotides and tetra-nucleotides, respectively. We experimented with larger values of k, but our preliminary results did not shown an improvement in the classification performance. We combine the GC content feature with the tri-nucleotide and tetra-nucleotide features within the ELM framework.

3.3 Model Parameters

The ELM framework involves choosing two parameters i.e., the activation function g(x) and the number of hidden nodes/neurons. In this study, we experimented with four activation functions: (i) sine, (ii) exponential, (iii) hyperbolic tangent and (iv) sigmoid. Our results showed that the sigmoid activation function consistently outperformed the other functions for the taxonomical classification function. As such, we report results only for the sigmoid activation function. The sigmoid activation function is defined as

$$g(x) = \frac{1}{1+\exp^{-x}}. \quad (6)$$

The number of hidden neurons is the other parameter that needs to be chosen within the ELM framework. We performed a grid search varying the number of neurons from 100 to 500 in increments of 100 with different features, evaluating the classification performance on a small validation set to select the best model. Having a large number of neurons makes the model complex and prone to over-fitting.

4 Experimental Evaluation

4.1 Dataset Description

We evaluated the performance of TAC-ELM on two previously defined benchmarks for taxonomic classification. The first dataset, referred to as FAMeS [10] is a simulated benchmark created to assess the accuracy of annotation and assembly methods developed for the analysis of metagenomic datasets. This dataset is constructed by combining sequence reads taken from 113 microbial genomes (available at the Integrated Microbial Genomes (IMG) database), sequenced independently. Metagenomes vary considerably in their compositions depending on the environment from which the reads were sampled. As such, the FAMeS dataset provides three distinct samples of varying complexities referred to as simLC (low complexity), simMC (medium complexity) and simHC (high complexity). The simLC dataset has a small number of species with high abundance whereas the simHC dataset has all species with similar abundance level making them hard to identify or separate.

Table 1: Dataset Description and Taxa Distribution.

	FAMeS		Phymm	
	# Classes	**# Test**	**# Classes**	**# Test**
Species	99	15000	539	2870
Genus	64	15000	53	3255
Family	49	15000	48	3335
Order	35	15000	34	4575
Class	16	15000	21	4390
Phylum	9	15000	14	4390

For the FAMeS dataset we perform a 10-fold cross-validation whereas for the Phymm dataset we have specific definitions for the the test and training set from the supplementary website [3]. See text for further details.

In this study we report results only for the most challenging simHC dataset, and provide the other results as part of supplementary information. We filter out reads that have unknown or incorrectly called nucleotides ('X' or 'N'), which leaves us with 15000 metagenomic reads of average length equal to 900 bp (representing Sanger-based reads). Some of the sequence reads are paired-end with clone sizes of either 3KB or 8KB. The sequence reads extracted in the FAMeS dataset can be taxonomically categorized from bottom to top in 99 species, 64 genera, 49 families, 35 orders, 16 classes and 9 phylum.

For comparison purposes, we evaluate the TAC-ELM algorithm on the dataset used for evaluation of the Phymm [3] method. We also follow an identical experimental protocol as the Phymm algorithm. This dataset was extracted from the genome repository available at NCBI RefSeq. Specifically, this benchmark consisted of 1,146 chromosomes and plasmids representing 539 bacterial and archael species. These 539 species were found in 53 possible genera, 48 distinct families, 34 orders, 21 classes and 14 phylum. In this paper, we refer to this benchmark as the Phymm dataset. The specific test sets were downloaded from the supplementary website [1]. The read lengths are varied from 100 bp to 1000 bp to simulate the current trends in sequencing technologies.

Table 1 shows the distribution of the categories (called clades) at each level of the hierarchy along with the

[1] http://www.cbcb.umd.edu/software/phymm/

Table 2: TAC-ELM parameter analysis on Phymm dataset using 100 bp sequence reads.

# Neurons	**Species**	**Genus**	**Family**	**Order**	**Class**	**Phylum**
GC + k-mer 3						
100	**56.1**%	**65.4**%	**67.7**%	**70.2**%	**75.4**%	**82.9**%
200	54.3%	64.1%	66.8%	69.4%	74.5%	81.2%
300	54.1%	63.3%	65.7%	68.8%	73.4%	80.5%
GC + k-mer 4						
100	63.2%	71.1%	71.8%	73.3%	77.2%	82.9%
200	**64.8**%	**72.2**%	**73.4**%	**75.1**%	**79.3**%	**85.2**%
300	64.4%	71.5%	72.1%	73.8%	77.5%	83.7%

The numbers in bold indicate the best classification accuracy for the feature set. In this experiment the same-species matches in test and training sets are masked.

total number of test reads for the different levels for the FAMeS and Phymm datasets. Since the distribution of genomes across the phylogeny varies significantly, there may be some clades that may have very few species or no sister clades. The Phymm dataset accounts for this under-representation of some clades by filtering any test samples so that the each clade had at least 2 children clades within the hierarchy [3]. The FAMeS dataset has same number of test reads across the different levels of hierarchy.

4.2 Experimental Protocol

The TAC-ELM algorithm proposed in this paper aims to classify metagenomic DNA reads of varying lengths into classes across the phylogeny. We essentially solve a multi-class classification problem for each of the different levels in the phylogeny or taxonomy tree. For example, we assess the performance of TAC-ELM algorithm in classifying a DNA read in one of the 9 phylum-level classes for the FAMeS dataset. For the FAMeS dataset we perform a 10-fold cross validation, and for each fold we run the TAC-ELM algorithm 10 times setting the input weight and bias vectors randomly at each iteration. The final classification results reported are averaged across the the ten iterations and ten folds. The average reads lengths for the FAMeS dataset is 900 base pairs and reflects the reads derived from Sanger-based sequencing. The classification accuracy is compared to a linear classifier which uses a PCA algorithm for feature selection [23].

For the Phymm dataset we report taxonomy classification performance across the different levels in the taxonomy tree as done in the FAMeS benchmark. The current generation of sequencing technologies produce sequence reads varying from 35 to 450 base pairs. As such, we vary the read lengths for the Phymm dataset from 100 to 1000 (but show results only for the 100 bp and 1000 bp reads). For different levels of the taxonomy tree we have specific test instances (distribution shown in Table 1) and defined on the Phymm website to account for under-representation within different clades in the taxonomy. Using a metagenomic simulator tool called Metasim [15] we simulate reads for the training and test sets with different read lengths. We repeat these experiments 10 times varying the input weights to the TAC-ELM algorithm and the simulated read sets, reporting the average classification accuracy.

4.2.1 Clade-Level Exclusion Experiment

A common scenario for metagenomic analysis is the discovery of new species that have not been laboratory cultured before and hence are not present in the genome reference databases. There is also a probability that a new clade i.e., a new class at a higher taxonomic level may be discovered in the metagenomic sample.

As discussed in the Phymm method [3] we follow a clade-level exclusion experiment. In this setting, for a specific clade having at least 2 or more children-clades, the reads derived from species belonging to one children-clade is kept aside for the testing phase. The models are then trained on the other children-clades. As an example, for the "family-excluded"-order classification problem, it is ensured that for every read within the test set, all the organisms from the same family as the one in the test set are excluded while training the order-level models. In essence, we are excluding a specific clade from the training sets and determining if we can predict the higher level clade in the taxonomy.

5 Results and Discussion

We perform a set of experiments evaluating the classification performance of TAC-ELM on two different datasets. For the Phymm benchmark, we present results evaluating short metagenomic reads varying from 100 bp to 1000 bp. We also evaluate the different parameters associated with the ELM scheme i.e., the number of hidden neurons and the k-mer size. In this paper we present only a subset of the experiments performed (due to space constraints) and also compare to several state-of-the-art taxonomy classification algorithms.

5.1 TAC-ELM parameters

Table 2 shows the classification performance measured using the K-way accuracy score across the different levels of taxonomy, for 100 bp reads on the Phymm dataset. We experiment with increasing the number of neurons from 100 to 300, and having feature combinations of GC content and tri-nucleotide (k-mer = 3) and GC content and tetra-nucleotide features (k-mer = 4).

We observe that the classification accuracy increases from lower levels of taxonomy i.e., Species-level to higher levels of taxonomy i.e., Phylum-level. This is because of decrease in the number of classes and the class definitions become more general from the lower to the higher levels. Further, the use of tetra-nucleotide

features and GC content features is superior than using the tri-nucleotide features and GC content. The percentage improvement of using the tetra-nucleotide versus tri-nucleotide features is 16% and 3% at the Species-level and Phylum-level, respectively. The best performance is achieved by using 100 hidden neurons and 200 hidden neurons for the tri-nucleotide and tetra-nucleotide features, respectively.

Increasing the *k*-mer size beyond 4, did not show a significant improvement in performance. We also tested different activation functions but the sigmoid function showed the best classification accuracy. We use the combination of tetra-nucleotides and GC content features, with a sigmoid activation function and 200 hidden neurons for the results reported in rest of this paper.

5.2 Clade-level exclusion

The clade-level exclusion experiment represents one of the main challenges of metagenomic analysis, where several species that have never been sequenced before will be found within the metagenomic DNA samples (Section 4.2.1). Table 3 shows the classification performance across different taxonomy levels with different clade-levels being excluded. As an example, in determining the class of species when the order is excluded, all organisms having the same order label as the organism from which the test read was sampled will be excluded from the training set. In this case of "order excluded"-class experiment TAC-ELM achieves an accuracy of 26.9%.

We also compare the results of the clade-level exclusion performance for TAC-ELM (Table 3) to the results achieved by Phymm [3], and shown in Table 4. Comparing the two tables, we notice that TAC-ELM outperforms Phymm across all the different taxonomy levels, and with different clades excluded. We highlight in bold, all entries in Table 3 where TAC-ELM outperforms Phymm by 10% (maximum improvement of 58% for the "genus excluded"-family experiment).

5.3 Comparison to other methods

In Table 5 we present taxonomy classification results for TAC-ELM, Phymm, BLAST, PhymmBL and PhyloPythia across the different levels of hierarchy for 1000 bp reads. We observe the same trend as noticed for the 100 bp reads in Section 5.2, that TAC-ELM outperforms Phymm. TAC-ELM also shows an approximate 2% improvement over BLAST at the phylum, class and order levels (higher taxa-levels). PhyloPythia is a SVM-based classifier and performs far poorly in comparison to the other classifiers. PhymmBL is a combination of Phymm and BLAST and shows the best taxonomy classification performance. This shows that if we combine BLAST results with TAC-ELM, we should expect a similar improvement as obtained for PhymmBL. Note, the results for all classifiers other than TAC-ELM were obtained from the Phymm study [3].

Table 3: Clade-Exclusion Classification Accuracy for TAC-ELM on Phymm dataset (100 bp reads).

	Species	Genus	Family	Order	Class	Phylum
All Matches Allowed	64.8	72.2	73.4	75.1	79.3	85.2
Species Excluded	–	35.2	39.2	43.2	50.8	62.5
Genus Excluded	–	–	**25.5**	**27.8**	38.2	62.1
Family Excluded	–	–	–	**21.3**	**35.1**	**58.5**
Order Excluded	–	–	–	–	**26.9**	**55.4**
Class Excluded	–	–	–	–	–	**49.7**

The numbers in bold show a percentage improvement of greater than 10% for TAC-ELM in comparison to Phymm (Table 4).

Table 6 shows the classification performance across the different taxonomic levels, after 10-fold cross validation of the FAMeS dataset. The average read length for the FAMeS dataset is 900 bp and we report results only for the high complexity dataset. We compare the performance of TAC-ELM to a PCA-based linear classifier [23]. TAC-ELM shows an overall improvement of 5% on classification accuracy and outperforms the PCA method in every level of hierarchy. The improvement in comparison to the PCA-base method for TAC-ELM is 8%, 9%, 6%, 5% and 4% across the phylum, class, order, family and genus levels, respectively.

Table 4: Clade-Exclusion Classification Accuracy for Phymm on Phymm dataset (100 bp reads).

	Species	Genus	Family	Order	Class	Phylum
All Matches Allowed	63.2	70.2	72.5	74.8	78.6	84.8
Species Excluded	–	32.1	36.4	41.2	49.1	60.0
Genus Excluded	–	–	16.1	21.6	35.6	56.7
Family Excluded	–	–	–	13.8	28.0	51.7
Order Excluded	–	–	–	–	21.8	49.1
Class Excluded	–	–	–	–	–	39.6

The results are taken from the Phymm paper [3].

6 Conclusions and Future Work

In this paper, we present TAC-ELM, a metagenomic taxonomic classifier that uses composition-based features within an extreme learning machine framework. We evaluate TAC-ELM on previously established benchmarks with varying complexities, and are able to demonstrate better classification performance for TAC-ELM in comparison to Phymm [3] and PhyloPythia [12]. We assessed TAC-ELM on short metagenomic reads, as produced by the current generation of sequencing technologies. The results suggest the promise of using TAC-ELM for metagenomic samples produced from different environments.

We were also able to show that TAC-ELM produced good classification accuracy, when specific species or clades were excluded from training the models. This reflects the pervasive case in metagenome analysis, where

Table 5: Comparative Performance on Phymm dataset (1000bp with same-species matches masked).

	TAC-ELM	Phymm	BLAST	PhmmBL	PhyloPythia
Genus	71.50%	71.10%	**73.80%**	78.40%	7.10%
Family	78.10%	77.50%	**79.20**%	84.80%	–
Order	**81.66%**	80.60%	80.80%	86.90%	25.10%
Class	**86.35%**	85.40%	84.10%	90.60%	30.80%
Phylum	**90.20%**	89.80%	88.00%	93.80%	50.30%

The results highlighted compare TAC-ELM with BLAST. TAC-ELM consistently outperforms Phymm and PhyloPythia. The results other than TAC-ELM are obtained from the Phymm paper [3].

several of the species have not been sequenced before and are not present in the reference databases. Our results also suggest that combining different classifiers like TAC-ELM with BLAST will lead to an improved classification performance across all levels of the hierarchy.

In the near future, we aim to integrate TAC-ELM with BLAST and also incorporate several more complementary features that will improve the classification performance. We are also working on a method that would allow us to automatically determine the optimal number of hidden neurons needed for the single-layer feedforward neural network within the ELM learning framework.

Table 6: Comparative Performance on FAMeS dataset.

	TAC-ELM	PCA-Linear
Genus	61.43 %	59.05 %
Family	64.74 %	61.80 %
Order	69.08 %	64.89 %
Class	74.56 %	68.62 %
Phylum	81.92 %	75.65 %

The results are reported after 10-fold cross-validation. TAC-ELM consistently outperforms PCA-Linear [23].

7 Acknowledgments

The work is supported by NSF grant IIS 0905117 and bioengineering seed grant awarded to HR.

References

[1] Stephen F. Altschul, Thomas L. Madden, Alejandro A. SchÃffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, September 1997.

[2] Florent E. Angly and et. al. The gaas metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. *PLoS Comput Biol*, 5(12):e1000593, 12 2009.

[3] Arthur Brady and Steven L. Salzberg. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nature methods*, 6(9):673–676, September 2009.

[4] J. R. Cole, Q. Wang, E. Cardenas, J. Fish, B. Chai, R. J. Farris, A. S. Kulam-Syed-Mohideen, D. M. McGarrell, T. Marsh, G. M. Garrity, and J. M. Tiedje. The ribosomal database project: improved alignments and new tools for rrna analysis. *Nucleic Acids Res*, Nov 2008.

[5] A. L. Delcher, D. Harmon, S. Kasif, O. White, and S. L. Salzberg. Improved microbial gene identification with glimmer. *Nucleic Acid Research*, 27(23):4436–4641, 1998.

[6] Elizabeth Dinsdale and et. al. Microbial ecology of four coral atolls in the northern line islands. *PLoS ONE*, 3(2):e1584, 02 2008.

[7] Micah Hamady and Rob Knight. Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome Research*, 19(7):1141–1152, 2009.

[8] Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew. Extreme learning machine: Theory and applications. *Neurocomputing*, 70(1-3):489 – 501, 2006. Neural Networks - Selected Papers from the 7th Brazilian Symposium on Neural Networks (SBRN '04), 7th Brazilian Symposium on Neural Networks.

[9] DH Huson, AF Auch, J Qi, and SC Schuster. Megan analysis of metagenomic data. *Genome Res*, 17(3):377–386, 2007.

[10] M. Konstantinos and et. al. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nature Methods*, 4(6):495–500, April 2007.

[11] Zongzhi Liu, Todd Z. DeSantis, Gary L. Andersen, and Rob Knight. Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucl. Acids Res.*, 36(18):e120–, 2008.

[12] Alice Carolyn McHardy, Hector Garcia Martin, Aristotelis Tsirigos, Philip Hugenholtz, and Isidore Rigoutsos. Accurate phylogenetic classification of variable-length dna fragments. *Nat Meth*, 4(1):63–72, 2007.

[13] F Meyer, D Paarmann, M D'Souza, R Olson, EM Glass, M Kubal, T Paczian, A Rodriguez, R Stevens, A Wilke, J Wilkening, and RA Edwards. The metagenomics rast server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, 9:386, 2008.

[14] Junjie Qin et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59–65, Mar 2010.

[15] Daniel C Richter, Felix Ott, Alexander F Auch, Ramona Schmid, and Daniel H Huson. Metasim: A sequencing simulator for genomics and metagenomics. *PLoS ONE*, 3(10):e3373+, 2008.

[16] Jay Shendure and Hanlee Ji. Next-generation dna sequencing. *Nat Biotechnol*, 26(10):1135–1145, Oct 2008.

[17] H Teeling, J Waldmann, T Lombardot, M Bauer, and FO Glockner. Tetra: a web-service and a standalone program for the analysis and comparison of tetranucleotide usage patterns in dna sequences. *BMC Bioinformatics*, 5:163, 2004.

[18] Peter J Turnbaugh, Ruth E Ley, Micah Hamady, Claire M Fraser-Liggett, Rob Knight, and Jeffrey I Gordon. The human microbiome project. *Nature*, 449(7164):804810, October 2007.

[19] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, 1995.

[20] J.C. Venter, K. Remington, J.F. Heidelberg, A.L. Halpern, D. Rusch, J.A. Eisen, D. Wu, I. Paulsen, K.E. Nelson, W. Nelson, et al. Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, 304(5667):66, 2004.

[21] C. R. Woese and G. E. Fox. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci U S A*, 74(11):5088–5090, 1977.

[22] K. Eric Wommack, Jaysheel Bhavsar, and Jacques Ravel. Metagenomics: Read length matters. *Appl. Environ. Microbiol.*, pages AEM.02181–07, 2008.

[23] Hongwei Wu. Pca-based linear combinations of oligonucleotide frequencies for metagenomic dna fragment binning. In *Computational Intelligence in Bioinformatics and Computational Biology, 2008. CIBCB '08. IEEE Symposium on*, pages 46 –53, 2008.

[24] Runxuan Zhang, Guang-Bin Huang, N. Sundararajan, and P. Saratchandran. Multicategory classification using an extreme learning machine for microarray gene expression cancer diagnosis. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 4:485–495, July 2007.