

Discovering Gene Regulatory Elements Using Coverage-Based Heuristics

Rami Al-Ouran, Robert Schmidt, Ashwini Naik, Jeffrey Jones, Frank Drews, David Juedes, Laura Elnitski, and Lonnie Welch

Abstract—Data mining algorithms and sequencing methods (such as RNA-seq and ChIP-seq) are being combined to discover genomic regulatory motifs that relate to a variety of phenotypes. However, motif discovery algorithms often produce very long lists of putative transcription factor binding sites, hindering the discovery of phenotype-related regulatory elements by making it difficult to select a manageable set of candidate motifs for experimental validation. To address this issue, the authors introduce the motif selection problem and provide coverage-based search heuristics for its solution. Analysis of 203 ChIP-seq experiments from the ENCyclopedia of DNA Elements project shows that our algorithms produce motifs that have high sensitivity and specificity and reveals new insights about the regulatory code of the human genome. The greedy algorithm performs the best, selecting a median of two motifs per ChIP-seq transcription factor group while achieving a median sensitivity of 77 percent.

Index Terms—Motif discovery, ChIP-seq, RNA-seq, biology of disease, ENCODE

1 INTRODUCTION

HUMAN disease association studies are often gene-centric and focus on identifying variants in genes. However, numerous diseases are caused by alterations in the non-coding, regulatory regions of the genome. Thus, discovery of genomic regulatory elements is not only important for understanding the biology of genomes, it is also critical for understanding the biology of disease [1]. RNA-seq, microarray and ChIP-seq experiments are used to discover disease-associated changes in gene expression and in transcription factor binding. Such experiments identify genomic areas (e.g., gene promoters and transcription factor binding regions) wherein disease-associated regulatory elements may be found.

Motif discovery, the *de novo* computational method for finding putative regulatory element binding sites, has several shortcomings. The *specificity problem* occurs when motif discovery methods produce too many motifs, causing a high false positive rate. The *coverage problem* occurs when motif discovery methods fail to find a single motif (or a small set of motifs) that covers all of the genomic sequences of interest (e.g., the binding regions from a ChIP-seq experiment).

These issues can be addressed by solving the *motif selection problem*, i.e., picking a small set of significant motifs from a large collection of discovered motifs. Since one might view each subset of the discovered motifs as a hypothesis concerning transcription factor binding or gene co-expression, by the principle of Occam's Razor, the simplest such hypothesis is preferred. Hence, the output of our method is viewed as a likely genomic mechanism to explain the common regulatory (or binding) properties of a sequence set.

In the remainder of this manuscript, the authors formally define the motif selection problem, present novel methods for solving the problem, and demonstrate the effectiveness of the methods by analyzing ChIP-seq data from the ENCyclopedia of DNA Elements (ENCODE) project [2]. Section 2 provides the biological motivation for the motif selection problem and reviews related algorithmic methods. In Section 3, the motif selection problem is formally defined and algorithms for the motif selection problem are presented. The effectiveness of the algorithms is demonstrated in Section 4 by providing analysis results for the ENCODE data.

2 BACKGROUND AND SIGNIFICANCE

Transcription factor proteins (TFs) and their DNA binding sites (TFBSs) are involved in the regulation of gene transcription. Thus, identifying TFBS-TF interactions assists in deciphering gene regulatory networks. Discovering TFBSs is considered a challenging problem in the fields of computer science and molecular biology, because the TFBSs are degenerate, TFBSs vary in length, and TFs work in a combinatorial manner [3]. Motif discovery is one of several methods used to help discover TFBSs.

Motif discovery is the process of finding short DNA patterns that are overrepresented in a set of DNA sequences

- R. Al-Ouran, R. Schmidt, F. Drews, D. Juedes, and L. Welch are with the Department of Electrical Engineering and Computer Science, Ohio University, Athens, OH 45701. E-mail: {ra102506, rs405111, drews, juedes, welch}@ohio.edu.
- A. Naik is with the Research Institute at Nationwide Children's Hospital, Columbus, OH 43110. E-mail: an911111@ohio.edu.
- J. Jones is with The Ohio State University, Columbus, OH 43210. E-mail: jones.5374@osu.edu.
- L. Elnitski is with the National Human Genome Research Institute, Bethesda, MD 20892. E-mail: elnitski@mail.nih.gov.

Manuscript received 16 May 2015; revised 6 Oct. 2015; accepted 13 Oct. 2015. Date of publication 30 Oct. 2015; date of current version 6 Aug. 2018. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TCBB.2015.2496261

which share a biological function (e.g., promoters of co-expressed genes) [4] or are bound by the same TF (as determined by ChIP-seq experiments). Motif discovery methods are divided into two classes. *Generative methods* discover motifs by contrasting their enrichments in a set of sequences with generated statistical background models (for example, a Markov background model generated from the input sequences [5]). Some well-known generative discovery methods include MEME [6] and Weeder [3].

Discriminative motif discovery methods compare two sets of sequences to find motifs that are overrepresented in the genomic regions of interest (the positive set) and are underrepresented in a different set of sequences (the negative set) which is provided by the user. The discriminative motif discovery problem was first introduced by Sinha [7], where a motif was considered a feature of the input sequences, and features that best discriminate the two sets were identified by applying classification techniques. Features can be ranked based on their power to discriminate between members of the two classes. Discriminative motif discovery methods include xxMotif [5], DECOD [8], DEME [9], and DME [10].

Tompa et al. [11] found that the sensitivity of motif discovery methods is very low. To address this problem, Tompa et al. suggested the use of an ensemble of multiple motif discovery tools. Ensembles implicitly solve the motif selection problem by choosing a set of motifs produced by a collection of motif discovery methods. Thus, the remainder of this section provides a review of motif discovery ensembles, including a description of the motif selection approach used by each ensemble.

In [12], five motif discovery tools were combined (AlignACE [13], MDScan [14], MEME [6], Trawler [15], and Weeder [3]). The motifs are ranked using an enrichment score that is computed by dividing the number of discovered motif instances by the number of shuffled control motif instances across the genomic regions studied.

W-ChIPMotifs [16], [17] uses three motif discovery tools (MEME [6], MaMF [18], and Weeder [3]) to discover motifs in ChIP-seq data. The candidate motifs are filtered for significance using a bootstrap resampling method and using p-values.

In CompleteMOTIFs [19], three motif discovery methods (MEME [6], Weeder [3], and ChIPMunk [20]) are used. The top 10 motifs are selected from each tool based on its scoring method. The candidate motifs (from all the tools) are scanned across a background data set generated by shuffling the original input data. CompleteMOTIFs reports the 10 motifs with the strongest q-values.

GimmeMotifs [21] incorporates nine motif discovery tools to discover motifs across ChIP-seq data. The tools are: BioProspector [22], GADEM [23], Improbizer [24], MDmodule [25], MEME [6], MoAn [26], MotifSampler [27], Trawler [15] and Weeder [3]. The input data are divided into a prediction set (20 percent of the original input set selected randomly) and a validation set. Additionally, two background data sets are generated to calculate motif statistical significance. One background data set is randomly generated from the input data while maintaining the dinucleotide frequency, and the second background data set is selected from the genome studied. The statistical significance of the non-redundant motifs is

then calculated using the validation set and the background data sets. The following statistical scores are calculated: absolute enrichment, hypergeometric p-value, ROC-AUC graph, and the Mean Normalized Conditional Probability. Finally, all the candidate motifs are clustered using a Weighted Information Content similarity score and the non-redundant motifs are reported.

The SCOPE ensemble [28], [29] uses three motif discovery algorithms (BEAM, PRISM, and SPACER), and a scoring metric, *Sig*, is used to rank the motifs. *Sig* is a statistical significance score that is based on three objective functions (motif overrepresentation, motif coverage, and motif positional bias). The motif coverage score is used to determine the statistical significance of only a single motif to assist in ranking the predicted motifs. This score compares the number of regions in the input set that contain the motif to the total number of regions in the entire genome that contain the motif.

Another motif ensemble method is MotifLab [30], wherein motif discovery is performed by popular tools chosen by the user and a p-value for over-representation is calculated for the top motifs produced by each tool.

In Ensemble Motif Discovery (EMD) [31], five motif discovery tools are used (AlignACE [13], Bio-Prospector [22], MDScan [14], MEME [6], and Motif-Sampler [27]). The ensemble approach for selecting motifs consists of five steps: collecting, grouping, voting, smoothing, and extracting. EMD runs the tools multiple times, where some tools are run with different parameters each time to produce different sets of motifs per run. The motifs are collected and grouped based on their scores, the groups are mapped onto the input sequences, and votes for each position across the sequences are counted. The final sites reported are the ones with the largest numbers of votes.

In MotifVoter [32], 10 motif discovery methods are used: MITRA [33], Weeder [3], SPACE [34], AlignACE [13], ANN-Spec [35], BioProspector [22], Improbizer [24], MDScan [14], MEME [6] and MotifSampler [27]. MotifVoter includes two stages: motif filtering and sites extraction. In the motif filtering step, similar motifs are clustered. In the site extraction step the goal is to identify the binding sites with the highest confidence based on how many motif methods report the site, where a binding site should be shared by at least two motif discovery methods. The final high confidence selected binding sites are aligned using MUSCLE [36] and a PWM is generated.

While motif discovery ensembles have been developed to address the problems identified by Tompa et al. [11], they tend to exacerbate the specificity problem and they do not consider the coverage problem. These challenges are addressed by explicitly defining and solving the motif selection problem, which incorporates both objectives.

3 METHODS

Whether using a single motif discovery method or an ensemble of motif discovery methods, one faces the challenge of selecting a biologically important subset of motifs from a large set of candidate motifs. This section provides a formal description of the motif selection problem and presents algorithms that solve the problem. The source code of the algorithms is available at <https://github.com/RamiOran/SeqCov.git>.

3.1 Formal Problem Definition

Given a set of motifs $M = \{m_1, m_2, \dots, m_k\}$ and a set of sequences $S = \{S_1, S_2, \dots, S_n\}$, the motif selection problem can be defined as follows:

$$\text{Minimize } \sum_{j=1}^k x_j. \quad (1)$$

Subject to:

$$\sum_{j=1}^k a_{ij}x_j \geq 1, \quad i = 1, \dots, n, \quad (2)$$

where x_i is defined as:

$$x_i = \begin{cases} 1 & \text{if } m_i \text{ is part of the solution} \\ 0 & \text{otherwise,} \end{cases} \quad (3)$$

where A is an $n \times k$ matrix representing the coverage of sequence set S by motif set M :

$$a_{ij} = \begin{cases} 1 & \text{if } S_i \in S_{m_j} \\ 0 & \text{otherwise,} \end{cases} \quad (4)$$

where $S_{m_i} \subseteq S$ is the set of sequences covered by motif m_i .

Note that equation (2) guarantees that each sequence in S is covered by at least one motif m_i . Note also that the motif selection problem can be modeled by the Set Covering Problem (SCP), and that the SCP decision problem is NP-complete and the SCP optimization problem is NP-hard [37].

A general solution procedure for the motif selection problem finds F^{\min} , a minimally sized set of motifs that covers all sequences in S . A feature set $F \subseteq M$ is generated by incrementally adding features (motifs) to the set, based on the heuristic rule of the algorithm used such that if a new motif m_i is added to F , then the corresponding set of sequences covered by m_i are added to S_F (the set of sequences covered by feature set F). This implies that it is never beneficial to add any motif that does not increase the size of S_F . The procedure terminates when $S_F = S$.

3.2 Relaxed Integer Linear Programming (RILP) Approximation Algorithm

The Set Cover Problem as well as the motif selection problem can be cast as a 0-1 integer linear program, wherein the goal is to find a 0-1 vector \vec{x} of length m satisfying the constraints $A\vec{x} \geq \vec{b}$ such that $Z = \vec{c} \cdot \vec{x}$ is minimized, where

- 1) \vec{c} is a vector of all 1's of length m
- 2) \vec{b} is a vector of all 1's of length n .

It is relatively straightforward to prove that the minimum set cover has K sets if and only if $Z = K$. To see this, notice that $\vec{x}[i] = 1$ corresponds to the motif m_i being part of the set cover, and $\vec{x}[i] = 0$ corresponds to the set m_i being left out of the set cover. Hence, $A\vec{x} \geq \vec{b}$ if and only if, for each $S_i \in S$, at least one of the sets m_i , where $\vec{x}[i] = 1$ contains S_i . Hence, if \vec{x} satisfies the constraint that $A\vec{x} \geq \vec{b}$, then \vec{x} corresponds to a valid set cover. The additional constraint that $\vec{c} \cdot \vec{x}$ is minimized means that the solution to the integer linear program

provides the optimal set cover. Hence, solvers for integer linear programs, such as those found in the GNU Linear Programming Kit (GLPK) [38], can be used to find the optimal solution to the set cover problem. However, these solvers may take a long time to find the optimal solution.

Now, while 0-1 integer linear programming is also NP-complete [37], it is possible to relax the constraint that $x_i \in \{0, 1\}$ and allow $x_i \in [0, 1]$. This relaxation converts the integer linear program into a linear program. Since linear programming can be solved in polynomial-time in the worst-case (e.g., the Ellipsoid Method), GLPK can be used [38] to solve the relaxed version. Furthermore, the relaxed version can be used to provide an approximate solution via randomized rounding [39]. The standard randomized rounding approach proceeds as follows: (i) construct the optimal solution \vec{x} to the relaxed version of the ILP problem, (ii) select set S_i be part of the cover C with probability $x[i]$, (iii) repeat step (ii) until C is a set cover. As shown in [39], this algorithm produces, with high probability, a set cover that is within $O(\log n)$ times the size of optimal solution. This is the approach used here to build good set covers.

3.3 Bounded Exact Search Algorithm

The GLPK toolkit [38] also provides a branch-and-cut algorithm which attempts to find exact solutions to certain stated integer linear programming problems by utilizing accepted trial solution methods (such as the simplex or primal-dual interior-point methods) and then successively computing cutting planes to reduce the size of the search space. This technique is applicable to our problem of interest since the set coverage problem can be described using only linear constraints, and the search space is convex. Branch and cut behaves heuristically in terms of the search space reduction, but provides exact answers to the linear programming problem. The principle drawback to branch and cut is that it demonstrates exponential runtime in the worst case, and therefore may not return any answer to specific problem instances within a reasonable time frame. The ILP characterization is provided in the previous section.

3.4 Greedy Algorithm

Greedy algorithms try to generate good solutions by employing simple rules. The strategy chosen here is to employ a "maximum uncovered-first" rule. According to this rule, a feature set F is constructed by incrementally adding motifs such that, at every iteration, the motif m_i that covers the largest number of uncovered sequences in S is added to F .

Two filtering steps are applied during the greedy motif selection process. The first filtering step is used to avoid selection of redundant features. During the iteration process, if feature m_i is similar to a previously selected feature m_k then feature m_i is discarded and the search continues for the next feature. The similarity of two motifs, m_i and m_k , is calculated using Tomtom [40], which assigns an E-value that characterizes the significance of the similarity. The significance of similarity between two motifs m_i and m_k is

defined as $\varepsilon(m_i, m_k)$. If $\varepsilon(m_i, m_k) < 0.05$, then the two motifs m_i and m_k are considered similar.

The second filtering step avoids selecting features which provide small incremental benefit, choosing features which add a minimum number of uncovered sequences. Let S_u be the set of uncovered sequences, let $|S_{m_i} \cap S_u|/|S|$ be the percentage of sequences covered by feature m_i , and let Δ be the minimum percentage of new sequences that must be added to the set cover. If $|S_{m_i} \cap S_u|/|S| < \Delta$, the greedy algorithm terminates (because no further improvement greater than Δ is possible by selecting any of the remaining motifs). This is beneficial since some features only add a small percentage of uncovered sequences. Although the filtering steps might result in partial coverage instead of full coverage, they produce a feature set which includes non-redundant features and avoids selection of features that add a small number of uncovered sequences. Algorithm 1 shows the pseudocode for the greedy algorithm.

Algorithm 1. Motif Selection Using the Greedy Algorithm.

```

1: procedure GREEDY ALGORITHM( $S, M, \Delta$ )
2:    $S_u = S$ 
3:    $F = \emptyset$ 
4:    $M_s = M$ 
5:    $j = 0$ 
6:   while  $S_u \neq \emptyset$  and  $j < |M|$  do
7:     Select an  $m_i \in M_s$  s.t.  $|S_{m_i} \cap S_u|$  is maximized
8:      $M_s = M_s - m_i$ 
9:      $j = j + 1$ 
10:    if there exists  $m_k \in F$  such that  $\varepsilon(m_k, m_i) < 0.05$  then
11:      continue
12:    end if
13:    if  $|S_{m_i} \cap S_u|/|S| < \Delta$  then
14:      break
15:    else
16:       $S_u = S_u - S_{m_i}$ 
17:       $F = F \cup m_i$ 
18:    end if
19:  end while
20:  Return  $F$ 
21: end procedure

```

4 RESULTS AND DISCUSSION

This section presents the results of applying our motif selection methods to the ENCODE ChIP-seq data [2]. Our results are compared to those described in [12].

In [12], the authors grouped 427 ChIP-seq experiments from the ENCODE project into 84 factor groups (based on homology and the presence of known motifs). For each ChIP-seq experiment in each factor group, the ChIP-seq peaks were divided into two parts, one for motif discovery by an ensemble of motif discovery tools and one for enrichment score calculation. The top 10 enriched motifs were selected for each factor group. Of the 84 factor groups, 56 groups have known TFBSs.

Our methods typically selected fewer motifs per factor group than reported in [12], and the TFBSs selected by our methods often cover higher percentages of the ENCODE ChIP-seq binding regions than did the motifs reported in [12]. Our methods are validated by their ability to

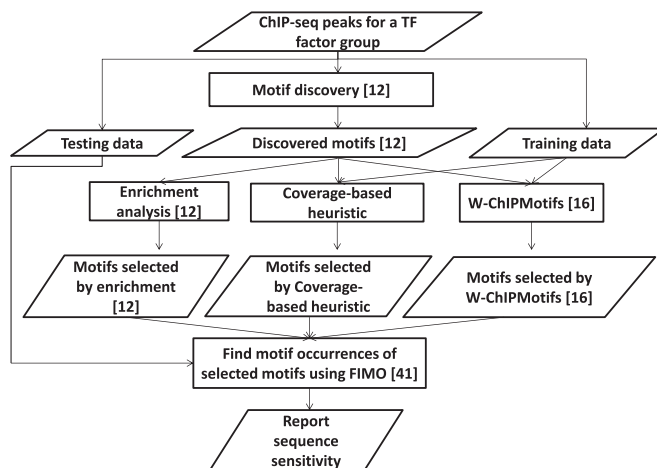


Fig. 1. Evaluation pipeline applied to the ENCODE ChIP-seq data. One thousand random peaks were selected per experiment per factor group for the training and testing data. Peaks selected for training data were not included in the testing data. The discovered motifs were reported in [12] using an ensemble of motif discovery tools. FIMO [41] was used for motif scanning.

rediscover known binding motifs for 38 factor groups. Interestingly, our methods rediscovered known motifs for one factor group (TCF12) for which the method in [12] failed to find known motifs.

The remainder of this section summarizes our key results. First, our evaluation pipeline is described and our evaluation metrics are defined. Focusing on the aforementioned 38 factor groups, the effectiveness of our motif selection methods is compared to the effectiveness of the method of Kheradpour and Kellis [12], as well as to the motif selection method used in a motif discovery ensemble (W-ChIP-Motifs [16]). Finally, we present new putative functional genomic elements discovered by our methods.

4.1 Evaluation Methodology

Fig. 1 shows the pipeline used for evaluating the motif selection methods. For each ChIP-seq experiment in each factor group, 1,000 randomly selected peaks were used as training data and 1,000 randomly selected peaks were used as testing data. The sets of all discovered motifs for each TF group were obtained from [12], and were provided as input to our motif selection algorithms and to the motif selection algorithm of W-ChIPMotifs. Filtering thresholds of $\Delta >= 5\%$ and $\varepsilon(m_i, m_k) < 0.05$ were used for the greedy algorithm.

The motif selection methods were evaluated in terms of the following metrics:

- 1) *Number of features selected (N)*: This measure indicates the number of motifs chosen by a motif selection method.
- 2) *Sequence sensitivity (sSn)*: This measure indicates the percentage of input sequences (ChIP-seq peaks) that were identified by the selected set of features reported by a method. Sequence sensitivity, **sSn**, is defined as $sSn = TP_s / (TP_s + FN_s)$, where the number of true positives, TP_s , is the number of sequences containing at least one selected motif (determined by using FIMO [41]) and the number of false negatives,

TABLE 1
Number of Features Used (Mean, Median, and SD)
across 38 TF Groups

Method	Mean	Median	SD
Greedy	2.5	2	0.86
Enrichment	4	3	2.2
W-ChIPMotifs	8.2	6	6.5
RILP	30.7	30	15.2
Bounded	30.7	30	15.2

FNs, is the number of sequences with no occurrence of any selected motif.

- 3) *Sensitivity in recovering known motifs (mSn)*: For each of the 56 TF groups with known motifs (see [12]), the known motifs were compared to the selected motifs. Motif sensitivity, **mSn**, is defined as $mSn = TPm / (TPm + FNm)$, where *TPm* is the number of TF groups with known motifs that are covered by the selected motifs, and *FNm* is the number of known motifs not matched by the predicted motifs (determined using Tomtom [40], with E-value threshold = 0.05).

4.2 Evaluation Results

Table 1 provides a summary comparison between the methods, in terms of the number of features selected. The greedy algorithm produces the smallest average number of features, followed by the enrichment method. The other three algorithms produce much larger feature sets. Fig. 2a provides a comparison between the greedy, enrichment, and W-ChIPMotifs methods, where it is clear the greedy has the lowest median number of features. The small number of features selected by the greedy algorithm and the concurrent high **sSn** indicate the strong specificity of the selected motifs.

The **sSn** measure is used to find the percentage of ChIP-seq peaks covered by the selected motifs. Table 3 shows a high-level comparison of all methods, and Fig. 2b shows a comparison between the greedy, enrichment, and W-ChIPMotifs methods. In terms of **sSn**, the RILP and bounded exact search algorithms performed the best. However, the higher sensitivity was obtained at the expense of selecting a very large number of motifs (see Table 1). The greedy algorithm and the enrichment method have much better motif-to-sensitivity ratios, and the greedy algorithm has the most favorable ratio overall. Table 2 shows the number of features and **sSn** values across the 38 TF groups.

The enrichment method reported known motifs for 37 TF groups (**mSn** = 66.1%). The RILP and bounded exact search algorithms reported known motifs for 38 TF groups (**mSn** = 67.9%). The W-ChIPMotifs method reported known motifs for 35 TF groups (**mSn** = 62.5%). The greedy algorithm achieved 64.3 percent **mSn** (reporting known motifs for 36 TF groups). All methods performed similarly with respect to the gold standard.

It is important to note that the greedy algorithm achieved this performance with fewer motifs, on average, than the other algorithms. In terms of running time, the greedy algorithm was the fastest with average run time

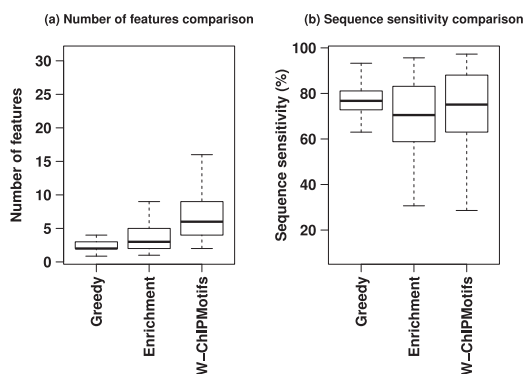


Fig. 2. Number of features and sequence sensitivity comparison.

of 103 seconds. The median was 57 seconds and the standard deviation was 123. The RILP and the bounded exact search algorithms had an average run time of 141 and 129 seconds, respectively. The median was 90 and 85 seconds and the standard deviation was 153 and 136 seconds, respectively.

4.3 Putative Functional Genomic Elements Discovered by Our Methods

The greedy method identified a number of putative regulatory elements. First, we present our findings for the TCF12 factor group, for which the previous study (see [12]) failed to rediscover a known motif. This is followed by a presentation of previously unreported motifs for the remaining 37 factor groups for which our methods were validated with respect to the gold standard.

The TCF12 TF (other names include HTF4 and HEB) is a member of the basic helix-loop-helix (bHLH) protein family and a member of the E-protein class which binds to the E-box sequence CANNTG [42], [43], [44].

The greedy algorithm reported three motifs for TCF12 group, with a **sSn** of 69.6 percent. The RILP, bounded exact search algorithms reported 29 motifs with a **sSn** of 96.2 percent. The W-ChIPMotifs method reported four motifs with a **sSn** of 55.3 percent. In [12], the enrichment method reported six motifs with a **sSn** of 72.3 percent.

Fig. 3 shows the matches found by comparing the predicted motifs by the greedy algorithm for TCF12 against the JASPAR database. In Fig. 3, the second selected motif matches other TFs from the JASPAR database which are likely co-factors of TCF12.

To study this factor group further, motif selection was performed on each of the TF ChIP-seq experiments individually, to identify cell line specific TFBSs. The TCF12 factor group consists of three experiments across three cell lines (HepG2, H1-hESC, and GM12878). Fig. 4 shows the motifs selected by the greedy algorithm for each experiment. The TCF12 binding regions of the two normal cell lines (GM12878 and H1-hESC) contain similar motifs, but the motifs selected for the binding regions of the hepatocellular carcinoma cell line (HepG2) are different. This suggests the presence of genomic regulatory elements that may be linked to hepatocellular carcinoma.

We examined the TCF12 for any overlapping Single Nucleotide Polymorphisms (SNPs), using the RegulomeDB

TABLE 2
Number of Features and sSn for Five Motif Selection Methods Across the 38 TF Groups

TF Group(P)(P(%))	Greedy		RILP and Bounded		W-ChIPMotifs		Enrichment	
	N	sSn(%)	N	sSn(%)	N	sSn(%)	N	sSn(%)
EGR1(2600)(97.0)	1	82.8	29	96.4	9	91.3	7	86.8
NRF1(4200)(98.9)	1	91.5	22	98.6	16	97.3	3	95.6
ATF3(2400)(89.5)	2	74.3	33	89.2	7	70.6	4	70.5
BHLHE40(1000)(81.2)	2	63.0	11	79.0	2	53.3	2	58.8
CEBPB(4000)(90.0)	2	64.7	32	88.5	4	60.5	2	30.6
E2F(8000)(98.8)	2	86.9	51	98.8	11	93.8	8	91.5
ELF1(3000)(93.3)	2	78.8	23	91.9	6	78.7	3	76.3
ETS(8200)(98.2)	2	87.5	55	97.8	18	89.8	9	90.1
FOXA(5000)(92.0)	2	63.0	35	92.3	6	63.0	5	58.4
HNF4(3000)(96.5)	2	76.1	26	96.1	4	72.5	5	77.1
MAF(4000)(97.2)	2	77.7	28	97.2	8	82.7	2	59.5
NFE2(1200)(96.1)	2	93.3	11	96.1	4	88.5	4	87.3
NFKB(10200)(97.3)	2	73.3	54	96.6	20	85.8	4	68.9
NFY(2000)(96.8)	2	93.2	13	97.3	4	92.0	1	83.2
POU2F2(4000)(85.5)	2	59.3	35	85.2	5	60.7	2	55.2
POU5F1(1000)(88.4)	2	76.4	7	86.0	3	68.1	2	74.6
PRDM1(1000)(91.1)	2	79.7	6	91.8	2	28.6	2	73.6
REST(10000)(97.4)	2	81.1	56	96.8	31	94.2	10	92.1
SPI1(3000)(98.7)	2	87.7	20	98.9	7	94.1	3	84.1
SRF(5000)(91.1)	2	70.3	40	90.7	11	81.3	2	57.5
TFAP2(2000)(96.8)	2	85.8	17	97.0	3	88.0	2	83.8
YY1(9200)(95.4)	2	79.1	49	95.2	18	88.0	5	83.1
ZEB1(1000)(85.9)	2	72.8	13	88.8	2	51.5	1	40.6
EBF1(2000)(87.0)	3	75.6	17	87.9	4	68.0	2	59.5
MEF2(2000)(86.0)	3	64.0	18	85.2	4	57.9	3	50.5
MXI1(2000)(88.2)	3	75.4	20	89.3	4	47.1	2	39.1
NR2C2(1600)(93.6)	3	86.8	20	92.5	5	56.6	3	67.9
PAX5(4000)(96.2)	3	74.7	41	95.7	8	77.0	5	71.4
RFX5(3200)(92.0)	3	76.1	33	92.1	6	67.9	3	54.1
SP1(4000)(88.5)	3	73.6	34	88.0	10	75.9	3	68.3
TCF12(2200)(96.3)	3	69.6	29	96.2	4	55.3	6	72.3
ESRRA(4200)(90.4)	4	65.4	44	89.2	9	74.5	4	62.3
GATA(8000)(99.0)	4	77.1	53	98.9	18	88.1	6	66.7
IRF(2650)(97.9)	4	80.5	31	97.7	4	75.5	6	85.3
NR3C1(4250)(95.7)	4	78.7	47	95.4	5	67.9	6	72.6
RXRA(3050)(94.5)	4	71.8	40	94.3	6	65.1	5	61.0
STAT(7200)(99.0)	4	77.4	60	98.7	20	86.6	7	79.6
TCF7L2(2000)(90.1)	4	83.0	12	90.9	4	75.1	2	45.6

P is the total number of peaks selected per TF group and *P* (percent) is the percentage of peaks with motif occurrences. *N* is the number of features selected by each method.

database [45], [1]. The first motif had 63 overlapping SNPs, the second motif had 164 overlapping SNPs, and the third motif had 23 overlapping SNPs. The second motif had one match in the genome-wide association study (GWAS) catalog; rs2293152 [46]. The disease associated with this SNP is multiple sclerosis [46] and the associated gene is STAT3.

TABLE 3
Sequence Sensitivity (Mean, Median, and SD)
across 38 TF Groups

Method	Mean	Median	SD
RILP	93.1	94.7	4.8
Bounded	93.1	94.7	4.8
Greedy	77.0	76.7	8.5
W-ChipMotifs	74.0	75.3	15.8
Enrichment	69.4	70.9	15.9

For the remaining 37 factor groups, Tables 4, 5, 6, and 7 show the motifs that were selected by the greedy algorithm but were not selected by the enrichment method.

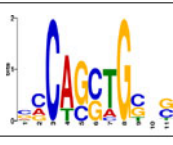
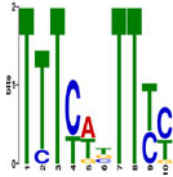



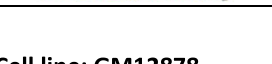

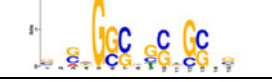

Predicted Motif	sSn(%)	JASPAR Matches	Species, class, family
	45.8	MA0522.1 (Tcf3)	Mus musculus, Other Alpha-Helix, High Mobility Group (Box)
		MA0521.1 (Tcf12)	Mus musculus, Zipper-Type, Helix-Loop-Helix
		MA0500.1 (Myog)	Mus musculus, Zipper-Type, Helix-Loop-Helix
		MA0499.1 (Myod1)	Mus musculus, Zipper-Type, Helix-Loop-Helix
	11.7	MA0517.1 (STAT2:STAT1)	Homo sapiens, Other, STAT
		MA0537.1 (BLMP-1)	Caenorhabditis elegans, Zinc-coordinating, BetaBetaAlpha-zinc finger
		MA0050.2 (IRF1)	Homo sapiens, Winged Helix-Turn-Helix, IRF
		MA0277.1 (AZF1)	Saccharomyces cerevisiae, Zinc-coordinating, BetaBetaAlpha-zinc finger
		MA0508.1 (PRDM1)	Homo sapiens, Zinc-coordinating, BetaBetaAlpha-zinc finger
		MA0554.1 (SOC1)	Arabidopsis thaliana, Other Alpha-Helix, MADS

Fig. 3. Motifs selected by the greedy algorithm for factor group TCF12 with JASPAR matches.

Cell line: HepG2

Motif	sSn(%)	Cumulative Coverage (%)
	30.5	30.5
	30	55
	27	64.5
	17.5	71

Cell line: GM12878

Motif	sSn(%)	Cumulative Coverage (%)
	50.7	50.7
	44.9	69.6
	15.7	77.2

Cell line: H1-hESC



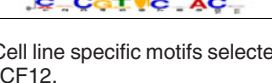
Motif	sSn(%)	Cumulative Coverage (%)
	53.9	53.9
	45.4	70.7
	48.9	76.8






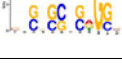









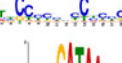





Fig. 4. Cell line specific motifs selected by the greedy algorithm for factor group TCF12.

We believe that these previously unreported motifs are important functional elements of the human genome, due to their ability to provide the simplest explanations for the ChIP-seq binding experiments.

5 CONCLUSIONS

This manuscript presents heuristics that employ the concept of sequence coverage to solve the motif selection problem, yielding a small, concise set of motifs with high coverage of the input sequences. Three motif selection algorithms were implemented and compared: greedy, relaxed integer linear programming (RILP), and bounded exact search. The proposed algorithms were also compared to two existing motif selection methods. The methods were compared in terms of the number of features (motifs) selected and the sequence sensitivity achieved by the chosen motifs. Even though the RILP and bounded exact search algorithms achieve the highest sequence sensitivity, that is obtained at the expense of a high number of motifs selected. Thus, the greedy algorithm is recommended because it produces

TABLE 4
Novel Motifs Discovered by the Greedy Algorithm

TF group	Motif	sSn(%)
BHLHE40		48.0
CEBPB		55.6
		22.1
EBF1		56.2
		35.5
		33.5
EGR1		82.7
ELF1		59.4
		57.0
ESRRA		39.4
		29.5
ETS		73.0
		65.2
FOXA		50.7
		20.7
GATA		40.3
		38.0
		39.0
		32.3
HNF4		66.9
		35.8

a small set of motifs that provides high sequence coverage, enhancing the feasibility of laboratory validation of the reported motifs.




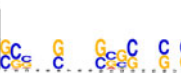














TABLE 5
Novel Motifs Discovered by the Greedy Algorithm

TF group	Motif	sSn(%)
IRF		49.1
		40.6
		12.4
MAF		70.8
		18.8
MEF2		39.1
		24.8
MXI1		55.9
		23.9
		19.6
NFE2		67.6
		29.8
NFKB		57.3
		40.7
NFY		86.8
		40.7
NR2C2		59.3
		47.4
		46.5

TABLE 6
Novel Motifs Discovered by the Greedy Algorithm

TF group	Motif	sSn(%)
NR3C1		41.0
		36.3
		24.1
NRF1		33.6
		91.7
PAX5		43.5
		27.4
POU2F2		36.7
		30.6
POU5F1		20.6
PRDM1		69.8
REST		62.8
RFX5		45.9
		38.6
RXRA		39.0
		27.6
SP1		38.0
		55.2

TABLE 7
Novel Motifs Discovered by the Greedy Algorithm

TF group	Motif	sSn(%)
SPI1		80.5
		33.4
SRF		43.2
		33.3
STAT		36.8
		29.5
		20.6
TCF12		47.0
		43.2
		12.3
TCF7L2		44.9
		38.8
		40.6
TFAP2		58.8
YY1		58.8
		53.7
ZEB1		49.1
		43.3

ACKNOWLEDGMENTS

The authors wish to thank Pouya Kheradpour who provided us with the set of discovered motifs and for helpful

discussions. They also thank Yichao Li, Liang Chen, and Yating Liu from the Ohio University Bioinformatics lab for helpful discussions. This work was supported by the Ohio University Graduate Research and Education Board (GERB).

REFERENCES

- [1] M. A. Schaub, A. P. Boyle, A. Kundaje, S. Batzoglou, and M. Snyder, "Linking disease associations with regulatory information in the human genome," *Genome Res.*, vol. 22, no. 9, pp. 1748–1759, 2012.
- [2] The ENCODE Consortium, "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 489, no. 7414, pp. 57–74, Sep. 2012.
- [3] G. Pavesi, G. Mauri, and G. Pesole, "An algorithm for finding signals of unknown length in DNA sequences," *Bioinformatics*, vol. 17, no. suppl. 1, pp. S207–S214, 2001.
- [4] F. Zambelli, G. Pesole, and G. Pavesi, "Motif discovery and transcription factor binding sites before and after the next-generation sequencing era," *Briefings Bioinf.*, vol. 14, no. 2, pp. 225–37, Mar. 2013.
- [5] H. Hartmann, E. W. Guthöhrlein, M. Siebert, S. Luehr, and J. Söding, "P-value-based regulatory motif discovery using positional weight matrices," *Genome Res.*, vol. 23, no. 1, pp. 181–94, Jan. 2013.
- [6] T. L. Bailey and C. Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers," in *Proc. Int. Conf. Intell. Syst. Molecular Biol.; Int. Conf. Intell. Syst. Molecular Biol.*, Jan. 1994, vol. 2, pp. 28–36.
- [7] S. Sinha, "Discriminative motifs," *J. Comput. Biol.*, vol. 10, nos. 3/4, pp. 599–615, Jan. 2003.
- [8] P. Huggins, S. Zhong, I. Shiff, R. Beckerman, O. Laptchenko, C. Prives, M. H. Schulz, I. Simon, and Z. Bar-Joseph, "DECOD: fast and accurate discriminative DNA motif finding," *Bioinformatics*, vol. 27, no. 17, pp. 2361–2367, Jul. 2011.
- [9] E. Redhead and T. L. Bailey, "Discriminative motif discovery in DNA and protein sequences using the DEME algorithm," *BMC Bioinf.*, vol. 8, p. 385, Jan. 2007.
- [10] A. D. Smith, P. Sumazin, and M. Q. Zhang, "Identifying tissue-selective transcription factor binding sites in vertebrate promoters," *Proc. Nat. Acad. Sci. US America*, vol. 102, no. 5, pp. 1560–1565, Feb. 2005.
- [11] M. Tompa, N. Li, T. L. Bailey, G. M. Church, B. De Moor, E. Eskin, A. V. Favorov, M. C. Frith, Y. Fu, W. J. Kent et al., "Assessing computational tools for the discovery of transcription factor binding sites," *Nature Biotechnol.*, vol. 23, no. 1, pp. 137–144, 2005.
- [12] P. Kheradpour and M. Kellis, "Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments," *Nucleic Acids Res.*, vol. 42, pp. 1–12, Dec. 2013.
- [13] F. P. Roth, J. D. Hughes, P. W. Estep, and G. M. Church, "Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation," *Nature Biotechnol.*, vol. 16, no. 10, pp. 939–945, 1998.
- [14] X. S. Liu, D. L. Brutlag, and J. S. Liu, "An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments," *Nature Biotechnol.*, vol. 20, no. 8, pp. 835–839, 2002.
- [15] L. Ettwiller, B. Paten, M. Ramialison, E. Birney, and J. Wittbrodt, "Trawler: De novo regulatory motif discovery pipeline for chromatin immunoprecipitation," *Nature Methods*, vol. 4, no. 7, pp. 563–565, 2007.
- [16] V. X. Jin, J. Apostolos, N. S. V. R. Nagisetty, and P. J. Farnham, "W-ChIPMotifs: A web application tool for de novo motif discovery from ChIP-based high-throughput data," *Bioinformatics*, vol. 25, no. 23, pp. 3191–3193, 2009.
- [17] B. A. Kennedy, X. Lan, T. H.-M. Huang, P. J. Farnham, and V. X. Jin, "Using ChIPMotifs for de novo motif discovery of OCT4 and ZNF263 based on ChIP-based high-throughput experiments," in *Next Generation Microarray Bioinformatics*. New York, NY, USA: Springer, 2012, pp. 323–334.
- [18] L. S. Hon and A. N. Jain, "A deterministic motif finding algorithm with application to the human genome," *Bioinformatics*, vol. 22, no. 9, pp. 1047–1054, 2006.
- [19] L. Kuttippurathu, M. Hsing, Y. Liu, B. Schmidt, D. L. Maskell, K. Lee, A. He, W. T. Pu, and S. W. Kong, "CompleteMOTIFs: DNA motif discovery platform for transcription factor binding experiments," *Bioinformatics*, vol. 27, no. 5, pp. 715–717, 2011.

- [20] I. V. Kulakovskiy, V. Boeva, A. V. Favorov, and V. Makeev, "Deep and wide digging for binding motifs in ChIP-Seq data," *Bioinformatics*, vol. 26, no. 20, pp. 2622–2623, 2010.
- [21] S. J. van Heeringen and G. J. C. Veenstra, "GimmeMotifs: A de novo motif prediction pipeline for ChIP-sequencing experiments," *Bioinformatics*, vol. 27, no. 2, pp. 270–271, 2011.
- [22] X. Liu, D. L. Brutlag, J. S. Liu et al., "BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes," in *Proc. Pacific Symp. Biocomput.*, 2001, vol. 6, no. 2001, pp. 127–138.
- [23] L. Li, "GADEM: A genetic algorithm guided formation of spaced dyads coupled with an EM algorithm for motif discovery," *J. Comput. Biol.*, vol. 16, no. 2, pp. 317–329, 2009.
- [24] W. Ao, J. Gaudet, W. J. Kent, S. Muttumu, and S. E. Mango, "Environmentally induced foregut remodeling by pPHA-4/FoxA and DAF-12/NHR," *Science*, vol. 305, no. 5691, pp. 1743–1746, 2004.
- [25] X. S. Liu, D. L. Brutlag, and J. S. Liu, "An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments," *Nature Biotechnol.*, vol. 20, no. 8, pp. 835–839, 2002.
- [26] E. Valen, A. Sandelin, O. Winther, and A. Krogh, "Discovery of regulatory elements is improved by a discriminatory approach," *PLoS Comput. Biol.*, vol. 5, no. 11, p. e1000562, 2009.
- [27] G. Thijs, K. Marchal, M. Lescot, S. Rombauts, B. De Moor, P. Rouz , and Y. Moreau, "A gibbs sampling method to detect over-represented motifs in the upstream regions of coexpressed genes," *J. Comput. Biol.*, vol. 9, no. 2, pp. 447–464, 2002.
- [28] A. Chakravarty, J. M. Carlson, R. S. Khetani, and R. H. Gross, "A novel ensemble learning method for de novo computational identification of DNA binding sites," *BMC Bioinf.*, vol. 8, no. 1, p. 249, 2007.
- [29] V. Martyanov and R. H. Gross, "Using SCOPE to identify potential regulatory motifs in coregulated genes," *J. Visualized Experiments*, vol. 51, pp. 1–7, 2011.
- [30] K. Klepper and F. Drabl s, "MotifLab: A tools and data integration workbench for motif discovery and regulatory sequence analysis," *BMC Bioinf.*, vol. 14, p. 9, 2013.
- [31] J. Hu, Y. D. Yang, and D. Kihara, "EMD: An ensemble algorithm for discovering regulatory motifs in DNA sequences," *BMC Bioinf.*, vol. 7, no. 1, p. 342, 2006.
- [32] E. Wijaya, S.-M. Yiu, N. T. Son, R. Kanagasabai, and W.-K. Sung, "MotifVoter: A novel ensemble method for fine-grained integration of generic motif finders," *Bioinformatics*, vol. 24, no. 20, pp. 2288–2295, 2008.
- [33] E. Eskin and P. A. Pevzner, "Finding composite regulatory patterns in DNA sequences," *Bioinformatics*, vol. 18, no. suppl. 1, pp. S354–S363, 2002.
- [34] E. Wijaya, K. Rajaraman, S.-M. Yiu, and W.-K. Sung, "Detection of generic spaced motifs using submotif pattern mining," *Bioinformatics*, vol. 23, no. 12, pp. 1476–1485, 2007.
- [35] C. Workman and G. Stormo, "ANN-spec: A method for discovering transcription factor binding sites with improved specificity," in *Proc. Pacific Symp. Biocomput.*, 2000, vol. 5, pp. 464–475.
- [36] R. C. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Res.*, vol. 32, no. 5, pp. 1792–1797, 2004.
- [37] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, 1st edit ed. San Francisco, CA, USA: Freeman, 1979.
- [38] (2015). GLPK GNU linear programming kit [Online]. Available: <http://www.gnu.org/software/glpk/>
- [39] V. V. Vazirani, *Approximation Algorithms*. New York, NY, USA: Springer-Verlag, 2001.
- [40] S. Gupta, J. A. Stamatoyannopoulos, T. L. Bailey, and W. S. Noble, "Quantifying similarity between motifs," *Genome Biol.*, vol. 8, no. 2, p. R24, Jan. 2007.
- [41] C. E. Grant, T. L. Bailey, and W. S. Noble, "FIMO: Scanning for occurrences of a given motif," *Bioinformatics*, vol. 27, no. 7, pp. 1017–1018, Apr. 2011.
- [42] C.-C. Lee, W.-S. Chen, C.-C. Chen, L.-L. Chen, Y.-S. Lin, C.-S. Fan, and T.-S. Huang, "TCF12 protein functions as transcriptional repressor of E-cadherin, and its overexpression is correlated with metastasis of colorectal cancer," *J. Biological Chemistry*, vol. 287, no. 4, pp. 2798–2809, 2012.
- [43] J.-S. Hu, E. Olson, and R. Kingston, "HEB, a helix-loop-helix protein related to E2A and ITF2 that can modulate the DNA-binding ability of myogenic regulatory factors," *Molecular Cellular Biol.*, vol. 12, no. 3, pp. 1031–1042, 1992.
- [44] Y. Zhang, J. Babin, A. L. Feldhaus, H. Singh, P. A. Sharp, and M. Bina, "HTF4: A new human helix-loop-helix protein," *Nucleic Acids Res.*, vol. 19, no. 16, p. 4555, 1991.
- [45] A. P. Boyle, E. L. Hong, M. Hariharan, Y. Cheng, M. A. Schaub, M. Kasowski, K. J. Karczewski, J. Park, B. C. Hitz, S. Weng et al., "Annotation of functional variation in personal genomes using RegulomeDB," *Genome Res.*, vol. 22, no. 9, pp. 1790–1797, 2012.
- [46] T. Burdett, P. Hall, E. Hasting, L. Hindorf, H. Junkins, A. Klemm, J. MacArthur, T. Manolio, J. Morales, H. Parkinson, and D. Welter. The NHGRI-EBI catalog of published genome-wide association studies [Online]. Available: www.ebi.ac.uk/gwas



Rami Al-Ouran received the BS degree in computer engineering from Mutah University, Jordan, and the MS degree in electrical engineering and computer science from Ohio University. He is currently working toward the PhD degree in electrical engineering and computer science at Ohio University. His research interests include bioinformatics and machine learning.



Robert Schmidt received the BS degree in computer science and engineering from The Ohio State University in 2012 and is currently working toward the MS degree in computer science at Ohio University. His research interests lie in bioinformatics, specifically in the analysis of genomic data and the identification of functional elements in genomic data.



Ashwini Naik received the BS degree in computer science from the M.V.S.R Engineering College, Andhra Pradesh, India, and the MS degree in computer science from Ohio University. She is currently a bioinformatics systems analyst at The Research Institute at Nationwide Children's Hospital. Her role involves the research and development of computational algorithms for the analysis and interpretation of next generation sequencing data, including human genome and exome sequencing and variant analysis, bacterial

genome sequencing, and workflow automation.



Jeffrey Jones received the BS degree in computer science from Ohio University in 1981, the MS degree in computer and information science from The Ohio State University in 1986, and the PhD degree in electrical engineering and computer science from Ohio University in 2015. He is currently a senior lecturer with The Ohio State University. Prior to his academic career, he served for 20 years as a vice president and president of Great Northern Consulting Services, Inc. During his early professional career, he was a systems engineer for Sun Microsystems as well as a member of technical staff for AT&T. He enjoys interdisciplinary research, and has current projects in the areas of high-performance computing and bioinformatics.



Frank Drews is a professor in the Electrical Engineering and Computer Science Department, Ohio University. He researches high-performance computing, real-time systems, and bioinformatics. He is an associate editor and member of the editorial board of the *International Journal of Computational Bioscience*, and he was a guest editor for the *Journal of Systems and Software's* Special Issue on Resource Management for Real-Time and Distributed Systems.



David Juedes received the BS degree in computer science and mathematics from the University of Wisconsin-La Crosse in 1988, and the MS and PhD degrees in computer science from Iowa State University in 1990 and 1994. He is a professor and chair in the School of Electrical Engineering and Computer Science, Ohio University. He performs research in algorithms and complexity theory. He is a senior member of the Association for Computing Machinery.



Laura Elnitski received the BS degree in molecular and cellular biology at The Pennsylvania State University (Penn State), with specialty research in chemical engineering. She received the PhD degree in biochemistry and molecular biology, also at Penn State, while pursuing one of the first projects to look at multispecies comparisons of noncoding regulatory elements. She joined the National Human Genome Research Institute in 2005 as a tenure track investigator.

She has participated in numerous genome sequencing projects including mouse, rat, cow, and chicken, as well as the ENCODE Consortium to elucidate functional elements in the human genome. Her work specializes in developing tools to identify and discern the mechanistic action of functional elements in the human genome



Lonnie Welch received the BS, MS, and PhD degrees in 1985, 1987, and 1990, respectively, in computer and information science from the Ohio State University. He is the Stuckey professor of electrical engineering and computer science at Ohio University and the director in the Bioinformatics Laboratory at Ohio University. He is a member of the Board of Directors in the International Society for Computational Biology (ISCB), is the founder and Steering Committee Chair of the ISMB Regulatory Genomics SIG, and is the founder and Steering Committee chair of the Great Lakes Bioinformatics Conference (an official ISCB conference).

▷ **For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/publications/dlib.**