# Context-Aware Semi-Supervised Motif Detection Approach

Rania Ibrahim, Nagia Ghanem and Mohamed A. Ismail[1]

*Abstract*— **Motif detection has raised as an important task in bioinformatics. Recently, the discovery of motifs that are localized relative to a certain biological area has become an important task in many applications. For example, it is used to discover regulatory sequences beside the transcription start site and the neighborhood of known transcription factor binding sites [1]. Therefore, the idea of context aware motif detection approach is needed. Moreover, there is an interest to use both labeled and unlabeled sets to enhance the motif detection approaches.**

**In this paper, three novel context aware semi-supervised motif detection approaches are proposed, which are self-learning, context aware and co-training context aware systems. In self-learning motif Hidden Markov Model (HMM) is enhanced independently using unlabeled sets. While in co-training, three different models are trained based on three different views which are pre-motif sequences, motif sequences and post-motif sequences. Moreover, our co-training context aware system is suitable for parallelization to enhance its execution time. The approaches were evaluated using human motif sequences and the results show that co-training context aware system has achieved the best results. The results also show that our approach outperforms other related works in [1], [2] and [3].**

## I. INTRODUCTION

Regulation of gene expression is usually occurred at the stage of transcription of the gene [1]. Transcription of genes is done by binding of transacting proteins called transcription factors (TFs) to DNA sequences in the regions of the gene. The TFs bind to short 520 bp segments of DNA called transcription factor binding sites (TFBSs). However, the main problem is that TFBSs are short of length and frequently have mutations, which makes them hard to be recognized. This problem can be solved using motif detection as motifs can be discovered within a set of sequences that are frequently occur. Therefore approaches usually looks for a short, conserved and repeated pattern called the motif in these sequences and which is likely to be the TFBS.

A number of approaches for motif detection have been introduced in the literature like in [1-5]. A useful piece of information that has not been adequately exploited in the existing motif finding algorithms is the position and location of motifs. Contextual property of motifs is an important property that has been limited in the research area. [1] utilizes the contextual characteristics of the motif position by using a novel scoring function and combining it with existing scoring functions like entropy and over-representation genes metrics. In addition, probabilistic methods, which search de novo for statistically overrepresented motifs in co-regulated have been proven successful for the prediction of regulatory motifs like in [4] and [5]. In this work, both context properties and statistically approaches are utilized to enhance motif detection results.

Moreover, in this paper, two semi-supervised machine learning approaches, namely self-learning [6] and co-training ([7], [8]) are used to enhance motif detection approaches by combining both labeled and unlabeled sets. In self-learning, a Hidden Markov Model (HMM) is initially trained using the labeled motif set, then its accuracy is enhanced by adding more data from unlabeled sets. In co-training, three HMMs are trained; each is specific to a different source of sequences (pre-motif, motif or post-motif), termed as three views of the data. Based on the three views, three HMM models are constructed and then used to train each other.

The proposed systems which are context aware and co-training context aware systems both perform an overhead computation by searching for the pre and post sequences to the motif and then training\running the one\three HMM(s) model(s). However, this problem can be solved by introducing parallelization. The three HMMs models can run in parallel and thus no additional execution time is introduced except in the searching for the pre and post motifs sequences.

The paper is organized as follows: section II discusses the related work, while section III describes the proposed approaches in detail and section IV shows experimental results. Finally section V concludes the paper.

## II. RELATED WORK

A number of attempts for motif detection have been introduced in the literature like in [1-5]. Recently, using contextual information to enhance motif detection have been explored in [1] using a novel scoring function. The scoring function uses the context information and combines it with other existing scoring functions to capture entropy and over-representation of the motif. However, in this paper, trained HMMs are automated to discover the hidden patterns in the motif sequences and its context sequences. Moreover, [4] starts by using a motif sampler component. Then, giving a DNA sequences and using the statistics the approach got, the approach gets the multiple possible motifs and then it applies fuzzy clustering to rank these motifs. After that, using existence motif databases, the approach compares the extracted motifs with the motif databases and try to align them together and the result is a group of aligned motifs. Finally, the aligned motifs are passed to motif locator to output the exact motifs. The previous approach didnt use the context information of the motif sequences, while our

[1] Rania Ibrahim, Nagia Ghanem and Mohamed A. Ismail are with Computer and Systems Engineering Department, Alexandria University, Alexandria 21544, Egypt rania.ibrahim.salama@gmail.com, nagia.ghanem@alexu.edu.eg and drmaismail@gmail.com

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

approach utilizes them to enhance motif detection accuracy. In addition [9] proposes an algorithm called SLUPC, which performs a separate-and-conquer searching method to discover discriminative one occurrence per sequence motifs. Also, the paper proposes E-SLUPC (Ensemble SLUPC), which uses SLUPC to search discriminative motifs from datasets that contains both labeled and unlabeled data. The paper didn't explore context information and only used self-learning semi-supervised machine learning approach to combine labeled and unlabeled sets. Moreover, [2] and [3] proposed a different methods for motif detection, they also didnt explore context information or the usage of the unlabeled sets.

Other semi-supervised machine learning approaches like self-learning and co-training were introduced in other domains. The heuristic approach of self-learning (also known as self-training) is one of the oldest approaches in semi-supervised learning and that was introduced in [6]. Self-learning was used in many applications as object detection [10] and word sense disambiguation [11]. Also, co-training is a semi-supervised approach that appeared in [7] and [8] and is also used in applications as word sense disambiguation [12] and email classification [13]. In addition, co-training has proved its ability for enhancing cancer sample classifiers based on combining both miRNAs and genes expression profiles, which was shown in [14]. Both contextual information and semi-supervised machine learning approaches are combined in our systems to enhance motif detection. The next sections explain the systems in detail and show their results in human sequences.

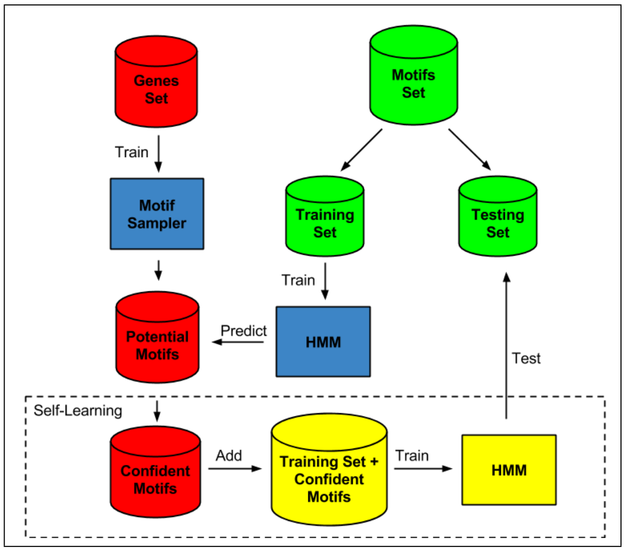

Fig. 1. Self-Learning System Overview

## III. System Overview

In this paper, three systems are proposed which are self-learning, context aware and co-training context-aware systems. Our objective is to use unlabeled sets to enhance motif detection accuracy, in addition to using context information. In self-learning, a Hidden Markov Model (HMM) is initially trained using the labeled motif set and enhanced by adding more data from unlabeled sets. While in co-training, three HMMs are trained; each on a different view (pre-motif, motif or post-motif). The next subsections explain the three proposed systems in detail. Fig. 1 and 2 show system overview of the self-learning and context aware systems respectively.

### *A. Self-Learning System*

Self-Learning [6] is a semi-supervised machine learning approach which is used to enhance classification accuracy. The main idea is to let the model teach itself by itself. The steps of the approach are described as follows:

1) Train an initial model using a small number of labeled samples.
2) Apply the model on unlabeled set(s).
3) Choose samples from unlabeled set with classification confident above a certain threshold and add them to the training set.
4) Repeat step (a) using the new training set.

The steps are repeated until either no further improvement is introduced or for a certain number of iterations. The main advantage of this method is that it combines both labeled and unlabeled set(s) and use the unlabeled set(s) to enhance the accuracy. Fig. 1 shows self-learning system overview and its adaption to work on the motif models. In order to construct the unlabeled set of motifs, unlabeled set of genes is chosen and then the motif sampler component in [4] is used to find a set of potential motifs using statistical sampling method. This generated set is considered as our unlabeled set where the initial HMM -trained on a small labeled set of motifs- is applied on it. Finally, potential motifs with classification confident above $\alpha$ are added to the training set and the HMM is re-trained. $\alpha$ was tuned experimentally and was finally set to value 10 in our experiments.

### *B. Context Aware System*

In order to enhance motif detection accuracy, context properties are used in our system. Context properties are captured by considering sequences that appear before and after the motifs. These sequences are named pre-motifs and post-motifs sequences respectively and are utilized to enhance the detection of the motifs. The steps of the context aware system are described as follows:

1) Retrieve pre-motifs sequences and post-motifs sequences by searching for the motif sequence in the gene unlabeled set and retrieving the sequence before and after it, within a certain window size.
2) Train three HMMs based on motif sequences, pre-motif sequences and post-motif sequences.
3) Combine HMMs decisions using union function. Other functions can be used for data fusion, but union function was explored in this paper.

As this approach captures the context properties of the motif, it allows for discovering local motifs. Fig. 2 shows context-aware system overview.

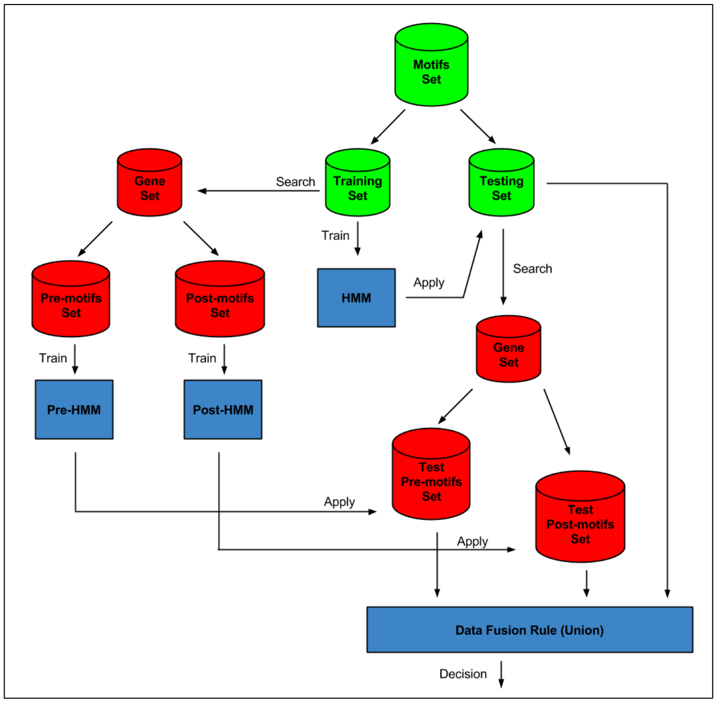

Fig. 2. Context Aware System Overview

### C. Co-Training Context Aware System

The final proposed system combines both the context information and unlabeled sets usage. In order to do so, the previous two systems are combined in harmony using the co-training semi-supervised approach. Co-Training ([7], [8]) is a semi-supervised approach which constructs a number of classifiers based on different views of the data and let them train each other. The steps of the system are described as follows:

1) Train three HMMs based on motif sequences, pre-motif sequences and post-motif sequences.
2) Apply each one of the three HMMs to a potential set of motifs to detect whether it is motif or not based on each view and identify each sample confident score.
3) Let each HMM teaches the other one, by selecting the confident motifs for each HMM and adding it to the training motifs of the other HMMs. Confident motifs are the motifs with confident score of HMM above $\alpha$. As in self-learning, $\alpha$ was tuned experimentally.
4) Combine HMMs decisions using union function.

Moreover, as shown in the steps, there is an overhead introduced in searching for pre and post sequences and for training\running the three HMMs. The training and running problems can be solved by introducing parallelization and training\running the three HMMs in parallel. However, there is still the overhead of searching for the pre and post motif sequences and reducing this overhead is considered as one of our future work.

Next section shows the experimental results in detail and also shows the improvements of the co-training context aware system.

## IV. Experimental Results

The first experiment is performed by comparing all the proposed systems to the baseline system, which is constructed by using an existing motif database to train Hidden Markov Model (HMM) and use the previous model to detect motifs in the test set. Experiments are conducted on the following datasets, first sequences of unlabeled dataset were used from [15] which contains the sequence of 104,763 genes. While, labeled motif datasets was used from [16] only 600 motifs were used from that set and the set was divided into 300 training motifs set and 300 test motifs set. Moreover, motif sampler online tool [17] was used which is a de novo motif detection tool designed to search for regulatory motifs in DNA sequences based on statistical measures of the given sequences. Hidden Markov Model (HMM) training tool is used from [18] to train motif sequence HMM, pre-motif sequence HMM and post-motif sequence HMM. Self-Learning and co-training systems were implemented in Java. Table I shows the number of detected motifs and confident detected motifs discovered by each one of the four systems, while fig. 3 shows the precision, recall and f-measure of our four systems.

Moreover, a comparison against 3 recent related works is conducted which are LOCALMOTIF [1], TRAWLER [2] and AMADEUS [3]. Table II shows the number of sequences at each set and the number of motifs detected by comparing them to our three proposed approaches in four datasets, which are OCT4, SOX2, NANOG and HNF6. The four sets have been recognized by ChIP-Chip experiments within -8 kb to +2 kb region flanking the TSS ([20], [21]).

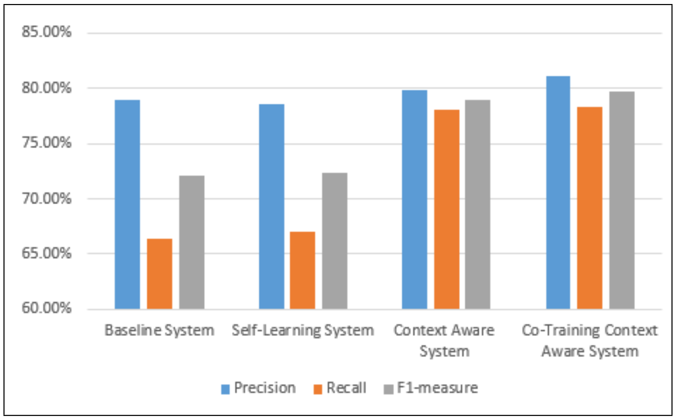

Fig. 3. Precision, recall and F1-measure of the four systems, using [15] and [16]

TABLE I

NUMBER OF DETECTED MOTIFS AND CONFIDENT DETECTED MOTIFS DISCOVERED BY EACH SYSTEM

| | No. of Detected Motifs | No. of Confident Detected Motifs (p-values <= 0.05) |
|---|---|---|
| **Baseline System** | 199 | 39 |
| **Self-Learning** | 201 | 41 |
| **Context-aware** | 234 | 71 |
| **Co-Training Context-aware** | **235** | **77** |

TABLE II

NUMBER OF MOTIFS DETECTED USING OUR APPROACHES AGAINST THREE RELATED WORKS ([1], [2], [3])

| | OCT4 | SOX2 | NANOG | HNF6 | CREB1 |
|---|---|---|---|---|---|
| **# of Sequences** | 372 | 785 | 1029 | 1238 | 2108 |
| **TRAWLER** | - | - | - | 1 | 1 |
| **AMADEUS** | - | - | - | - | 1 |
| **LOCAL MOTIF** | - | 8 | 8 | 6 | 1 |
| **Self-Learning** | - | - | - | - | - |
| **Context Aware** | 7 | 7 | 66 | 18 | 24 |
| **Co-Training Context Aware** | **31** | **29** | **69** | **23** | **14** |

## V. CONCLUSIONS

In this paper, three novel motif detection systems were introduced, which are self-learning, context aware and co-training context aware systems. The main idea was to combine both contextual properties of the motif in addition to combining labeled and unlabeled sets using semi-supervised machine learning approaches. Results show that context aware co-training system was able to achieve an improvement of 12% in the recall and 7.5% increase in the f1-measure over the baseline system. The results also show that our approach outperforms other related works in [1], [2] and [3].

## REFERENCES

[1] V. Narang, A. Mittal and W. Sung. Localized motif discovery in gene regulatory sequences. Bioinformatics. (29):9, pp. 1152-1159, 2010.

[2] L. Ettwiller. Trawler: de novo regulatory motif discovery pipeline for chromatin immunoprecipitation. Nat. Methods, 4, 563565., 2007.

[3] C. Linhart. Transcription factor and microRNA motif discovery: the Amadeus platform and a compendium of metazoan target sets. Genome Res., 18, 11801189. 2008.

[4] G. Thijs. A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. J. Comput. Biol., 9, 447464, 2002.

[5] T. L. Bailey. MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res., 34, W369W373, 2006.

[6] O. Chapelle, B. Schlkopf and A. Zien. Semi-supervised learning. Cambridge, Mass., MIT Press, 2006.

[7] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. Proceedings of the Workshop on Computational Learning Theory (COLT), pp. 92-100, Wisconsin, USA, 1998.

[8] T. Mitchell. The role of unlabeled data in supervised learning. Proceedings of the Sixth International Colloquium on Cognitive Science (ICCS), San Sebastian, Spain, 1999.

[9] T. Nhan Le, T. Bao Ho, S. Kawasak. A Semi-Supervised Ensemble Learning Method for Finding Discriminative Motifs and its Application. Journal of Universal Computer Science, (19):4, 2013.

[10] C. Rosenberg, M. Hebert and H. Schneiderman. Semi-Supervised Self-Training of Object Detection Models. 7th IEEE Workshop on Applications of Computer Vision / IEEE Workshop on Motion and Video Computing (WACV/MOTION), pp. 29-36, Breckenridge, CO, USA , 2005.

[11] R. Mihalcea. Co-training and Self-training for Word Sence Disambiguation. In Proceedings of CoNLL, pp. 33 40, Boston, MA, USA, 2004.

[12] R. Mihalcea. Co-training and Self-training for Word Sense Disambiguation. In Proceedings of the Conference on Natural Language Learning (CoNLL), pp. 33-40, Boston, USA, May 2004.

[13] S. Kiritchenko and S. Matwin. Email Classification with Co-Training. In Proceedings of Conference of the Center for Advanced Studies on Colleborative Research (CASCON), pp. 301-312, 2011.

[14] R. Ibrahim, N. A. Yousri, M. A. Ismail and N. M, El-Makky. miRNA and Gene Expression based Cancer Classification using Self-Learning and Co-Training Approaches. IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 495-498, 2013.

[15] `ftp://ftp.ensembl.org/pub/release-74/fasta/homo_sapiens/cds/` [last seen 16/3/2014]

[16] `http://johnsonlab.ucsf.edu/mochiView/motif-library-downloads.html` [last seen 16/3/2014]

[17] `http://bioinformatics.psb.ugent.be/webtools/MotifSuite/motifshampler.php` [last seen 16/3/2014]

[18] `http://meme.nbcr.net/meme/` [last seen 16/3/2014]

[19] `http://www.lncrnadb.org/Detail.aspx?TKeyID=229` [last seen 16/3/2014]

[20] L. A. Boyer.Core transcriptional regulatory circuitry in human embryonicstem cells. Cell, 122, 947956. 2005.