2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

Study of transductive learning and unsupervised feature construction methods for biological sequence classification

Ana Stanescu
Computing and Information Sciences
Kansas State University,
Manhattan, KS, U.S.A.
Email: anas@ksu.edu

Karthik Tangirala
Computing and Information Sciences
Kansas State University,
Manhattan, KS, U.S.A.
Email: karthikt@ksu.edu

Doina Caragea
Computing and Information Sciences
Kansas State University,
Manhattan, KS, U.S.A.
Email: dcaragea@ksu.edu

***Abstract*—Next Generation Sequencing (NGS) technologies have led to fast and inexpensive production of large amounts of biological sequence data, including nucleotide sequences and derived protein sequences. These fast-increasing volumes of data pose challenges to computational methods for annotation. Machine learning approaches, primarily supervised algorithms, have been widely used to assist with classification tasks in bioinformatics. However, supervised algorithms rely on large amounts of labeled data in order to produce quality predictors. Oftentimes, labeled data is difficult and expensive to acquire in sufficiently large quantities. When only limited amounts of labeled data but considerably larger amounts of unlabeled data are available for a specific annotation problem, semi-supervised learning approaches represent a cost-effective alternative. In this work, we focus on a special case of semi-supervised learning, namely transductive learning, in which the algorithm has access during the training phase to the instances that need to be labeled. Transduction is particularly suitable for biological sequence classification, where the goal is generally to label a given set of unlabeled instances. However, a challenge that needs to be addressed in this context consists of identification of compact sets of informative features. Given the lack of labeled data, standard supervised feature selection methods may result in unreliable features. Therefore, we study recently proposed unsupervised feature construction approaches together with transductive learning. Experimental results on two classification problems, namely cassette exon identification and protein localization, show that the unsupervised features result in better performance than the supervised features.**

I. Introduction

Successful advancements in biotechnology have resulted in powerful high throughput sequencing instruments that can produce biological data (such as DNA sequences and derived protein sequences) rapidly and inexpensively. Annotation can no longer be handled solely by wet-lab experiments and it necessarily requires computational methods for automating the annotation process. Traditionally, supervised machine learning has been successfully used for classification or prediction problems in the field of bioinformatics. Supervised methods, however, require large amounts of labeled data for training in order to induce valuable models, but in many cases obtaining a sufficiently large number of labeled instances is infeasible due to the costs involved. Unlabeled instances are more accessible and usually they are available in much larger quantities than labeled instances. Therefore, semi-supervised machine learning algorithms, which can use unlabeled data in conjunction with the labeled data to produce classifiers that can surpass the predictive power of supervised algorithms, are preferable, and can constitute a cost-effective solution to manual (expert) annotation.

Both supervised [1], [2], [3] and semi-supervised [4], [5] algorithms typically produce a classifier that is used to predict the labels of new unlabeled (test) instances, not encountered before in the learning phase. Sometimes, it is not necessary to produce an inductive model. For example, some bioinformatics problems, such as genome annotation, require the labeling (annotation) of the unlabeled instances and not necessarily the labeling of future unseen instances. We argue that in such cases, transductive learning [6], [7] is a more suitable approach as opposed to semi-supervised learning, which solves a "harder" problem by producing an inductive model. More specifically, a transductive algorithm has access, during the training phase to the instances that need to be labeled, in addition to the originally labeled data. In other words, the unlabeled instances also represent the actual test data.

Any classifier requires a formal representation of the instances, usually given in vectorial form (a vector of features). The predictive quality of the features with respect to the classification problem strongly influences the quality of the model [8]. In order to produce an effective and efficient model, a good representation comprises a compact set of informative (predictive) features that correlate well with the class variable and are independent of each other. Bioinformatics tasks such as cassette exon prediction or protein localization (which we explore in this study) can be formulated as sequence classification problems, for which it is ideal to have biologically known features, *e.g.*, DNA motifs or protein domains, to represent the instances. For example, the problem of classifying exons as alternatively spliced (cassette) or as constitutive has been successfully addressed in the supervised [9], semi-supervised

IEEE/ACM ASONAM 2016, August 18-21, 2016, San Francisco, CA, USA
978-1-5090-2846-7/16/$31.00

[10], and transductive [11] learning frameworks using a representation based on biologically significant features (DNA motifs), such as exonic splicing enhancers and intronic regulatory sequences.

However, biologically significant features are expensive and time-consuming to identify, and may not be readily available for certain problems or for newly sequenced organisms. Oftentimes, practitioners resort to the most straightforward feature construction technique, based on the sliding window approach, in which a window of size k traverses and fragments the sequence into substrings of length k, referred to as k-mers, or k-grams. These unique substrings obtained from the training instances comprise the set of k-mers used to represent the data. Varying k, the size of the window, produces k-mers of various lengths, which is desirable, because it is believed that real motifs have variable length that carry more information as compared to fixed-length motifs. Variable length k-mer representations result in high-dimensional spaces, posing problems to learning algorithms in terms of computational resources and tractability, while many k-mers may not be informative. Feature selection is often used to reduce the dimensionality by discarding features based on feature-class dependency scores. The remaining features exhibit high mutual information with the class. Clearly, the class is used in this approach. In transductive learning, in which only a small number of labeled examples are available, feature selection may produce features that are not informative and miss many features that are informative.

In summary, there are two main challenges that machine learning faces when used for biological sequence classification tasks. The first challenge is posed by the insufficiency of the labeled data, and can potentially be addressed using transductive learning approaches. The second challenge is feature construction, which is a domain-specific task, particularly difficult for bioinformatics, especially in a semi-supervised/transductive learning framework, and can potentially be addressed using unsupervised feature construction approaches.

The aim of this work is to gain insights into transductive learning in relation to unsupervised feature construction for both DNA and protein sequence classification tasks. Our study focuses on the empirical comparison of three popular transductive algorithms and their compatibility with two unsupervised feature construction techniques. We use one large margin classifier, namely Transductive Support Vector Machines (TSVM) [6], and two graph-based approaches, namely Label Propagation (LP) [12] and Modified Adsorption (MAD) [13]. To construct features, we use the approach recently introduced in [14], based on Burrows-Wheeler Transform, and the approach proposed in [15], based on a community-detection algorithm. We evaluate the performance of these methods on two biological problems formulated as classification tasks, identification of cassette exons and prediction of protein localization, for which we use two DNA sequence datasets and four protein sequence datasets, respectively. We compare the results obtained using the unsupervised feature construction approaches with the results obtained using supervised feature selection.

The rest of the paper is organized as follows. A brief review of related work is presented in Section II. Section III (Methods) presents the transductive algorithms and the unsupervised feature generation techniques. The biological problems addressed and the datasets are described in Section IV (Data). The experimental setup is detailed in Section V. We discuss the results of the proposed approaches in Section VI (Results). We conclude the study and propose several future research directions in Section VII.

II. Related Work

Transductive learning algorithms have enjoyed great popularity in many domains that suffer from labeled data insufficiency and have many successful utilizations in various domains. Examples of applications include text classification, natural language processing, sentiment analysis, movie and video recommendation.

In bioinformatics, transductive algorithms, mainly Transductive Support Vector Machines algorithm, have been applied to many prediction problems, including promoter recognition [16] and gene expression classification [17]. In [18], the authors address the classification of proteins into SCP (Structural Classification of Proteins) super-families, using cluster kernels (bagged mismatch and neighborhood mismatch kernels) to utilize unlabeled data and labeled data. Kondratovich *et al.* [19] utilized Transductive Support Vector Machines algorithm for the problem of molecule activity prediction.

Transductive graph-based approaches have also been utilized in bioinformatics. For example, in order to predict functional classes of yeast proteins (a multiclass prediction problem), Shin *et al.* [20] used spectral clustering on a graph created by combining multiple graphs obtained from several independent and complementary sources of information. Yu *et al.* [21] used a graph-based approach to predict yeast protein functions.

The MAD algorithm (which we also use in this work), has been previously applied to gene prioritization [22].

Studies that compare transductive learning algorithms have been developed for sentiment classification. Among others, it is worth mentioning a study by Yong *et al.* [23], who compare MAD and LP on the classification of sentiment polarity in documents from underresourced languages. Their experiments on three domains (hotels, notebooks, and books) revealed that MAD outperformed LP. Our objective is also a comparison of transductive approaches, in relation to unsupervised feature construction methods, for biological sequence classification.

In our previous work [11], we compared these approaches on the problem of cassette exon identification for the *C. elegans* dataset, which we also use in the experiments of this work. Our preliminary study was focused on various kernels and biologically relevant features; the results obtained were in favor of TSVM with biologically significant motifs, and MAD with 6-mers (obtained using the sliding window-based approach). The latter motivated us to further analyze other features derived directly from the data, if/when biologically

relevant motifs are unavailable. In particular, we resort to recently proposed unsupervised feature construction techniques that do not require labeled data [14], [24], [25].

III. Methods

In this section, we describe the three transductive algorithms applied in our work (Section III-A), and the two unsupervised feature generation techniques (Section III-B) used for data representation.

A. Transductive Algorithms

The first learning approach we use in our study is the Transductive Support Vector Machines algorithm (TSVM) [6], which is an extension to the classical supervised SVM algorithm. TSVM, just like its supervised counterpart SVM, is also based on the assumption that nearby points should share the same label and that the separation hyperplane should reside in a low density region of the space. The second and third transductive approaches used are two graph-based approaches, built on the "smoothness" assumption, which states that nodes connected by a strong edge, thereby being very similar, are more likely to share the same label. The training data, comprised of labeled instances $\{(x_1, y_1), ..., (x_l, y_l)\}$ and unlabeled (or test) instances $\{(x_{l+1}, y_{l+1}), ..., (x_u, y_u)\}$, where usually $l \ll u$, are represented as nodes in an undirected graph. The graph is defined as $G = \{V, E, W\}$, where V represents the set of nodes (vertices), $E = V \times V$ is the set of edges corresponding to pairs of nodes, and W is the set of weights associated with edges. The edge weights indicate similarity scores between connected nodes.

The Label Propagation (LP) algorithm [12] spreads the labels of the originally labeled nodes through the graph with the objective of classifying the unlabeled nodes. The smoothness assumption can be formulated as an optimization problem (Equation 1), in which labels $\widehat{y}_i$ and $\widehat{y}_j$ of vertices v_i and v_j, respectively, should be similar for a large W_{ij} in order to minimize the *energy* function, a standard objective function used in graph-based methods [26]. The *energy* function aims to minimize the inconsistencies resulting from the similarity (*i.e.*, edge weight W_{ij}) between examples and their label assignment ($\widehat{y}_i - \widehat{y}_j$). The LP approach also ensures that the original labels are maintained ($\widehat{Y}_l = Y_l$).

$$\min \sum_{i,j} W_{ij} (\widehat{y}_i - \widehat{y}_j)^2, \; s.t. \; \widehat{Y}_l = Y_l. \tag{1}$$

Each unlabeled node receives "soft" labels, or a class distribution, while the labeled nodes maintain their original labels. The actual propagation is realized by means of an iterative algorithm that is repeated until convergence, *i.e.*, until the newly assigned labels do not vary much from one iteration to the next, indicating that the propagation is complete.

Modified Adsorption (MAD) [13] is the third transductive learning algorithm evaluated in our study. It is based on the original Adsorption [27] algorithm and it resembles the concepts of LP [12]. MAD has a well-defined optimization function (Equation 2) that can be solved iteratively in matrix form using the Jacobi method. MAD is a controlled "random walk"-type approach, in which labels are propagated throughout the graph by the means of three probabilities. In order for a vertex v to be labeled, the random walk has three choices associated with probabilities: injection (p_v^{inj}, to stop and return), continuation (p_v^{cont}, to continue the walk to one of v_i's neighbors, v_j), and termination (p_v^{term}, to abandon the walk). Unlike LP, MAD does not reinforce the original labels of the instances, an approach that can potentially alleviate noise in the original labeled training data. However, the first term of the MAD cost function captures the constraint that the inferred labels ($\widehat{Y_{vk}}$) should not significantly differ from the original labels (Y_{vk}). MAD also outputs the uncertainty with respect to newly labeled instances by means of a "dummy variable" initialized to null in the beginning of the algorithm's iterations and assigned a default termination probability if the label propagation process is abandoned at a certain iteration. The second term of MAD's cost function ensures the "smoothness" assumption (that similar nodes should share class labels) and the third term is a regularizer that discourages uncertainty. The importance of each term is controlled by three tunable hyperparameters, μ_1, μ_2, and μ_3.

$$\begin{aligned} \min \sum_v [\mu_1 \sum_k p_v^{inj} (Y_{vk} - \widehat{Y_{vk}})^2 + \\ \mu_2 \sum_v \sum_j p_v^{cont} W_{vj} (\widehat{Y_{vk}} - \widehat{Y_{jk}})^2 + \\ \mu_3 \sum_k p_v^{term} (\widehat{Y_{vk}} - R_{vk})^2] \end{aligned} \tag{2}$$

B. Unsupervised Feature Generation

The Burrows-Wheeler Transform (BWT) [28] is a popular algorithm used in compression, because of the reversible nature of its transformation. In bioinformatics, BWT has been used for sequence alignment, in programs such as Bowtie, BWA, and SOAP2. In our work, we use BWT to construct variable length features, an approach that has been proven to be successful in various learning paradigms, such as supervised, semi-supervised, and domain adaptation [14], [24], [25]. The algorithm mainly identifies multiple occurrences of a subsequence and groups them based on the similarity of the corresponding suffixes. A sequence of size n can have at most n rotations. The rotations are sorted alphanumerically and the last column of the sorted rotations represents the BWT transform of the input sequence. In the transform string, we search for repetitions and obtain features associated with the repetitions by extracting the common prefixes corresponding to each repetition from the sorted rotations. The fact that the BWT algorithm groups prefixes based on lexicographically similar suffixes helps identify variable length features that occur multiple times in a given sequence. In our work, we use features that occur at least twice in at least one input sequence. For more details about the BWT feature generation approach, the reader is directed to [14], [24].

TABLE I: Class labels and number of samples per class for the four protein and two DNA datasets.

Gram-negative		**Gram-positive**		**Plant**		**Non-plant**		***C. elegans***		***D. melanogaester***	
cytoplasm	278	cytoplasmic	194	mitochondrial	368	mitochondrial	371	cassette	487	cassette	164
cytoplasmic membrane	309	cytoplasmic membrane	103	secretory pathway	269	secretory pathway	715	constitutively spliced	2531	constitutively spliced	1246
periplasm	276	cellwall	61	chloroplast	141	other	1652				
outer membrane	391	extracellular	183	other	162						
extracellular	190										
total	**1444**		**541**		**940**		**2738**		**3018**		**1410**

The second feature construction approach that we evaluate with respect to transductive learning is the community detection-based approach (CDA), previously used as a means to obtain variable length features in [15] in an unsupervised manner. This approach leverages communities within a network in which nodes are subsequences (of the instances to be classified) of a certain length. A community thus reflects a subgroup of closely related nodes and can be further refined to form a motif. To identify communities, we use a multi-step technique from [29], based on modularity gain.

We compare the above-mentioned unsupervised approaches to feature construction with supervised feature selection from the set of all variable length k-mers obtained using the traditional sliding window-based approach. The supervised feature selection approach is denoted FSK. All three transductive approaches studied require a pairwise similarity measurement of the instances, given in the form of a kernel function for TSVM, or as a similarity matrix for the graph-based approaches. We use a similarity measure obtained as the opposite of the Euclidean distance.

IV. Biological Problems and Data

We conducted experiments using two DNA datasets and four protein datasets. The DNA datasets consist of nucleotide sequences of exons and their flanking introns, from *C. elegans*, published by Rätsch *et al.* [30], and *D. melanogaster*, constructed in our lab using ALEXA [31]. The *C. elegans* dataset contains 3018 sequences belonging to one of two classes: cassette (487) and constitutive (2531) exons. The *D. melanogaster* dataset contains 1410 sequences belonging to one of the two classes: cassette (164) or constitutive (1246) exons.

The protein datasets contain amino-acid sequences of proteins. The PSORTb v2.0 [32] datasets consist of proteins from Gram-positive and Gram-negative bacteria, and the TargetP datasets [33] consists of proteins from Plant and Non-plant organisms. The Gram-positive dataset contains 541 sequences belonging to one of the following four classes, based on the proteins' localization: cytoplasm (194), cytoplasmic membrane (103), cellwall (61), and extracellular (183). The Gram-negative dataset contains 1444 sequences belonging to one of five classes: cytoplasm (278), cytoplasmic membrane (309), periplasm (276), outer membrane (391), and extracellular (190). The Plant dataset contains 940 sequences belonging to one of four classes: chloroplast (141), mitochondrial (368), secretory pathway/signal peptide (269), and other (consisting of 54 proteins labeled nuclear and 108 examples labeled cytosolic). The Non-plant dataset contains 2738 sequences belonging to one of three classes: mitochondrial (371), secretory pathway/signal peptide (715), and 'other' (consisting of 1224 proteins labeled as nuclear and 438 proteins labeled as cytosolic). The total samples and the number of samples per class for each dataset are summarized in Table I.

V. Experimental Setup

We evaluate the performance of three transductive learning algorithms, TSVM, LP, and MAD, in relation to two unsupervised feature construction methods, BWT and CDA, and one supervised feature selection method, FSK, on two biological sequence classification problems, namely cassette exon identification and protein localization prediction.

The experimental setup is specifically designed to answer the following research questions:

1) What is the most useful feature set for transductive learning, in general?
2) What is the relation between feature sets and specific transductive algorithms?

Typically, the effect of the labeled data on the classification ability, in semi-supervised and transductive frameworks, is far more significant than the effect that the unlabeled data has [34]. Furthermore, supervised feature selection methods benefit from more labeled data. For these reasons, we limit the amount of labeled data to 20% of the total dataset, and the unlabeled instances represent the remaining 80%. We also vary the labeled data from 20% to 5%, by randomly discarding some instances, while the 80% unlabeled data remains fixed.

The datasets of our study are relatively imbalanced, thus measuring the performance in terms of accuracy would not reliably reflect the quality of the classifiers [35]. Therefore, we report the performance in terms of area under the Receiver Operating Characteristic curve (auROC) [36] and area under the Precision-Recall curve (auPRC) [37]. The latter is considered a more appropriate (sensitive) measure for skewed class distributions [38].

The results are averages of a five-fold cross validation procedure, utilized in order to avoid sampling bias. For the protein datasets, which are multi-class problems, we use the 'one class versus all the other classes' approach to evaluate the performance.

We use the SVMLight [34] implementation of TSVM. SVMLight resembles the classical "self-training" [39] approach, in which a completely supervised SVM is built on the labeled data, and next the newly assigned labels of the

TABLE II: Features obtained from 5%, 10%, 15%, and 20% labeled data and all the unlabeled data. The features were generated using the Burrow Wheeler Transform (BWT) approach, the Community Detection Algorithm (CDA), and feature selection over the set of all k-mers of length 2, 3, and 4 obtained with the sliding window (FSK). The number of all k-mers of length 6, 7, and 8 obtained with the sliding window approach ($K_{\{6,7,8\}}$) is also shown.

	C. elegans				*D. melanogaster*			
	5%	**10%**	**15%**	**20%**	**5%**	**10%**	**15%**	**20%**
BWT	3040	3139	3150	3228	3235	3274	3386	3416
CDA	4701	4705	4705	4645	7201	7058	7138	7163
FSK	4701	4705	4705	4645	7201	7058	7138	7163
$K_{\{6,7,8\}}$	66467	66786	67031	67599	73549	74300	74895	74936

(a) Number of features for the DNA datasets

	Gram-positive				Gram-negative			
	5%	**10%**	**15%**	**20%**	**5%**	**10%**	**15%**	**20%**
BWT	3492	3574	3731	3679	3456	3663	3830	3903
CDA	1902	1906	1860	1875	1798	1895	1906	1976
FSK	3492	3574	3731	3679	3456	3663	3830	3903
$K_{\{6,7,8\}}$	80863	82670	84089	84071	69532	72071	74832	76373

(b) Number of features for the bacteria datasets

	Plant				Non-plant			
	5%	**10%**	**15%**	**20%**	**5%**	**10%**	**15%**	**20%**
BWT	5235	5276	5508	5436	2920	2998	3030	3089
CDA	1791	1782	1807	1869	1695	1672	1643	1677
FSK	5235	5276	5508	5436	2920	2998	3030	3089
$K_{\{6,7,8\}}$	101275	101877	103911	103776	75160	77213	77759	79080

(c) Number of features for the Plant/Non-plant datasets

unlabeled (test) data are "switched" in order to optimize the objective function while consistently classifying the originally labeled examples. For the graph-based approaches, we use the Junto Label Propagation Toolkit [13] and we maintain the default parameters.

DNA signals are relatively short, usually 6-14 nucleotides long, and, for this reason, we set the length of features at 6, 7, and 8 nucleotides for the DNA datasets. Protein domains are even shorter, and thus, in the case of proteins, the feature length is set at 2, 3, and 4 amino-acids. We compare the set of features obtained using the unsupervised BWT and CDA approaches with features obtained using supervised feature selection over the set of all k-mers (FSK). To obtain the FSK features, we employ the entropy-based feature selection technique from [40], originally proposed for text categorization, which calculates feature-class correlations based on the labeled data.

We refer to the features obtained using the BWT approach as b-mers, the features obtained with CDA as c-mers, and the features obtained using FSK as f-mers. For a particular dataset, the number of all k-mers is significantly larger than the number of b-mers or the number of c-mers, while the number of b-mers is relatively comparable to the number of c-mers. Therefore, in our study, the number of features selected using the FSK approach was chosen to match the maximum number of features between b-mers and c-mers, *i.e.*, $|\text{f-mers}| = max(|\text{b-mers}|, |\text{c-mers}|)$.

Table I shows the exact sizes of the feature sets obtained from each feature construction technique, when varying the amount of labeled data from 5% to 20%: table IIa for the DNA datasets *C. elegans* and *D. melanogaster*, table IIb for the Gram-positive and Gram-negative bacteria, and table IIc for the Plant and Non-Plant datasets.

VI. Results

In this section, we present the results of our transductive experiments for the six datasets described in Section IV. For each dataset, we present experiments with increasingly larger amounts of labeled data (while keeping the unlabeled data at 80% of the total training amount). For each transductive learning algorithm described in Section III-A, we evaluate the three feature generation techniques described in Section III-B, and present their performance in terms of auROC and auPRC. The results are shown in Table III for DNA (table IIIa for *C. elegans* and table IIIb for *D. melanogaster*), and Table IV for the protein datasets (table IVa for Gram-negative bacteria, table IVb for Gram-positive bacteria, table IVc for Plant, and IVd for Non-plant). The values in bold font represent the best performance for a given algorithm (TSVM, MAD, or LP) and a given amount of labeled data, and the highlighted (shaded) cells show the best result overall, for a given experiment (*i.e.,* a given amount of labeled data). Next, we answer the research questions enumerated in the beginning of Section V.

1) What is the most useful feature set for transductive learning, in general?

BWT is generally the best unsupervised feature construction technique, finding good features especially for protein datasets. For DNA datasets, supervised feature selection (FSK) gives the best performance, followed by BWT. To gain insights into these results, first, we should note that for the DNA datasets, the number of FSK features is almost twice the number of BWT features (as it equals the larger number of CDA features), and that gives some advantage to the FSK features. Furthermore, as can be seen from the results Table

TABLE III: Results for the DNA datasets, in terms of auROC and auPRC averaged values over the five folds, for increasingly larger amounts of labeled data, while maintaining the unlabeled data at a fixed 80%. For each transductive learning algorithm TSVM, LP, and MAD, we have evaluated the four feature sets based on BWT, CDA, and FSK. The values in bold font represent the best performance for a given algorithm and the highlighted (shaded) cells show the best result overall.

C. elegans										
		TSVM (auROC)			**LP (auROC)**			**MAD (auROC)**		
Labeled	#Samples	*BWT*	*CDA*	*FSK*	*BWT*	*CDA*	*FSK*	*BWT*	*CDA*	*FSK*
5%	150	0.560	0.540	**0.591**	0.638	0.626	**0.700**	0.649	0.632	**0.722**
10%	301	0.623	0.605	**0.659**	0.690	0.683	**0.809**	0.702	0.694	**0.814**
15%	452	0.660	0.664	**0.714**	0.735	0.731	**0.860**	0.746	0.742	**0.862**
20%	603	0.712	0.701	**0.737**	0.776	0.763	**0.852**	0.779	0.769	**0.856**
		TSVM (auPRC)			**LP (auPRC)**			**MAD (auPRC)**		
Labeled	#Samples	*BWT*	*CDA*	*FSK*	*BWT*	*CDA*	*FSK*	*BWT*	*CDA*	*FSK*
5%	150	0.538	0.525	**0.552**	0.596	0.586	**0.659**	0.610	0.596	**0.675**
10%	301	0.564	0.557	**0.605**	0.625	0.620	**0.767**	0.638	0.632	**0.764**
15%	452	0.601	0.605	**0.653**	0.668	0.667	**0.828**	0.679	0.676	**0.822**
20%	603	0.642	0.627	**0.675**	0.710	0.697	**0.811**	0.716	0.705	**0.813**

(a) Averages of auROC and auPRC values over the five folds for the *C. elegans* dataset

D. melanogaster										
		TSVM (auROC)			**LP (auROC)**			**MAD (auROC)**		
Labeled	#Samples	*BWT*	*CDA*	*FSK*	*BWT*	*CDA*	*FSK*	*BWT*	*CDA*	*FSK*
5%	70	**0.601**	0.580	0.590	**0.639**	0.627	0.633	**0.642**	0.631	0.633
10%	141	0.598	0.589	**0.611**	**0.690**	0.675	0.746	0.697	0.682	**0.750**
15%	211	**0.648**	0.632	0.638	0.723	0.710	**0.838**	0.725	0.713	**0.839**
20%	282	**0.659**	0.648	0.622	0.754	0.743	**0.890**	0.753	0.744	**0.915**
		TSVM (auPRC)			**LP (auPRC)**			**MAD (auPRC)**		
Labeled	#Samples	*BWT*	*CDA*	*FSK*	*BWT*	*CDA*	*FSK*	*BWT*	*CDA*	*FSK*
5%	70	**0.538**	0.529	0.532	0.556	0.554	**0.562**	0.558	0.556	**0.561**
10%	141	0.535	0.534	**0.549**	0.590	0.581	**0.645**	0.595	0.584	**0.649**
15%	211	**0.568**	0.562	0.563	0.627	0.617	**0.771**	0.628	0.619	**0.778**
20%	282	**0.574**	0.567	0.563	0.646	0.638	**0.846**	0.642	0.635	**0.862**

(b) Averages of auROC and auPRC values over the five folds for the *D. melanogaster* dataset

III, the cases when BWT is better than FSK for the DNA problems, generally, correspond to smaller amounts of labeled data (*i.e.*, 5% or 10%) in *D. melanogaster*. We should also note that the data available for *D. melanogaster* in our experiments is less than half the data available for *C. elegans*. Thus, our results suggest that for *C. elegans*, 5% of the data is still enough to obtain good correlations between k-mers and the class variable, while this is not the case for *D. melanogaster*, which has less data (as well as a higher class imbalance) in the first place. Similarly, as the protein datasets are smaller, FSK does not usually identify an informative set of features for those datasets.

2) What is the relation between feature sets and specific transductive algorithms?

When studying the relation between different sets of features and different transductive algorithms, we see that generally, if the set of features is good, almost all algorithms will benefit from them. For example, all algorithms give their best performance when used with the FSK features for all the *C. elegans* experiments, while they give their best performance when used with BTW features for the Gram-negative bacteria, and when used with the CDA features for Plant, suggesting that those features are informative for all algorithms. Furthermore, it can be seen that the relation between features and transductive algorithms does not depend on the amount of labeled data, although, not surprisingly, larger amounts of labeled data lead to better classifiers.

VII. Conclusion

In this study, we have evaluated the performance of three transductive learning algorithms and their relation to unsupervised feature construction techniques for biological sequence classification. More specifically, we have used transductive SVM, Label Propagation, and Modified Adsorption algorithms for the problems of cassette exon identification and protein localization. We have formulated the problems as classification tasks and experimented with two DNA datasets and four protein datasets. We have used methods based on Burrows-Wheeler Transform and Community Detection and compared their performance with a supervised technique, namely feature selection on the total number of derived k-mers. Our results show that transduction is applicable to such problems in the presence of limited labeled data and can achieve good classification performance with compact feature sets obtained using unsupervised feature construction methods.

In future work, we plan to address other DNA classification problems and also evaluate graph-based algorithms on larger DNA datasets, for which biologically relevant features are not easily available, thus making unsupervised feature generation a potential solution.

TABLE IV: Results for the protein datasets, in terms of auROC and auPRC averaged values over the five folds, for increasingly larger amounts of labeled data, while maintaining the unlabeled data at a fixed 80%. For each transductive learning algorithm TSVM, LP, and MAD, we have evaluated the four feature sets based on BWT, CDA, and FSK. The values in bold font represent the best performance for a given algorithm and the highlighted (shaded) cells show the best result overall.

Gram-negative bacteria										
		TSVM (auROC)			**LP (auROC)**			**MAD (auROC)**		
Labeled	#Samples	*BWT*	*CDA*	*FSK*	*BWT*	*CDA*	*FSK*	*BWT*	*CDA*	*FSK*
5%	72	**0.526**	0.516	0.519	**0.870**	0.847	0.830	**0.870**	0.848	0.839
10%	144	**0.648**	0.596	0.636	**0.886**	0.867	0.828	**0.891**	0.873	0.833
15%	216	**0.757**	0.644	0.645	**0.898**	0.878	0.830	**0.903**	0.882	0.837
20%	288	**0.848**	0.789	0.741	**0.901**	0.888	0.863	**0.906**	0.892	0.862
		TSVM (auPRC)			**LP (auPRC)**			**MAD (auPRC)**		
Labeled	#Samples	*BWT*	*CDA*	*FSK*	*BWT*	*CDA*	*FSK*	*BWT*	*CDA*	*FSK*
5%	72	**0.519**	0.512	0.511	**0.659**	0.614	0.576	**0.661**	0.624	0.590
10%	144	**0.578**	0.556	0.575	**0.673**	0.649	0.559	**0.688**	0.669	0.562
15%	216	**0.651**	0.583	0.585	**0.705**	0.668	0.538	**0.721**	0.682	0.553
20%	288	**0.757**	0.698	0.659	**0.722**	0.692	0.605	**0.736**	0.701	0.600

(a) Averages of auROC and auPRC values over the five folds for the Gram-negative bacteria dataset

Gram-positive bacteria										
		TSVM (auROC)			**LP (auROC)**			**MAD (auROC)**		
Labeled	#Samples	*BWT*	*CDA*	*FSK*	*BWT*	*CDA*	*FSK*	*BWT*	*CDA*	*FSK*
5%	27	**0.928**	0.895	0.880	0.877	**0.882**	0.802	0.893	**0.901**	0.807
10%	54	0.912	0.871	**0.913**	0.905	**0.920**	0.795	0.925	**0.929**	0.836
15%	81	**0.927**	0.889	0.916	0.914	**0.924**	0.766	0.932	**0.936**	0.848
20%	108	**0.946**	0.912	0.908	0.929	**0.940**	0.811	0.942	**0.945**	0.863
		TSVM (auPRC)			**LP (auPRC)**			**MAD (auPRC)**		
Labeled	#Samples	*BWT*	*CDA*	*FSK*	*BWT*	*CDA*	*FSK*	*BWT*	*CDA*	*FSK*
5%	27	**0.847**	0.805	0.833	**0.711**	0.710	0.580	0.727	**0.739**	0.582
10%	54	**0.851**	0.812	0.842	0.763	**0.771**	0.565	**0.802**	0.797	0.613
15%	81	0.849	0.813	**0.852**	0.770	**0.778**	0.553	0.809	**0.812**	0.627
20%	108	0.877	0.839	**0.882**	0.806	**0.826**	0.606	0.829	**0.834**	0.661

(b) Averages of auROC and auPRC values over the five folds for the Gram-positive bacteria dataset

Plant										
		TSVM (auROC)			**LP (auROC)**			**MAD (auROC)**		
Labeled	#Samples	*BWT*	*CDA*	*FSK*	*BWT*	*CDA*	*FSK*	*BWT*	*CDA*	*FSK*
5%	47	0.632	**0.738**	0.676	0.699	**0.773**	0.699	0.741	**0.776**	0.721
10%	94	0.694	**0.739**	0.692	0.762	**0.793**	0.723	0.795	**0.817**	0.746
15%	141	**0.852**	0.797	0.700	0.808	**0.837**	0.777	0.820	**0.846**	0.810
20%	188	**0.890**	0.884	0.766	0.840	**0.849**	0.819	0.847	**0.863**	0.812
		TSVM (auPRC)			**LP (auPRC)**			**MAD (auPRC)**		
Labeled	#Samples	*BWT*	*CDA*	*FSK*	*BWT*	*CDA*	*FSK*	*BWT*	*CDA*	*FSK*
5%	47	0.613	**0.703**	0.645	0.445	**0.521**	0.442	0.493	**0.526**	0.470
10%	94	0.671	**0.702**	0.657	0.515	**0.536**	0.472	0.561	**0.579**	0.493
15%	141	**0.817**	0.764	0.660	0.574	**0.600**	0.508	0.601	**0.620**	0.539
20%	188	**0.865**	0.860	0.734	0.613	**0.617**	0.549	0.634	**0.649**	0.552

(c) Averages of auROC and auPRC values over the five folds for the Plant dataset

Non-plant										
		TSVM (auROC)			**LP (auROC)**			**MAD (auROC)**		
Labeled	#Samples	*BWT*	*CDA*	*FSK*	*BWT*	*CDA*	*FSK*	*BWT*	*CDA*	*FSK*
5%	137	0.538	0.606	**0.634**	**0.782**	0.758	0.716	**0.794**	0.765	0.725
10%	273	0.705	**0.770**	0.733	**0.809**	0.808	0.752	**0.839**	0.823	0.784
15%	410	**0.851**	0.737	0.765	**0.836**	0.819	0.822	**0.855**	0.828	0.827
20%	547	**0.875**	0.813	0.818	**0.839**	0.822	0.817	**0.861**	0.840	0.830
		TSVM (auPRC)			**LP (auPRC)**			**MAD (auPRC)**		
Labeled	#Samples	*BWT*	*CDA*	*FSK*	*BWT*	*CDA*	*FSK*	*BWT*	*CDA*	*FSK*
5%	137	0.534	0.583	**0.605**	0.576	**0.578**	0.519	**0.595**	0.581	0.529
10%	273	0.664	**0.715**	0.687	0.611	**0.651**	0.553	0.659	**0.678**	0.604
15%	410	**0.810**	0.683	0.722	0.659	**0.675**	0.641	0.688	**0.697**	0.652
20%	547	**0.837**	0.776	0.782	0.657	**0.680**	0.632	0.693	**0.714**	0.657

(d) Averages of auROC and auPRC values over the five folds for the Non-plant dataset

ACKNOWLEDGMENT

The computing for this project was performed on the Beocat Research Cluster at Kansas State University, which is funded in part by NSF grants ACI-1341026, CNS-1126709, CNS-1006860, and EPS-0919443.

REFERENCES

[1] T. M. Mitchell, *Machine learning*. McGraw-Hill, 1997.

[2] S. Kotsiantis, "Supervised machine learning: A review of classification techniques," in *Proceedings of the 2007 conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*. IOS Press, 2007, pp. 3–24.

[3] R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, *Machine learning: An artificial intelligence approach*. Springer Science & Business Media, 2013.

[4] X. Zhu and A. B. Goldberg, "Introduction to semi-supervised learning," *Synthesis lectures on artificial intelligence and machine learning*, vol. 3, no. 1, pp. 1–130, 2009.

[5] O. Chapelle, B. Schlkopf, and A. Zien, *Semi-Supervised Learning*, 1st ed. The MIT Press, 2010.

[6] V. N. Vapnik and V. Vapnik, *Statistical learning theory*. Wiley New York, 1998, vol. 2.

[7] A. Gammerman, V. Vovk, and V. Vapnik, "Learning by transduction," in *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, ser. UAI'98. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998, pp. 148–155.

[8] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, ser. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001.

[9] J. Xia, D. Caragea, and S. J. Brown, "Prediction of alternatively spliced exons using support vector machines," *International Journal of Data Mining and Bioinformatics*, vol. 4, no. 4, pp. 411–430, Jul. 2010.

[10] A. Stanescu, K. Tangirala, and D. Caragea, "Predicting alternatively spliced exons using semi-supervised learning," *International Journal of Data Mining and Bioinformatics*, vol. 14, no. 1, pp. 1–21, 2016.

[11] A. Stanescu and D. Caragea, "Predicting cassette exons using transductive learning approaches," in *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2015 IEEE Conference on*. IEEE, 2015, pp. 1–8.

[12] X. Zhu and Z. Ghahramani, "Learning from labeled and unlabeled data with label propagation," CMU-CALD-02-107, Carnegie Mellon University, Tech. Rep., 2002.

[13] P. P. Talukdar and K. Crammer, "New regularized algorithms for transductive learning," in *Machine Learning and Knowledge Discovery in Databases*. Springer Berlin Heidelberg, 2009, pp. 442–457.

[14] K. Tangirala and D. Caragea, "Generating features using Burrows Wheeler Transformation for biological sequences," in *Proceedings of the 5th International Conference on Bioinformatics Models, Methods and Algorithms*, ser. BIOINFORMATICS 2014, 2014, pp. 185–192.

[15] K. Tangirala and D. Caragea, "Community detection-based features for sequence classification," in *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM, 2014, pp. 559–568.

[16] N. Kasabov and S. Pang, "Transductive support vector machines and applications in bioinformatics for promoter recognition," in *Neural Networks and Signal Processing. Proceedings of the 2003 International Conference on*, vol. 1, 2003, pp. 1–6.

[17] S. Pang and N. Kasabov, "Inductive vs transductive inference, global vs local models: SVM, TSVM, and SVMT for gene expression classification problems," in *International Joint Conference on Neural Networks*, vol. 2. IEEE, 2004, pp. 1197–1202.

[18] J. Weston, C. Leslie, E. Ie, D. Zhou, A. Elisseeff, and W. S. Noble, "Semi-supervised protein classification using cluster kernels," *Bioinformatics*, vol. 21, no. 15, pp. 3241–3247, 2005.

[19] E. Kondratovich, I. I. Baskin, and A. Varnek, "Transductive SVM: Promising Approach to Model Small and Unbalanced Datasets," *Molecular Informatics*, vol. 32, no. 3, pp. 261–266, 2013.

[20] H. Shin, K. Tsuda, and B. Schölkopf, "Protein functional class prediction with a combined graph," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3284–3292, 2009.

[21] G. Yu, C. Domeniconi, H. Rangwala, G. Zhang, and Z. Yu, "Transductive multi-label ensemble classification for protein function prediction," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 1077–1085.

[22] L. De Baets, "Identifying novel neuroblastoma oncogenes using machine learning," Master's thesis, Department of Information Technology, Faculty of Engineering and Architecture, Universiteit Gent, 2014.

[23] R. Yong, K. Nobuhiro, N. Yoshinaga, and M. Kitsuregawa, "Sentiment classification in under-resourced languages using graph-based semi-supervised learning methods," *IEICE TRANSACTIONS on Information and Systems*, vol. 97, no. 4, pp. 790–797, 2014.

[24] K. Tangirala and D. Caragea, "Semi-supervised classification of protein sequences using burrows wheeler transformation-based features," in *Proceedings of the 6th International Conference on Bioinformatics and Computational Biology*, ser. BICoB 2014, 2014, pp. 21–26.

[25] N. Herndon, K. Tangirala, and D. Caragea, "Predicting protein localization using a domain adaptation na¨ ıve bayes classifier with burrows wheeler transform features," in *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on*. IEEE, 2014, pp. 501–504.

[26] A. Agovic and A. Banerjee, "A unified view of graph-based semi-supervised learning: Label propagation, graph-cuts, and embeddings," *University of Minnesota, Tech. Rep. CSE*, pp. 09–012, 2009.

[27] S. Baluja, R. Seth, D. Sivakumar, Y. Jing, J. Yagnik, S. Kumar, D. Ravichandran, and M. Aly, "Video suggestion and discovery for youtube: taking random walks through the view graph," in *Proceedings of the Seventeenth International Conference on World Wide Web*. ACM, 2008, pp. 895–904.

[28] M. Burrows and D. J. Wheeler, "A block-sorting lossless data compression algorithm." Digital Equipment Corp., Palo Alto, CA, Tech. Rep. 124, 1994.

[29] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, 2008.

[30] G. Rätsch, S. Sonnenburg, and B. Schölkopf, "RASE: recognition of alternatively spliced exons in *C. elegans*," in *Proceedings of 13th International Conference on Intelligent Systems for Molecular Biology (ISMB)*, vol. 21, 2005, pp. 369–377.

[31] M. Griffith, M. Tang, O. Griffith, R. Morin, S. Chan, J. Asano, T. Zeng, S. Flibotte, A. Ally, A. Baross *et al.*, "ALEXA: a microarray design platform for alternative expression analysis." *Nature Methods*, vol. 5, no. 2, p. 118, 2008.

[32] J. L. Gardy, M. R. Laird, F. Chen, S. Rey, C. Walsh, M. Ester, and F. S. Brinkman, "Psortb v. 2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis," *Bioinformatics*, vol. 21, no. 5, pp. 617–623, 2005.

[33] O. Emanuelsson, H. Nielsen, S. Brunak, and G. von Heijne, "Predicting subcellular localization of proteins based on their n-terminal amino acid sequence," *Journal of Molecular Biology*, vol. 300, no. 4, pp. 1005–1016, 2000.

[34] T. Joachims, "SVMLight: Support vector machine," *SVM-Light Support Vector Machine http://svmlight.joachims.org, University of Dortmund*, vol. 19, no. 4, 1999.

[35] F. J. Provost, T. Fawcett, and R. Kohavi, "The case against accuracy estimation for comparing induction algorithms," in *Proceedings of the 15th International Conference on Machine Learning*, ser. ICML '98. Morgan Kaufmann Publishers Inc., 1998.

[36] J. Huang and C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, pp. 299–310, 2005.

[37] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in *Proceedings of The Twenty Third International Conference on Machine Learning*. ACM, 2006, pp. 233–240.

[38] L. Jeni, J. Cohn, and F. de la Torre, "Facing imbalanced data–recommendations for the use of performance metrics," in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, Sept 2013, pp. 245–251.

[39] D. Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods," in *Proceedings of The Thirty Third Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 1995, pp. 189–196.

[40] C. Largeron, C. Moulin, and M. Gèry, "Entropy based feature selection for text categorization," in *Proc. of the 2011 ACM Symp. on Applied Computing*, ser. SAC '11. New York, NY, USA: ACM, 2011, pp. 924–928.