Motif Evaluation by Leave-one-out Scoring

Audrey Girouard, Noah W. Smith and Donna K. Slonim
Department of Computer Science
Tufts University
Medford, MA 02155, USA
Email: {audrey.girouard, noah.smith, donna.slonim@tufts.edu}

***Abstract*— We propose a new method for collecting information on regulatory elements found by any motif discovery program. We suggest that combining the results of n leave-one-out motif discovery runs provides additional information. By examining motifs found in $n-1$ of the sequences and scoring them on the remaining sequence, we overcome some of the issues arising from noisy data to identify more high-quality motifs.**

We describe preliminary investigations of this approach, using MEME for motif discovery. We show that the Leave-one-out method highlights different motifs than a single MEME run would. We demonstrate that our method increases the power of small datasets. We also explore how the information gain of the method changes as the number of sequences increases. Our approach may be generalized to any number of sequences, and may be applied with any motif-inference package that generates a final population of solutions and scores.

I. Introduction

This paper introduces an alternative system for ranking the results of software packages that identify common regulatory elements in sets of sequences. The question of whether several related runs of a system analyzed together could provide more information about the overall strength of a motif than a single such run was examined. This work is a preliminary exploration of the effects of using a Leave-one-out Scoring strategy inspired by n-way cross-validation, similar to the Jackknife approach used by statisticians to improve error estimation [15], [6]. By observing the changes caused by adding or removing a sequence, we are able to further evaluate the motifs found.

Building upon existing motif discovery systems, Leave-one-out Scoring provides a new method of evaluating common regulatory elements. Here, we describe preliminary investigations of this approach, using MEME for motif discovery. We study the behavior of Leave-one-out Scoring on four sequence clusters that are expected to share common regulatory elements: two sets of orthologous promoter regions spanning a wide range of organisms, and two sets upstream of *C. elegans* genes tightly co-expressed in multiple time series.

We show that the collection of motifs found by combining multple Leave-one-out runs is larger than the set found by a single traditional run, but not hugely so, reflecting substantial overlap of the motifs found in each run. The intuition behind the method is that the overlap information provides more information about the motifs than their original score in a single set of sequences. Overall, we are interested in whether this method provides sufficient new information to justify the added expense of computation.

II. Background

A DNA sequence can be represented as a string over the four letter language of nucleotides $\{A, C, G, T\}$, and genes can (simplistically) be viewed as substrings of a DNA sequence. These substrings are used by living cells as the blueprints for making specific proteins, which in turn carry out all the functions of cellular life. While all cells in an organism contain the same DNA sequences, different cells make radically different sets of proteins, reflecting both the cells' function in the organism and current conditions or needs. The first step in the process of making a protein from its gene blueprint is called transcription or gene expression.

Understanding how cells control the expression of various genes gives us clues into the nature of the relevant proteins, leading to a better understanding of biology, evolution, and cellular responses to stress or disease. Embedded in the non-coding DNA of an organism are many control sequences that influence when a gene is expressed and which potentially-coding portions of the gene (exons) are spliced together to form the gene product.

Commonly, the DNA region just upstream of a particular gene is thought to contain many short substrings of DNA that do not actually encode any part of a protein, but that may be used to control the associated gene's expression. For example, molecules called transcription factors may bind to these regions of the DNA, either inhibiting or facilitating the expression of the gene. These DNA sequences are known as transcription factor binding sites (TFBSs). These and other sequences that may be used to regulate gene expression are collectively known as regulatory elements or *regulatory motifs*; in this manuscript we will just refer to them as *motifs*.

Though a given transcription factor may have a preferred TFBS sequence that it binds to, identification of such motifs is non-trivial, in part because the motifs are short (lengths of 6-8 base pairs are common), their locations may be anywhere within 1-2 kilobases of the gene, and any particular binding site may be only an inexact match to the ideal sequence. Thus, distinguishing biologically functional motifs from those appearing just by chance is difficult.

There are many different methods designed to identify motifs from sets of sequences thought to share common or homologous regulatory proteins (see reviews [11], [16]). Popular approaches include searching clusters of co-regulated genes (often those whose expression patterns are similar across many conditions) for statistically over-represented motifs, using techniques such as expectation maximization[1] or Gibbs sampling[19]. Alternatively, one can examine multiple orthologous sequences for motifs that are preferentially conserved throughout evolution [4], [5]. Interspecies comparison is a powerful tool to distinguish actively conserved sequences (an indicator of functionality) from sequences conserved due to shared ancestry [7], [17], [18]. The best such methods take advantage of phylogentic information to model the evolution of a putative motif.

In this manuscript, we consider both types of sequence sets - those from co-expressed genes and those from orthologous sequences corresponding to the promoter regions of the same gene in multiple species. For this study, we use MEME to find the motifs[1]. MEME uses an EM algorithm to identify motifs of maximum likelihood with respect to a probabilistic model of the sequence. While MEME does not make use of any phylogenetic information, it and other similar methods have successfully been used to identify conserved motifs from orthologous sequences[4].

III. Approach

Suppose that an arbitrary motif-finding program $\mathcal{P}$ is used to find the best motifs, according to some scoring system, common to a set of n sequences. While the motifs identified this way are the best according to the given scoring system, it is not clear that they are always the most biologically meaningful. Furthermore, when the sequences are derived from different species or from potentially-noisy clusters of co-expressed genes, it is possible that some functional regulatory motifs are not actually well represented in all sequences. Such motifs might not score very well on the full set of n sequences, or might not even be found at all.

Suppose that instead, we run the same program $\mathcal{P}$ on $n-1$ of the sequences, to identify a new set of motifs. We can do this n times, leaving out each sequence in turn, just as in leave-one-out cross-validation[13]. If we find a motif that scores well in the $n-1$ sequences, especially if it does so multiple times, it may be interesting and functional even if it does not score well enough in all n sequences to be detected by $\mathcal{P}$. This is the intuition behind the approach we investigate here.

First, let us introduce some terminology. We use the term *AllMax* to denote the motif set found in the n-species run, because it represents motifs found by *all* the sequences, using the *maximal* set of sequences available (rather than a subset of it). We also consider subsets of sequences, and run the same leave-one-out approach (called *LOO* hereafter in the text) on those subsets. We therefore call a run of $\mathcal{P}$ (in this case, MEME) on a subset of k sequences an *AllSubset* run, and we can then talk about LOO runs on the subset of size k as well.

To assess the quality of the motifs we discover using the LOO method is challenging without biological validation of the motifs. However, we address this question in part by evaluating whether LOO can be used on small subsets of sequences (of size $k < n$) to help approximate the results of an AllMax run. The idea here is that better motifs are found as the data set grows in size, and that if LOO can help identify these motifs with a small subset of the data, it is improving the quality of the motif-discovery process throughout.

	Cluster 60				Cluster 177			
Gene	All (17)	3	4	5	All (8)	3	4	5
alh-9	X							
aqp-2	X	X	X	X				
C09D4.2	X	X	X	X				
C46H11.2	X							
cwn-1	X							
F35D2.3	X							
hnd-1	X			X				
lin-17	X							
lin-18	X	X	X	X				
R02D3.1	X							
spp-10	X		X	X				
T22B7.3	X							
T27D12.1	X							
vab-8	X							
ZK1307.1	X							
ZK593.1	X							
ZK622.3a	X							
C34B2.7					X			X
cyn-7					X			
F25H2.5					X			
gly-3					X	X	X	X
R06C7.4					X		X	X
rskn-1					X	X	X	X
T05H10.7					X			
T12D8.8					X	X	X	X

TABLE I

Genes used for the different runs of Leave-one-out Scoring

IV. Methodology and Data

A. Data Sets

The Leave-one-out motif-finding procedure (LOO) was tested on two types of datasets: orthologous regulatory sequences from a range of species, and regulatory sequences from clusters of co-expressed genes.

Two sets of co-expressed genes were taken from [2], which studied gene expression during embryonic development in the worm *C. elegans*. To be in the same cluster, genes displayed similar expression patterns in 10 timepoints spanning early embryonic development in each of the wild-type and two RNAi knockout models. Clustering was done as in [3]. Briefly, genes were grouped using a quality-based clustering method [9], but only if the co-expressed genes were consistently co-clustered under a noise model fit to the expression data. Such an approach ensures that clusters are particularly tight and robust. Since the knockout experiment was designed to highlight genes likely to be directly or indirectly regulated by the transcription factor *pal-1*, we focused here on clusters likely to contain *pal-1* targets. Clusters 60 and 177 were chosen from those that showed a clean *pal-1*-target expression pattern because of their sizes, respectively 17 and 8 genes. The 1000bp region upstream of the transcription start site for each gene was downloaded from WormBase using WormMart (www.wormbase.org/biomart/martview). Table I displays the genes found in these clusters.

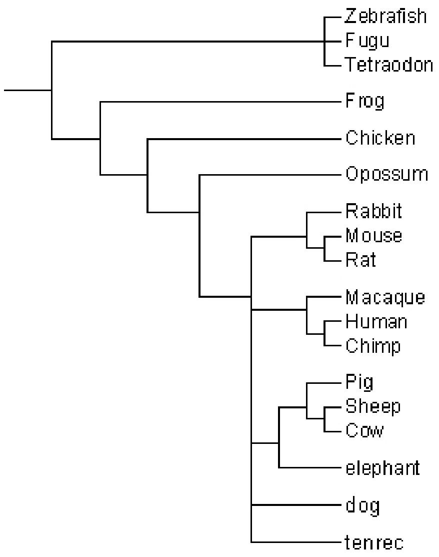

Fig. 1. Phylogeny for the selected vertebrate species. *This evolutionary tree describes relationships among the set of vertebrates selected for Leave-one-out Scoring. The phylogeny was constructed using data from [12], [14], [17].*

	SDF4				CCNL2			
Species	All (15)	3	4	5	All (11)	3	4	5
Human	X	X	X	X	X	X	X	X
Chimp	X			X	X			X
Macaque	X				X			
Cow	X				X			
Elephant	X				X			
Dog	X		X	X	X		X	X
Tenrec	X				X			
Rabbit	X							
Mouse	X	X	X	X	X	X	X	X
Rat	X	X	X	X	X	X	X	X
Opossum	X				X			
Chicken	X				X			
Frog	X							
Zebrafish	X							
Tetraodon	X							

TABLE II

SPECIES USED FOR THE DIFFERENT RUNS OF LEAVE-ONE-OUT SCORING

Two other sets of sequences were chosen from aligned upstream regions of orthologous genes (SDF4 and CCNL2) in multiple species. The data came from the multiple alignment of 15 vertebrate genomes to the human genome available at the UCSC Genome web site, release 18 (March, 2006) [10]. These particular genes were chosen because they had at least 1000bp in all available species that aligned well with the 2000bp upstream of the transcription start site in human. Not all species had unique orthologs for each gene, so there are 15 sequences in the SDF4 but only 11 for CCNL2. Table II shows which species were used for these genes; Figure 1 shows the putative evolutionary relationships between these species.

B. Method

For the purposes of this study, regulatory motifs were identified using MEME[1], version 3.5.3. For all experiments with MEME, we searched for a maximum of 100 motifs, exactly 6 bp in length, under the assumption that a motif appeared at most once in each sequence (i.e., using the zero-or-one option). The E-value cutoff was 1e-100, and motifs were allowed to occur on either the positive or negative strand.

Note that any other motif identification scheme with a system for scoring a given motif on a given sequence could be used. We chose to use MEME's reported information content as a way of scoring the motif on the input sequences. Given a position weight matrix W for a motif, let $W_{i,a}$ be the frequency of base a at position i, and let b_a be the background frequency of base a in the input sequence. Then the information content of that motif [16] is

$$\sum_{position\ i} \left(\sum_{letter\ a} W_{i,a} log \frac{W_{i,a}}{b_a} \right) \tag{1}$$

For each LOO run, a subset of the sequences was chosen. Within that subset, MEME was run with all sequences but one, returning the best motifs and their information content. We then scored those motifs in the left-out species as well, by computing their information content for the best possible motif match in the remaining sequence. The left-out information content was calculated using the background letter frequencies and the position weight matrix from the corresponding $n-1$ sequence run. This was repeated for all possible combinations (to leave out every sequence once), for each set of sequences.

As a baseline experiment, MEME was also run with all the sequences for comparison purposes, to obtain the AllMax set (or AllSubset in the case of a subset).

The results of all the runs were then compiled to produce a *LOO score* for each motif (equation 2).

$$\frac{s_{n-1} * (n-1) + s_1}{n} \tag{2}$$

The LOO score is based on the information content (score) the motif received in each of the runs. In equation 2, s_{n-1} refers to the score obtained in the $n-1$ sequence run, while s_1 refers to the left out sequence score, and n is the total number of sequences tested. If a motif was found in more than one run, the scores s_{n-1} and s_1 were obtained by taking the average score of each run. Taking the minimal or maximal value has the potential for skewing the score too much, while taking the average smooths out these irregularities. We then ranked the motifs according to their LOO score. A *frequency* index for each motif was also calculated: this figure indicates the percentage of the n LOO runs where the motif was found.

For each dataset, four different LOO tests were conducted. First, we ran LOO with the maximal number of sequences available (between 8 and 17). Then, subsets of 3, of 4 and of 5 sequences were selected from each of the datasets and tested with the LOO method. Tables I and II indicate which sequences were selected for each test, with each dataset.

V. Results and Discussion

In this section, we will address issues of motif quality using LOO and the stability of the LOO approach as n increases.

Overall, the number of distinct motifs discovered by the LOO method does not increase linearly with the number of sequences in the subset (Figure 2). Rather, in most cases, we observe an eventual decrease when LOO is executed with all the sequences. The number of motifs found by LOO is considerably less than k times the number of motifs in the single run, proving there are strong similarities among the motifs found by each of the runs.

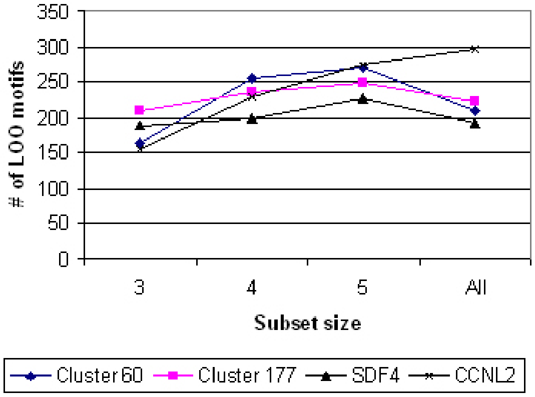

Fig. 2. Total number of LOO motifs found for 3, 4, and 5 species subsets.

A. LOO increases power of small data sets

Hoping to estimate motif quality without actually knowing which motifs are biologically meaningful, we investigated whether the use of LOO with only a few sequences would allow us to identify more of the motifs that could be identified with larger amounts of data. If so, LOO could be particularly useful both for finding promoter signals controlling only a small number of target genes, and for cross-species motif finding when the relevant sequences are only available in a few species.

The idea that this test reflects motif quality relies on the assumption that motifs found in the full AllMax run are more likely to be biologically relevant than those identified from only a small subset of the data. This assumption might not always hold true, especially when the subset runs contain only sequences from a few closely related mammals, while the AllMax data set includes fish and birds. To address this possibility, while we chose only mammals for the subset runs, we attempted to include some diversity within the mammalian population rather than picking the closest trio of mammals available (see Table II for details). For the co-expressed gene clusters, where subsets were chosen arbitrarily and the clusters are designed to exclude spuriously correlated genes, the argument seems more compelling. However, as we found consistent results across all the sequence sets, we suspect that the assumption does hold for all of them.

We found that the AllSubset runs, for subsets of size 3 to 5, typically detected only 20 to 30 percent of the motifs found by the AllMax run (Figure 3a). By considering both AllSubset and the additional motifs found only in one or more LOO runs, we were able to approximately double the number of the AllMax motifs identified, using just 3, 4, or 5 sequences instead of n. Thus, it appears that this approach allows us to extract more information from smaller datasets. Figure 3a shows the average results over all four data sets. Figures 3b, c, and d show the details for each data set using just LOO, just AllSubset, and the two combined, respectively.

These results show that LOO motif-finding using small amounts of data can help approximate the results of having a much larger set of related sequences. This suggests that the LOO method can help identify more meaningful motifs in a variety of contexts.

B. Gain from LOO changes as n increases

If LOO were run on a larger set of sequences than our trial data sets (whose sizes range from 8 to 17 sequences), the doubling of the percentage of AllMax motifs that we obtain using LOO (as shown in Figure 3) might no longer be seen. For example, if we limited our attention to motifs detected with reasonable frequency in the LOO runs, (that is, not just in one of n runs), we might expect that the results of LOO for a sufficiently large subset or cluster would not be very different from the AllMax results themselves. We decided to investigate whether the sequence sets we chose were large enough that this was the case.

To do so, we compared the number of distinct motifs we found from the combined LOO runs (with frequency of at least $1/3$) to the number of motifs found in the full AllSubset run, when the subset size was 3, 4, 5, or n. (That is, in the last case, we ran LOO on each full sequence cluster). We call motifs found in at least $1/3$ of the LOO runs *frequent* motifs. The results are reported in Figure 4 as the percentage of new motifs found with LOO compared to the corresponding AllSubset run. We call this value the *gain* of the LOO run. We expected to see the gain converging from a large percentage (over 100) for a 3-sequence subset to a fairly low percentage when run on the full n sequences.

While the trend in the Figure 4 generally agrees with the expected pattern, several interesting points arise. The first is that this “convergence” is not consistent across all sequence sets. Though there is a downward trend across the co-expressed cluster data as the number of sequences used increases, the data from the multiple species alignment actually shows a higher gain for n species than for 4 or 5 species, particularly for the gene SDF4. Partially this is due to the fact that the number of motifs found in the AllMax run for SDF4 drops significantly, but there may be other factors at work as well.

One possible contributing factor is that the subset runs were selected to contain only mammalian species. While convergence may be expected when more and more closely

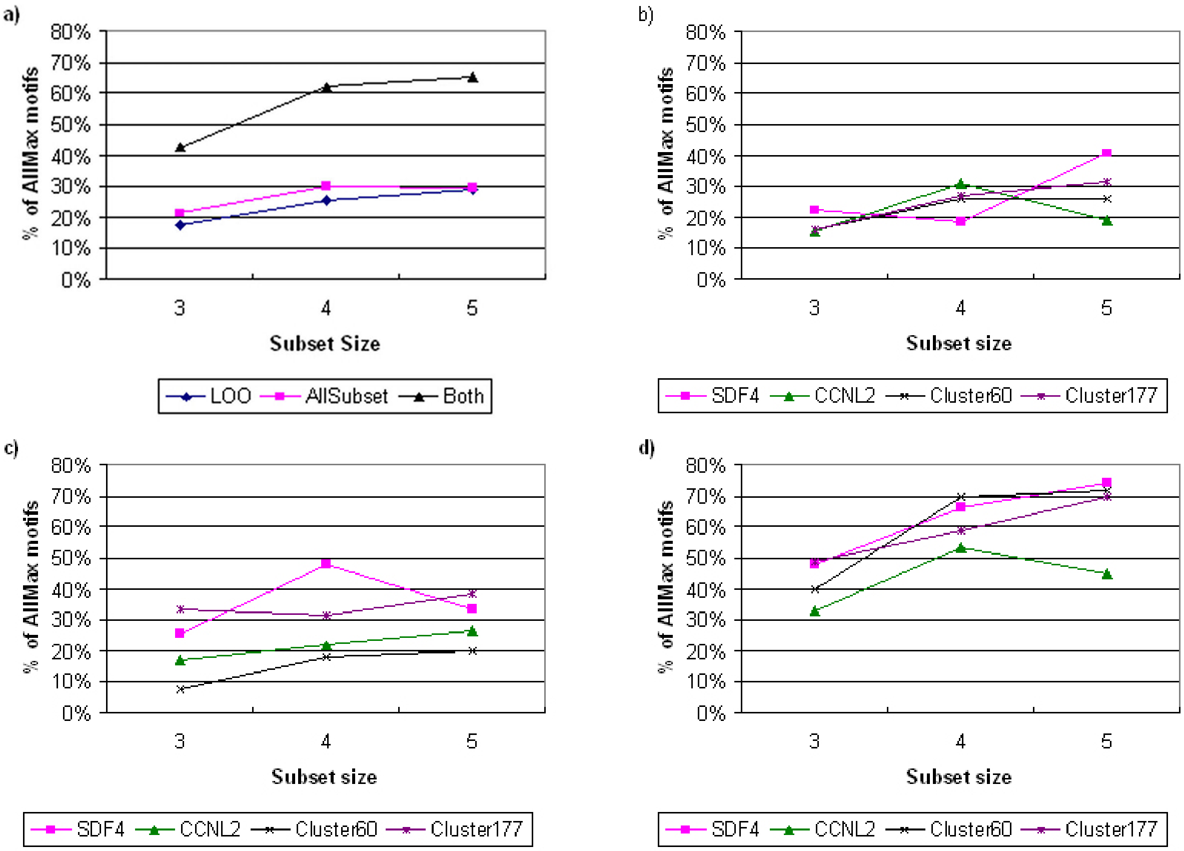

Fig. 3. Percent of AllMax motifs found for 3, 4, and 5 species subsets. a) Average (over all 4 data sets) percent of AllMax motifs found using just AllSubset, just LOO, and both methods combined. b) LOO details for all 4 data sets. c) AllSubset details. d) Combined LOO and AllSubset details.

related sequences are added to a data set, it is not expected when the sequences being added are noisy, or more distantly related. Thus, we see a trend reasonably consistent with the convergence theory for CCNL2, whose eleven sequences include ten mammals and a bird. However, for SDF4, adding in the full set of sequences means adding two fish and an amphibian sequence to the bird and eleven mammalian sequences. Thus, fewer motifs meet the significance criteria for the the AllMax run, and the higher gain for the n species run is perhaps to be expected. Also consistent with this theory is the fact that the co-expressed clusters of worm genes better fit the expected convergence pattern, and that the pattern is strongest for Cluster 60, the largest cluster.

Another interesting point arising from this experiment is that even for Cluster 60, with 17 sequences, using LOO on all 17 sequences yields a 44% increase in the set of frequent motifs (those found in at least $1/3$ of the LOO runs) over the sequences found by using AllMax alone. This suggests that even a cluster of 10-20 sequences is not sufficiently large for this process to have converged. While one possible cause is that there is an increasing amount of noisy data in the full cluster, it seems equally plausible that clusters need to be considerably larger before this process converges.

For example, Table III shows some sample motifs found by the LOO approach on all 11 CCNL2 sequences. In the table,

TABLE III

20 BEST MOTIFS FOUND IN CCNL2. *Score refers to the LOO score obtained from equation 2; Rank to the ranking according to Score. AllScore is the score obtained with the All-sequence run (no LOO), AllRank to the ranking according to AllScore. Blank scores and ranks in AllScore and AllRank indicate the motif was not ranked as one of the 100 best motifs found in the All-sequence run.*

Motif	Score	Rank	AllScore	AllRank	Freq
TTAAAA	14.57	1	14.8	1	0.91
AAAATA	14.54	2	14.8	1	0.82
TTATTT	14.49	3			0.64
TTAATA	14.49	4	14.8	1	0.82
TTTATT	14.11	5			0.09
TTTTAT	13.93	6			0.55
TGAAAA	13.8	7			1
GTTTTT	13.71	8			0.36
TTTCTT	13.64	9	12.6	30	0.82
TTCATT	13.55	10			0.27
AAACAT	13.54	11			0.45
TACTTA	13.53	12	14	4	1
TTCTTT	13.53	13	13.5	11	0.73
TTTTTC	13.52	14	14	4	0.55
GTAAAA	13.52	14			0.55
AAAAGA	13.52	16	12.6	30	0.45
TGAATT	13.51	17	14	4	0.45
TTTTCT	13.48	18	13	27	0.27
TGTTTT	13.47	19	14	4	0.45
TGAATA	13.46	20			0.09

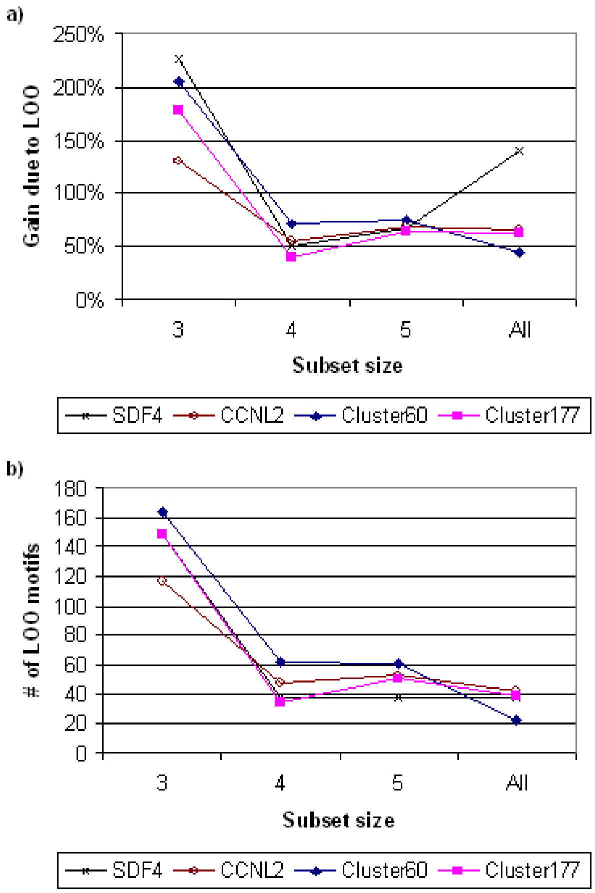

Fig. 4. a) Frequent new motifs found by the combined LOO runs, as a percentage of those found by the corresponding AllSubset run b) Raw numbers of frequent motifs found by each AllSubset run (the numerator in part a).

“Rank” refers to the sequence rank by information content (score) in the original 11-species AllMax run. AllMax motifs with the same information content are listed as having the same Rank. The table shows examples of many motifs with high information content that are found with high frequency in the LOO runs, but that would not have been found at all using a single MEME run on all the sequences. Thus, it seems worth considering the LOO approach even for sequence clusters of reasonable size.

VI. Conclusions and Future Directions

A new method of motif analysis, Leave-one-out Scoring, was developed and tested. Using this strategy, it may be possible to achieve “better” rankings of motifs than those obtained using a single run of a given motif finding method. We believe that the additional information derived from the approach will often justify the added computational expense.

The behavior of the LOO method as n increases suggests that typical clusters containing tens of co-expressed genes may not yet have “converged” to the point where LOO provides only redundant information. As motif finding is often done in much less conservatively-formed clusters than those described here (generally a full partitioning or hierarchical clustering method is used, and no noise-tolerance filter is applied), typical clusters of co-expressed genes are likely to be much noisier than those we tested. Thus, even larger cluster sizes are likely to benefit from the Leave-one-out approach.

Furthermore, the LOO approach may allow the extraction of better motifs using a single motif-finding method, without relying on the consensus of a number of different methods. Thus, radically novel motif-detection programs that identify regulatory elements rarely found by other methods might particularly benefit from the use of Leave-one-out Scoring.

Future work should extend the preliminary findings reported here to larger data sets and additional methods. At a minimum, we would like to investigate the effects of the LOO method using several common motif-recognition algorithms. We would also like to further investigate convergence behavior by examining a much greater range of cluster sizes, and to compare the behavior of these methods on the promoter regions of co-expressed genes to those on aligned sequences from multiple species. Another interesting possibility would be to extend Leave-one-out Scoring to Leave-k-out Scoring. This new method would examine motifs in $n - k$ ($k > 1$) sequences. Finally, to combat the rising computational costs as n and k grow, one could perform only random subsets of the desired LOO runs (essentially a constrained form of Bootstrapping).

The ideal test of this method would be a labeled data set with “correct” motifs. While this is not available as such, one possible strategy would be to start with known, experimentally-validated transcription factor binding sites, and to work backwards to show that this approach validates a higher percentage of those than the straightforward application of any of a variety of motif detection techniques. The genes with the greatest number of long orthologous promoter regions in the most species, those we selected for this study, did not happen to contain clusters of experimentally validated binding sites. However, as sequence availability increases, it should be possible to find some good candidate sequence sets for this experiment.

With datasets from orthologous sequences, adding a parameter concerning the evolutionary distance of each species would allow us to make decisions about the validity of a motif based on the distance to the evolutionary norm of the held-out species. To enforce this rule, it would perhaps be possible to eliminate motifs that are outside of a certain range, defined by a multiplier of the standard deviation. Certainly, a number of motif detection programs make use of evolutionary data [4], [8]; it would be interesting to explore the effects of integrating the LOO method with such algorithms in a way that takes full advantage of the phylogenetic data.

References

[1] T L Bailey and C Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Second International Conference on Intelligent Systems for Molecular Biology*, pages 28–36. AAAI Press, Menlo Park, CA, 1994.

[2] LR Baugh, AA Hill, JM Claggett, K Hill-Harfe, JC Wen, DK Slonim, EL Brown, and CP Hunter. The homeodomain protein PAL-1 specifies a lineage-specific regulatory network in the *C. elegans* embryo. *Development*, 132(8):1843–54, March 2005.

[3] LR Baugh, AA Hill, DK Slonim, EL Brown, and CP Hunter. Composition and dynamics of the *Caenorhabditis elegans* early embryonic transcriptome. *Development*, 130(5):889–900, March 2003.

[4] M. Blanchette and M. Tompa. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Research*, 12(5):739–748, 2002.

[5] CB Congdon, CW Fizer, NW Smith, HR Gaskins, J Aman, G Nava, and C Mattingly. Preliminary results for GAMI: A genetic algorithms approach to motif inference. In *2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB-2005)*, November 2005.

[6] Bradley Efron and Gail Gong. A leisurely look at the bootstrap, the jackknife, and cross-validation. *JSTOR*, 37(1):36–48, 1983.

[7] KA Frazer, JB Sheehan, RP Stokowski, X Chen, R Hosseini, JF Cheng, SP Fodor, DR Cox, and N. Patil. Evolutionarily conserved sequences on human chromosome 21. *Genome Research*, 11:1651–1659, 2001.

[8] R Van Hellemont, P Monsieurs, G Thijs, B de Moor, Y Van de Peer, and K Marchal. A novel approach to identifying regulatory motifs in distantly related genomes. *Genome Biology*, 6(13):R113, 2005.

[9] LJ Heyer, S Kruglyak, and S Yooseph. Exploring expression data: identification and analysis of coexpressed genes. *Genome Research*, 9(11):1106–1115, 1999.

[10] W J Kent, C W Sugnet, T S Furey, K M Roskin, K M Pringle, A M Zahler, and D Haussler. The human genome browser at UCSC. *Genome Research*, 12(6):996–1006, 2002.

[11] M A Lones and A M Tyrrell. The evolutionary computation approach to motif discovery in biological sequences. In *Genetic and Evolutionary Computation Conference (GECCO-2005); Workshop on Biological Applications of Genetic and Evolutionary Computation*. ACM SIGEVO, 2005.

[12] D. R. Maddison and K.-S. Schulz (ed.). The tree of life web project, 2004.

[13] Tom Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[14] W. J. Murphy, E. Eizirik, W. E. Johnson, Y. P. Zhang, O. A. Ryder, and S. J. O'Brien. Molecular phylogenetics and the origins of placental mammals. *Nature*, 409(6820):614–618, February 2001.

[15] Jun Shao and Dongsheng Tu. *The Jackknife and Bootstrap*. Springer-Verlag, 1995.

[16] G B Stormo. DNA binding sites: representation and discovery. *Bioinformatics*, 16(1):16–23, 2000.

[17] J W Thomas and J W Touchman. Vertebrate genome sequencing: building a backbone for comparative genomics. *Trends Genet*, 18:104–108, 2002.

[18] J W Thomas, J W Touchman, R W Blakesley, G G Bouffard, S M Beckstrom-Sternberg, and E H Margulies. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature*, 424:788–793, 2003.

[19] W Thompson, E C Rouchka, and C E Lawrence. Gibbs recursive sampler: finding transcription factor binding sites. *Nucleic Acids Research*, 31(13):3580–3585, 2003.