IEM: An Algorithm for Iterative Enhancement of Motifs Using Comparative Genomics Data

Erliang Zeng[1], Kalai Mathee[2], and Giri Narasimhan[1*],
[1] Bioinformatics Research Group (BioRG), School of Computing and Information Sciences, Florida International University, Miami, Florida, 33199, USA, and [2]Department of Biological Sciences, Florida International University, Miami, Florida, 33199, USA.

ABSTRACT

Understanding gene regulation is a key step to investigating gene functions and their relationships. Many algorithms have been developed to discover transcription factor binding sites (TFBS); they are predominantly located in upstream regions of genes and contribute to transcription regulation if they are bound by a specific transcription factor. However, traditional methods focusing on finding motifs have shortcomings, which can be overcome by using comparative genomics data that is now increasingly available. Traditional methods to score motifs also have their limitations. In this paper, we propose a new algorithm called IEM to refine motifs using comparative genomics data. We show the effectiveness of our techniques with several data sets. Two sets of experiments were performed with comparative genomics data on five strains of P. aeruginosa. One set of experiments were performed with similar data on four species of yeast. The weighted conservation score proposed in this paper is an improvement over existing motif scores.

Keyword: Comparative Genomics, Motif, EM algorithm

1 INTRODUCTION

Gene expression is a fundamental biological process. The first step in this process called *transcription* transmits genetic information from DNA to messenger RNA (mRNA). A transcription factor (TF) is a protein that regulates transcription of a gene by interacting with specific short DNA sequences, located often in the upstream region of the regulated genes. Such short DNA sequences are called transcription factor binding sites (TFBS) or regulatory elements. The regulatory elements can be described as sequence signatures and will be referred to in this paper as *motifs*. One TF can regulate a large set of genes, and a single gene may be regulated by the combination of several TFs. The upstream region of each gene regulated by the same TF must have at least one binding site specific for that particular TF. These binding sites must be specific enough so that the TF can “recognize” them and bind to them. However, it is well known that different sites bound by the same TF are not necessarily identical. The computational challenge is to find these sites and to succinctly and accurately describe all such binding sites.

The simplest way to describe a binding site is to write down its consensus sequence. However, this is very imprecise and does not do justice to the complexity of the sequence signature. A sequence alignment of all known binding sites captures its complexity, but is not succinct enough. A logo format (Schneider and Stephens 1990; Crooks, Hon et al. 2004) is succinct enough, but is merely visual. The appropriate description is a *profile*, which is also referred to as a position-specific scoring matrix (PSSM) or a position weight matrix (PWM) (Werner 1999; Stormo 2000). A profile is a 4 × K matrix (K is the length of the binding site) whose entries give a measure of the preference of a base appearing at any given position.

Examples of sophisticated algorithms to identify TF binding sites include MEME (Bailey and Elkan 1994), AlignACE (Hertz and Stormo 1999), Bioprospector (Liu, Brutlag et al. 2001), MDscan

[*] To whom correspondence should be addressed.

(Liu, Brutlag et al. 2002), YMF (Sinha and Tompa 2003), Weeder (Pavesi, Mereghetti et al. 2004) and many more. All these methods attempt to find sequence signatures that are significantly overrepresented in the upstream regions of a given gene set (typically a cluster of co-regulated genes from analyzing microarray data, or a gene set inferred from a ChIP-Chip experiment) when compared to an appropriately chosen background.

Despite the successful application of the algorithms listed above, each of them has certain limitations (Hu, Li et al. 2005; Tompa, Li et al. 2005; GuhaThakurta 2006; MacIsaac and Fraenkel 2006; Sandve and Drablos 2006). First, all these methods are prone to predict a large number of motifs, many of which are false-positives, partly because TFs show remarkable flexibility in the binding sites they can potentially bind to. Second, all these methods report statistically over-represented motifs. However, statistical significance of motifs need not be synonymous with biological relevance of motifs. Binding of TFs to their binding sites is a complex process and may be assisted or hindered by many other unexplained factors.

Comparative genomics data is a promising new source of information that can help to improve motif prediction. With the availability of an increasing number of whole genome sequences of evolutionarily-related genomes, it is practical to incorporate the comparative genomics data into the motif discovery process. The basic assumption is that transcription factors and transcriptional mechanisms involved in fundamental cellular processes are likely to be conserved among evolutionary-related genomes. Consequently, the binding sites for such TFs are also likely to be conserved. Therefore, availability of comparative genomics data is likely to provide additional support to the predictions of binding sites. The simplest way to deal with data on additional genomes is to pool together the upstream regions of all available genomes and to apply traditional motif detection methods. However, this is not an optimal utilization of the comparative genomics data. The "phylogenetic footprinting" strategy is a sophisticated method used to find motifs that are conserved for a particular gene across related organisms (Blanchette and Tompa 2002). Several subtle approaches such as PhyloCon (Wang and Stormo 2003), orthoMEME (Prakash, Blanchette et al. 2004), CompareProspector (Liu, Liu et al. 2004), EMnEM (Moses, Chiang et al. 2004), PhyME (Sinha, Blanchette et al. 2004), and PhyloGibbs (Siddharthan, Siggia et al. 2005) were developed recently to solve this problem. In these approaches, either an EM-based algorithm, a greedy algorithm or a Gibbs Sampling strategy was applied to optimize an objective function, while taking the phylogenetic relationships into account. The main problem with these methods is that phylogenetic relationships are often not easy to infer and not very reliable. Also, any motif that is unique to particular genomes or in upstream regions of genes with no orthologs in some related genomes will not be detected. Most of above methods also need an alignment of the input sequence. Like phylogenetic relationships, alignments are also often unreliable. Inaccurate alignments (or phylogeneties) lead to errors in profile matrices, and ultimately in motif prediction.

Another challenge in motif prediction is to develop scoring functions that reflect biological significance. Several popular scoring functions include IC (information content), MAP, Group Specificity score, LLBG (least likely under the background model) and Bayesian scoring function. However, as explained earlier, algorithms that use these scoring schemes end up with a large number of false positives in their predictions. When dealing with multiple genomes, the degree of conservation of the 'hits' of a profile across the many genomes can be used as a crude surrogate for the significance of the motif. However, this metric has its shortcomings. In this paper, we propose a metric to measure such biological significance.

In this paper, we propose a new algorithm called *IEM* (**I**teratively **E**nhancing **M**otif **D**iscovery). *IEM* is an iterative version of an earlier algorithm called *EMR* (Enhancing Motif Refinement) (Zeng and Narasimhan 2007). It differs from other earlier approaches in that no attempt is made to perform *de novo* detection of motifs (although that would be easy to incorporate). Instead, comparative genomics data is used to "enhance" any given motif. These motifs may have been discovered by other computational methods, or may have been identified by laboratory techniques. Thus our method leverages the best-known motif discovery methods, or utilizes the (potentially incomplete) knowledge of previous studies while incorporating newly available comparative genomics data.

The research described here is significant for the following reasons. First, there is a clear need to reduce the number of false positives predicted by traditional tools. Second, our method can make use of partial information (on one or more binding sites), which may be available as a result of biological experiments. Third, with the availability of high throughput gene expression techniques like Microarrays and ChIP-Chip experiments, it is possible to get sets of co-expressed genes involved in the same metabolic pathway (and, therefore, potentially coregulated). Finally, our results show that the IEM algorithm has superior ability to overcome the shortcoming of previous methods and to effectively utilize any available comparative genomics data.

2 METHODS

2.1 Algorithm

The *IEM* algorithm takes as input an "unrefined" motif for a given genome Γ_1 (called the reference genome); this motif may have been generated using any reasonable existing motif detection method. Alternatively, the input could be a known binding site or a crude approximation based loosely on some experiments. Using one or more additional genomes Γ_2 (referred to as the *related* genomes), and the corresponding orthology information between Γ_1 and Γ_2, the algorithm returns an enhanced motif. The refinement procedure is EM-based, as described below in Section 2.1.3.

2.1.1 Basic Expectation Maximization (EM) Algorithm Since our algorithm is EM-based, we first present an adaptation of the classical EM algorithm (Dempster, Laird et al. 1977) for *ab initio* motif discovery (Lawrence and Reilly 1990). Motif prediction can be thought of as a parameter estimation process for a mixture model: (1) a model for the motif and (2) a model for the background. Roughly speaking, the algorithm can be described as follows: In the (Expectation) E-Step, for every site, the likelihood that it belongs to either model of the mixture is computed. And, in the (Maximization) M-Step, a set of parameters (i.e., the entries of the profile) for the individual models (motif model and background model) are recomputed using the likelihood values computed in the E-step as weights in the calculation. Upon convergence, we end up with two models: one for motif and one for background. We randomly initialize parameters for the motif model (by randomly choosing the locations of the binding sites), and then the E-step and M-step are iterated until convergence.

2.1.2 Improvements in MEME: The original version of EM as proposed by Lawrence and Relly (Lawrence and Reilly 1990) suffers from several limitations. For example, it does not state how to choose a starting point: It assumes that each sequence in the dataset contains exactly one occurrence of the motif; it also assumes that there is only one instance of the motif in each upstream region and does not attempt to find multiple instances. Bailey and Elkan proposed a modified EM method called MEME to eliminate these limitations (Bailey and Elkan 1994). Their method used sequences from the input as random start points. The method allows multiple instances of a motif in one upstream region. Furthermore, once the algorithm converges upon a motif, it is eliminated from consideration and then the algorithm restarts to look for other motifs.

MEME works reasonably well on many data sets, and is widely used. However, it has shortcomings. First, even though it choses a start point form among the subsequences of the input sequence, it may not converge upon a desired motif. Thus, it is not suitable for finding motifs for which we may know partial information. Second, the only way it can deal with comparative genomics data is by merely pooling the input sequences from multiple genomes. However, as mentioned before, this leaves the comparative genomics data underutilized. Our proposed IEM method considers comparative genomic data in a "dual" manner.

2.1.3 IEM Algorithm The IEM algorithm is described below in Figure 1. Assume the input consists of profile $M_1 = (m_{ij})$, which is a 4 × K matrix. K is the length of the motif and m_{ij} is the entry in the i^{th} row and j^{th} column of M_1. Let the indicator variable matrix be defined as $Z = (z_{pq})$: where $z_{pq} = 1$, if an instance of the motif starts from p^{th} position in the upstream region of the q^{th} gene, and is equal to 0 otherwise. These indicator variables approximate the probability that a specific site (i.e., the sequence starting from the p^{th} position in the upstream region of the q^{th} gene) is a binding site according to the profile matrix. The IEM algorithm estimates the indicator variable matrix Z_1 and profile matrix M_1 in the reference genome and the indicator variable matrix Z_2 and profile M_2 in the related genomes iteratively. The estimation process is similar to that in MEME (Bailey and Elkan 1994). However, in IEM a dual-step estimation is applied by incorporating comparative genomics data. Given indicator variable z_{pq} in one data source (either the reference genome or the related genomes) and a motif model (i.e., profile matrix) M for the entire data set (merged from M_1 and M_2), we can calculate the probability of observing a given upstream region U_q as follows:

$$P(U_q \mid Z_{pq}, M) = \prod_{i=1}^{l-nk+1} m_{a0} \prod_{j=1}^{nk} m_{aj}\,, \qquad (1)$$

where m_{a0} is background frequency for base a, m_{aj} is frequency for base a at position j in the motif model, k is the motif length, n is number of 1s in Z_{pq}, and l is the length of upstream sequence. Then by Bayes' rule, we can calculate the probability that the site at position p in upstream region q is a binding site as follows:

$$P(Z_{pq} \mid M, U_q) = \frac{P(U_q \mid Z_{pq}, M)}{\sum_{r=1}^{l-k+1} P(U_q \mid Z_{rq}, M)} \qquad (2)$$

Intuitively, the *IEM* algorithm tries to refine a motif in each iteration in two successive EM steps. In each step, it computes the likelihood for each site in one data set over a model M (not merely M_1 or M_2), which is arrived at by the previous maximization step applied over all the data sets. Comin *et al.* reported a subtle motif discovery method using a similar two-step strategy (Comin and Parida 2007). The differences are twofold. First, we incorporate comparative genomics data, and second, we use profiles instead of consensus sequences to represent the motifs.

```
Input: a) Profile $M_1$, motif length $l$, and associ-
          ated gene set $G_1$ from genome $\Gamma_1$
       b) upstream sequences of the ORFs in $G_1$
       c) Additional genome(s) $\Gamma_2$,.and the
          orthology map for all the genomes
       d) upstream sequences of the ORFs in $G_2$,
          the orthologs of $G_1$ in $\Gamma_2$
Output: Refined motif weight matrix Mr
Algorithm:
1.   Estimate $Z_2$ in $G_2$ from $M_1$.
while (not converged) do
2.   Re-estimate $M_2$ in $G_2$ from $Z_2$.
3.   $M$ = merge($M_1$ , $M_2$)
4.   Re-estimate $Z_1$ in $G_1$ from $M$.
5.   Re-estimate $M_1$ in $G_1$ from $Z_1$.
6.   $M$ = merge($M_1$ , $M_2$)
7.   Re-estimate $Z_2$ in $G_2$ from $M$.
endwhile
9.   Return $M_2$.
```

Figure 1. *IEM* Algorithm

In summary, IEM algorithm does the following 4 steps iteratively:

1. In the first E step, the probabilities that each site in the reference genome belongs to the profile M_1 are computed by using formula (2).
2. In the first M step, the new profile M_1 is estimated by using every (indicated) binding site in the reference genome (i.e., weighted with Z_{pq}). Profile M is updated using the new sites.
3. In the second E step, the probabilities that each site in the related genomes belong to the profile M_2 are computed by using formula (2).

4. In the second M step, the new profile M_2 is estimated by using every (indicated) binding site in the related genomes (i.e., weighted with Z_{pq}). Profile M is updated by using the new sites.

The "merge" operation mentioned in the algorithm is achieved by creating the profile matrix from the instances of the sites with indicator value 1 from all the genomes. Note that a generalization of the merging step is possible where the sites are weighted by the probability of that site belonging to a model (i.e., its score against the profile).

2.2 Evaluation Approaches

Evaluation of the IEM algorithm is a nontrivial task because very little experimentally verified data is available. Even the available experimentally verified data is often only partial information. In one of the experiments described below, we consider the critical regulation activities in the arginine metabolic pathways in the bacterium *P. aeruginosa* (PAO1). We show that our algorithm, with the help of the complete genomes of six strains of *P. aeruginosa*, produces refined motifs with improved accuracy (see the Results section for details). The performance in such cases can be measured in terms of true positives and false positives from the available partial information. Here the true positives measure indicates the number of known binding sites that are predicted, while the false positives are the number of known non-binding sites that are predicted.

In another experiment, where no experimentally verified data was available, we have proposed two approaches to evaluate our results. One approach is to investigate the functional enrichment of the genes whose upstream regions have a predicted binding site. Using gene ontology analysis, we observed that the terms that were enriched were closely related to what is known about the regulator.

Another approach is to compute meaningful measures of motif scores. Traditional ones such as MAP and IC scores are not well-suited for comparative genomics data. A better approach is to use scores based on how well the predicted binding site is conserved across all the genomes under consideration. The simplest measure along these lines is what we will refer to as the *conservation score*. It is the average number of genomes in which any given predicted binding site occurs simultaneously in the upstream sequences of orthologous genes. This value ranges between 0 and m, where m is the number of genomes (besides the reference genome) being analyzed. Such a measure was proposed earlier (Gertz, Riles et al. 2005). Let m denote the number of genomes (besides the reference genome) being considered. Let n be the total number of genes in the reference genome whose upstream sequence has at least one predicted site of the motif, and let s_i be the number of genomes in which the ortholog of gene i contains a site in its upstream region. Then the conservation score S is defined as:

$$S = \sum_{i=1}^{n} \frac{s_i}{n} \qquad (3)$$

The weakness of this conservation score is that it does not account for some key facts. In the following discussion, let A and B be two predicted motifs with the same conservation score, i.e., same average hits per genome.

(1) If A has more instances than B in which s_i equals to m, it should be considered more significant.

(2) If A has more hits than B in the reference genome, then it should be considered more significant.

To overcome the above disadvantages, we propose a new score, which we refer to as the weighted conservation score. It is given as:

$$S_c = \log[mn] \frac{\sum_{i=1}^{m} i w_i n_i}{n \sum_{i=1}^{m} w_i}, \; w_i > w_{i-1} \quad \forall i, \qquad (4)$$

where m is the number of genomes being considered, n is the number of genes in the reference genome whose upstream regions contain at least one instance of the predicted motif, n_i is the number of genes that has i number of genomes in which the corresponding ortholog contains at least one instance of the motif in its upstream region, and w_i is a suitable weight constant that satisfies $w_i > w_{i-1}$ for all i, implying that if a motif instance occurs in more orthologs then it should be weighted higher. w_i is chosen to be i in following example.

We highlight the differences between the conservation score and the weighted conservation score using simple examples. In Figure 2, motifs A and B have the same conservation score. Unlike motif B, motif A has instances across all related genomes in the upstream regions of three orthologous gene sets. We argue that motif A is more conserved than motif B. The weighted conservation score reflects this intuition. Motif C, with the same conservation score as motif D, has more instances in the reference genome, which may indicate a more important biological role. The weighted conservation score rewards motifs A and C.

3 RESULTS

3.1 Results on the arginine metabolic pathway study

Metabolic pathways have been widely studied. They can be extremely complex, and may involve large numbers of genes. Often every path in the network involves one or more TFs and the genes regulated by them. However, only a few of genes and TFs in the pathways may have been identified, and even fewer of the TF binding sites may be known. A useful problem is to identify the genes and TFs and their binding sites specifically involved in a specific pathway. Starting from one or two experimentally verified binding sites, can we predict the rest of the relevant binding sites of the genes in the pathway? Furthermore, can we identify such a gene set? We will show that our IEM algorithm can help to address these questions.

In order to evaluate our results, we used a well studied pathway - the arginine metabolic pathway in *P. aeruginosa*, as an example. It is already known that *P. aeruginosa* possesses four different pathways for utilization of arginine (Lu, Yang et al. 2004): the arginine deiminase (ADI) pathway, the arginine succinyltransferase (AST) pathway, the arginine decarboxylase (ADC) pathway, and the arginine dehydrogenase (ADH) pathway. Under anaerobic conditions, arginine can be used as a direct source of ATP via the ADI pathway. ArgR is a TF in the ADH pathway. Lu *et al.* used microarray experiments to identify candidate genes for the ArgR regulon (Lu, Yang et al. 2004). It was reported that ArgR regulated 37 (28 induced and 9 repressed) genes from 17 operons. Eighteen of the 28 arginine-inducible genes are in 4 transcriptional units that have been reported previously as members of the ArgR regulon (Itoh 1997; Park, Lu et al. 1997; Nishijyo, Park et al. 1998; Lu, Winteler et al. 1999; Lu and Abdelal 2001; Hashim, Kwon et al. 2004). Lu *et al.* also identified several new ArgR regulon members among these 37 genes, and verified them by wet lab experiments. Since the ArgR system is well studied, we used it to test the IEM algorithm.

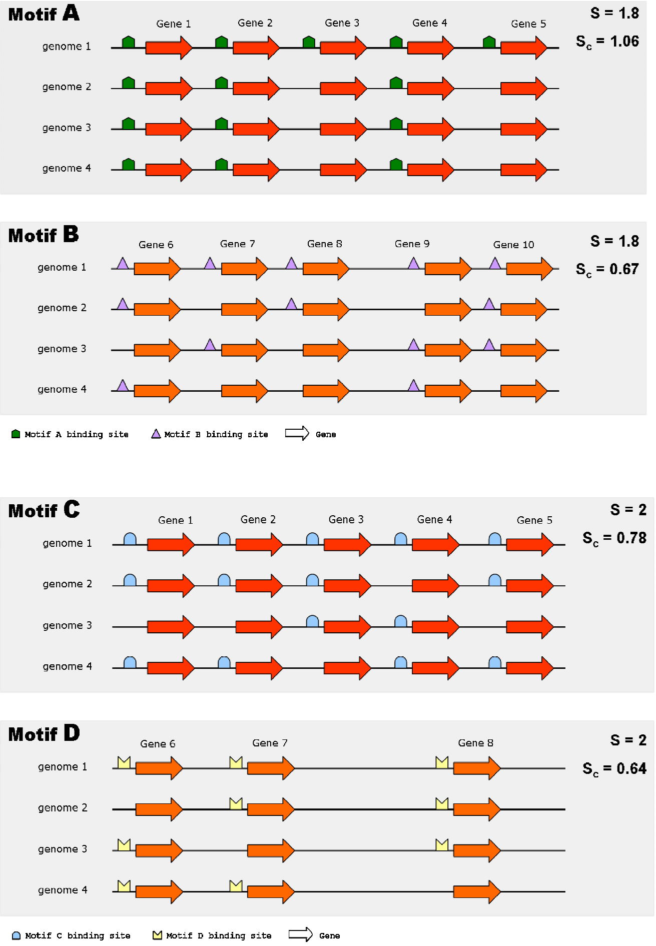


Figure 2. Shown are examples that highlight the differences between the conservation score, S, and the weighted conservation scores, S_c.

3.1.1 Arginine pathway data set Upstream regions of the 17 transcriptional units (operons) were obtained for five strains of *P. aeruginosa* (PAO1, PA14, PACS2, PA2192, and PA3719). We also included 6 genes involved in the ADC pathway and the ADH pathways that were known not to bind to ArgR.

3.1.2 Prediction Comparison Procedure To show the power of our technique, we assumed for our experiments that we know only one (randomly chosen) instance of a binding site for ArgR. We used a subset of the operons mentioned above (12 out of 17 from ADI pathways and all 6 from ADC/ADH pathways). We then set out to see if the algorithm successful in locating previously known binding sites in the remaining 5 operons. On an average the refined motif missed 1.2 of the 5 known binding sites.

We applied MEME, AlignACE, and IEM to the same data set. The results were compared for an experiment with data from two genomes (PAO1 and PA14) and another experiment with data from five genomes (PAO1, PA14, PACS2, PA2192, and PAC3719). The idea was to get a sense of how much the comparative genomics data helped in the task. MEME and AlignACE were applied to the pooled data. For IEM, the initial profile was created using the motif instance. The frequency of the base from the consensus sequence was set at 0.7, and the frequencies of other bases were set at 0.1. Each of the three programs was run 10 times for the data set introduced earlier. We counted the number of true predictions (TP, True Positives), the number of false predictions (FP, False Positives) and the motif scores IC (Information Content), MAP (maximum *a posteriori* probability) and the weighted conservation scores S_c.

3.1.3 Arginine pathway prediction comparisons results Tables 1 and 2 present the results from two experiments (two genome case vs five genome case) for the 10 runs. The three columns present the results with the three programs. In cases where a motif was reported, the number of TPs and FPs along with three measures of quality of the motif are reported. The IEM algorithm finds the ArgR binding motif in every instance. In the experiments involving two genomes, the motif AlignACE. However, when four genomes were

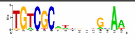
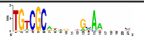
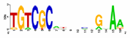
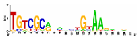
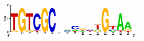
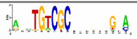
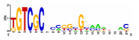
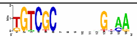
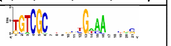
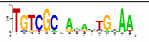
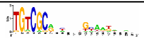
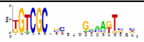
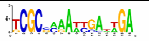
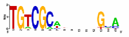
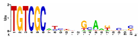
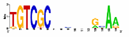
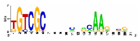
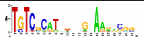
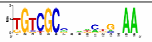
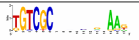
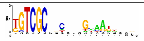
	MEME (TP, FP, IC, MAP, S_c)	AlignACE (TP, FP, IC, MAP, S_c)	IEMD (TP, FP, IC, MAP, S_c)
1	no report	 (11, 4, 7.72, 1.64, 2.30)	 (13, 5, 7.90, 1.34, 2.30)
2	no rcport	 (11, 4, 7.52, 1.60, 2.30)	 (7, 4, 7.23, 1.09, 1.79)
3	 (5, 1, 7.17, 1.11, 1.20)	 (10, 8, 7.19, 1,44, 2.39)	 (3, 5, 7.27, 0.96, 1.79)
4	no report	 (8, 8, 7.12, 1.54, 2.39)	 (7, 3, 7.78, 1.11, 1.79)
5	 (4, 1, 7.60, 1.09, 0.64)	no rcport	 (6, 7, 6.98, 1.08, 2.08)
6	no report	no report	 (5, 4, 7.38, 1.02, 1.95)
7	 (2, 3, 7.92, 1.09, 1.29)	 (10, 10, 7.08, 1.74, 2.07)	 (9, 5, 8.05, 1.28, 1.95)
8	no rcport	 (13, 4, 7.50, 1.65, 2.39)	 (9, 7, 7.21, 1.17, 2.20)
9	no report	no report	 (9, 2, 7.95, 1.17, 1.95)
10	 (5, 1, 6.97, 1.15, 1.50)	 (10, 10, 7.22, 1.57, 2.56)	 (7, 6, 7.08, 1.10, 1.95)

Table 1 Motif predicted by IEM, MEME, and AlignACE using data on 2 strains of *P. aeruginosa* (PA01and PA14).

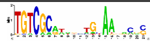
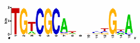
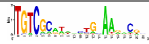
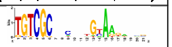
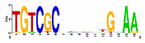
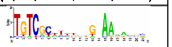
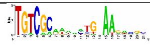
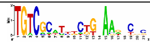
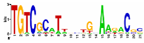
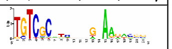
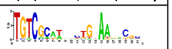
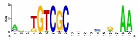
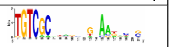
	MEME (TP, FP, IC, MAP, S_c)	AlignACE (TP, FP, IC, MAP, S_c)	IEMD (TP, FP, IC, MAP, S_c)
1	no rcport	no rcport	 (10, 3, 9.65, 1.86, 3.33)
2	no rcport	 (5, 1, 8.46, 2.11, 2.40)	 (11, 2, 9.59, 1.85, 3.46)
3	no report	no report	 (9, 4, 8.89, 1.73, 3.33)
4	no report	 (2, 3, 8.42, 1.95, 2.52)	 (12, 10, 8.55, 1.83, 3.30)
5	no report	no report	 (8, 2, 9.78, 1.78, 3.56)
6	no report	no report	 (8, 2, 9.80, 1.79, 3.07)
7	no report	no report	 (9, 2, 10.13, 1.91, 3.56)
8	no report	no report	 (17, 6, 8.68, 1.98, 3.61)
9	no report	no report	 (10, 2, 9.66, 1.83, 3.56)
10	no rcport	 (2, 10, 9.33, 0.83, 2.61)	 (11, 5, 8.76, 1.80, 3.09)

Table 2 Motif predicted by IEM, MEME, and AlignACE using data on 5 strains of *P. aeruginosa* (PA01, PA14, PA2192, PACS2, and PAC3719).

used, the scores using the IEM algorithm was markedly superior to those with the other two methods (when they were reported).

3.2 Results on ampR

In this section, we discuss our experiments with the IEM algorithm applied to data from experiments on the transcription factor, AmpR, in *P. aeruginosa*. AmpR was recently reported as a global transcription factor that regulates the expression of many virulence factors (Kong, Jayawardena et al. 2005). To better understand the regulon of AmpR, the consensus sequence (5'-TCTGCTGCAAATTT-3') of AmpR binding sites in *C. freundii* and *E. cloacae* was used by Kong et al. to find an exactly conserved sequence site within the upstream region of ampC in PAO1 (Kong, Jayawardena et al. 2005). They also analyzed the upstream regions of all the genes putatively regulated by AmpR with the hope of finding a potential AmpR binding site. Tools such as MEME and AlignACE failed to find anything resembling the binding site from the upstream region of ampC..

The IEM algorithm was then applied using the consensus sequence mentioned above, a potential hand-crafted list of 10 genes possibly regulated by AmpR, and newly available comparative genomics data sets from four closely related strains of *Pseudomonas* (PA14, PA2192, PACS2, and PAC3719). As mentioned in the previous section, a crude motif profile was constructed based on the consensus sequence. The results before and after applying the IEM algorithm are shown in Table 3. The refined motif showed improved scores according to three different motif scores. After refinement, we found that putative AmpR binding site appears only in 3 of the 10 genes mentioned above (*lasA*, *lasR*, and *ampC*) across all five strains of *P. aeruginosa*. Support for these 3 predictions was obtained using lacZ fusions in the Mathee lab. Further experimental verification is needed and work is underway in the Mathee lab. We conjecture that the remaining 7 genes are only indirectly regulated by AmpR.

We then used the refined motif to scan the entire PAO1 genome for instances of the motif in the upstream regions. Based on the likelihood value calculated in formula (2), we ranked the "hits" and chose the top 150 genes and followed it up with gene function enrichment analysis. See Table 4 for the results. The term with the top hit, i.e., the lowest P-value was "periplasmic space". This is considered significant because, ampR is known to be involved in cell-wall recycling. A similar search with the motif before refinement did not find this GO-term.

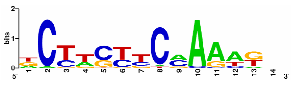
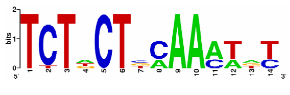
Consensus Sequence (5'-TCTGCTCCAAATTT-3')	
Motif before refinement (IC,MAP,Sc)	Motif After refinement (IC,MAP,Sc)
 (4.57, 1.20, 2.84)	 (8.26, 1.60, 2.98)

Table 3 Characteristics of motif before and after refinement

3.3 Results on whole genomic data

Next we discuss our experiments with yeast data sets. Recently, Kellis et al. compared five yeast species to identify regulatory elements in the entire genome by searching for conserved segments across different yeast species (Kellis, Patterson et al. 2003). They developed a motif score called MCS (Motif Conservation Score) to measure the conservation ratio of a motif compared to the random patterns of the same length and degeneracy (Kellis, Patterson et al. 2003). A list of 72 full motifs having MCS at least 4 was reported. These 72 predicted motifs showed strong overlap with 28 of the 33 known motifs in yeast. However, the motifs used in the paper were represented using generalized consensus sequences (i.e., using IUPAC codes to represent nucleotide degeneracy) instead of the more powerful profile matrix. We set out to consider whether the IEM algorithm could improve the predictions from that work.

Starting from the results of Kellis et al., we used IEM to refine each of the 72 motifs mentioned above. Data from four yeast genomes (*S. cerevisiae*, *S. paradoxus*, *S. mikatae* and *S. bayanus*) were used. Complete results on the refined motifs are available at our supplementary results website:

[http://biorg.cs.fiu.edu/IEM/]. Below we show some of the highlights in Table 5. In each case the number of hits went down after the refinement.

Category Name	Total	Enriched	P-Value	GO ID
periplasmic space	3	2	0.0085	GO:0042597
periplasmic space (sensu Proteobateria)	3	2	0.0085	GO:0030288
3-dehydroquinate dehydratase activity	3	2	0.0085	GO:0003855
nucleotide kinase activity	4	2	0.0164	GO:0019201
phosphotransferase activity,phosphate group as acceptor	5	2	0.0264	GO:0016776
nucleobase,nucleoside,nucleotide and nucleic acid metabolism	290	23	0.0273	GO:0006139
response to stimulus	152	14	0.0293	GO:0050896
transcription	97	10	0.0328	GO:0006350
purine nucleotide binding	14	3	0.0367	GO:0017076
primary metabolism	671	43	0.04	GO:0044238

Table 4: Go enrichment analysis for the AmpR experiments.

4 DISCUSSION AND CONCLUSIONS

In this paper we propose a new algorithm to refine motifs with the help of comparative genomics data. The algorithm incorporates an improved scoring scheme that is sensitive to hits in the related genomes. The algorithm is inspired by the technique of "co-training" from the field of data mining, where lessons learnt from one data source is iteratively used to model the situation for another data source. The results show clear improvements in the quality of the motifs output.

The IEM algorithm does have its own shortcomings, which we continue to improve. First, it does not attempt to change the length of the motif from the initial motif it started with. Second, it works best if the genomes considered are very closely related and is useful in cases where the phylogenetic relationships between the genomes are not known. If phylogentic information is available, then the algorithm can be modified to factor this in, along the lines of several previous algorithms.

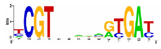
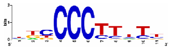
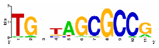
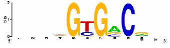
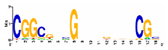
Motif Number	Motif	# of ORFs	Motif Score (IC, MAP)
1	YCGTnnnnnRYGAY 	796 668	1.89, 0.40 9.83, 5.61
29	hRCCCYTWDt 	442 284	1.93, 0.53 6.78, 4.80
42	TGnKAGCGCCG 	72 60	2.49, 0.67 7.76, 3.98
51	WGTGACg 	202 180	1.90, 0.52 4.90, 3.50
57	CGGCnnMGnnnnnnnCGC 	84 52	2.03, 0.34 5.68, 1.81

Table 5 Results of motif refinement for the yeast data set. For each of the five motifs, the upper row is the consensus sequence from Kellis et al., while the lower row is the result after refinement by the IEM algorithm.

ACKNOWLEDGMENTS

The work of GN was supported in part by a grant from NIH under NIH/NIGMS S06 GM008205. We thank Camilo Valdes for helping us compile the upstream sequence data for the five strains of P. aeruginosa and for his help with Figure 2 in the paper.

REFERENCES

Bailey, T. L. and C. Elkan (1994). "Fitting a mixture model by expectation maximization to discover motifs in biopolymers." Proc Int Conf Intell Syst Mol Biol **2**: 28-36.

Blanchette, M. and M. Tompa (2002). "Discovery of regulatory elements by a computational method for phylogenetic footprinting." Genome Res **12**(5): 739-48.

Comin, M. and L. Parida (2007). Subtle Motif Discovery for Detection of DNA regulatory sites. Asia Pacific Bioinformatics Conference (APBC2007), Hong Kong.

Crooks, G. E., G. Hon, et al. (2004). "WebLogo: A sequence logo generator." Genome Research **14**(6): 1188-1190.

Dempster, A. P., N. M. Laird, et al. (1977). "Maximum likelihood estimation from incomplete data via the EM algorithm." J. R.Statist. Soc. B **39**: 1-38.

Gertz, J., L. Riles, et al. (2005). "Discovery, validation, and genetic dissection of transcription factor binding sites by comparative and functional genomics." Genome Res **15**(8): 1145-52.

GuhaThakurta, D. (2006). "Computational identification of transcriptional regulatory elements in DNA sequence." Nucleic Acids Res **34**(12): 3585-98.

Hashim, S., D. H. Kwon, et al. (2004). "The arginine regulatory protein mediates repression by arginine of the operons encoding glutamate synthase and anabolic glutamate dehydrogenase in Pseudomonas aeruginosa." J Bacteriol **186**(12): 3848-54.

Hertz, G. Z. and G. D. Stormo (1999). "Identifying DNA and protein patterns with statistically significant alignments of multiple sequences." Bioinformatics **15**(7-8): 563-77.

Hu, J., B. Li, et al. (2005). "Limitations and potentials of current motif discovery algorithms." Nucleic Acids Res **33**(15): 4899-913.

Itoh, Y. (1997). "Cloning and characterization of the aru genes encoding enzymes of the catabolic arginine succinyltransferase pathway in Pseudomonas aeruginosa." J Bacteriol **179**(23): 7280-90.

Kellis, M., N. Patterson, et al. (2003). "Sequencing and comparison of yeast species to identify genes and regulatory elements." Nature **423**(6937): 241-54.

Kong, K. F., S. R. Jayawardena, et al. (2005). "Pseudomonas aeruginosa AmpR is a global transcriptional factor that regulates expression of AmpC and PoxB beta-lactamases, proteases, quorum sensing, and other virulence factors." Antimicrob Agents Chemother **49**(11): 4567-75.

Lawrence, C. E. and A. A. Reilly (1990). "An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences." Proteins **7**(1): 41-51.

Liu, X., D. L. Brutlag, et al. (2001). "BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes." Pac Symp Biocomput: 127-38.

Liu, X. S., D. L. Brutlag, et al. (2002). "An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments." Nat Biotechnol **20**(8): 835-9.

Liu, Y., X. S. Liu, et al. (2004). "Eukaryotic regulatory element conservation analysis and identification using comparative genomics." Genome Res **14**(3): 451-8.

Lu, C. D. and A. T. Abdelal (2001). "The gdhB gene of Pseudomonas aeruginosa encodes an arginine-inducible NAD(+)-dependent glutamate dehydrogenase which is subject to allosteric regulation." J Bacteriol **183**(2): 490-9.

Lu, C. D., H. Winteler, et al. (1999). "The ArgR regulatory protein, a helper to the anaerobic regulator ANR during transcriptional activation of the arcD promoter in Pseudomonas aeruginosa." J Bacteriol **181**(8): 2459-64.

Lu, C. D., Z. Yang, et al. (2004). "Transcriptome analysis of the ArgR regulon in Pseudomonas aeruginosa." J Bacteriol **186**(12): 3855-61.

MacIsaac, K. D. and E. Fraenkel (2006). "Practical strategies for discovering regulatory DNA sequence motifs." PLoS Comput Biol **2**(4): e36.

Moses, A. M., D. Y. Chiang, et al. (2004). "Phylogenetic motif detection by expectation-maximization on evolutionary mixtures." Pac Symp Biocomput: 324-35.

Nishijyo, T., S. M. Park, et al. (1998). "Molecular characterization and regulation of an operon encoding a system for transport of arginine and ornithine and the ArgR regulatory protein in Pseudomonas aeruginosa." J Bacteriol **180**(21): 5559-66.

Park, S. M., C. D. Lu, et al. (1997). "Cloning and characterization of argR, a gene that participates in regulation of arginine biosynthesis and catabolism in Pseudomonas aeruginosa PAO1." J Bacteriol **179**(17): 5300-8.

Pavesi, G., P. Mereghetti, et al. (2004). "Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes." Nucleic Acids Res **32**(Web Server issue): W199-203.

Prakash, A., M. Blanchette, et al. (2004). "Motif discovery in heterogeneous sequence data." Pac Symp Biocomput: 348-59.

Sandve, G. K. and F. Drablos (2006). "A survey of motif discovery methods in an integrated framework." Biol Direct **1**: 11.

Schneider, T. D. and R. M. Stephens (1990). "Sequence logos: a new way to display consensus sequences." Nucleic Acids Res **18**(20): 6097-100.

Siddharthan, R., E. D. Siggia, et al. (2005). "PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny." PLoS Comput Biol **1**(7): e67.

Sinha, S., M. Blanchette, et al. (2004). "PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences." BMC Bioinformatics **5**: 170.

Sinha, S. and M. Tompa (2003). "YMF: A program for discovery of novel transcription factor binding sites by statistical overrepresentation." Nucleic Acids Res **31**(13): 3586-8.

Stormo, G. D. (2000). "DNA binding sites: representation and discovery." Bioinformatics **16**(1): 16-23.

Tompa, M., N. Li, et al. (2005). "Assessing computational tools for the discovery of transcription factor binding sites." Nat Biotechnol **23**(1): 137-44.

Wang, T. and G. D. Stormo (2003). "Combining phylogenetic data with co-regulated genes to identify regulatory motifs." Bioinformatics **19**(18): 2369-80.

Werner, T. (1999). "Models for prediction and recognition of eukaryotic promoters." Mamm Genome **10**(2): 168-75.

Zeng, E. and G. Narasomhan (2007). "Enhancing motif refinement by incorporating comparative genomic data." ISBRA: To Appear.