LASAGNA-Search: an integrated web tool for transcription factor binding site search and visualization

Chih Lee and Chun-Hsi Huang
Department of Computer Science and Engineering, University of Connecticut, Storrs, CT, USA

BioTechniques 54:141-153 (March 2013) doi 10.2144/000113999

Keywords: transcription factor binding site; motif search; gene regulatory network; visualization
Supplementary material for this article is available at www.BioTechniques.com/article/113999.

The release of ChIP-seq data from the ENCyclopedia Of DNA Elements (ENCODE) and Model Organism ENCyclopedia Of DNA Elements (modENCODE) projects has significantly increased the amount of transcription factor (TF) binding affinity information available to researchers. However, scientists still routinely use TF binding site (TFBS) search tools to scan unannotated sequences for TFBSs, particularly when searching for lesser-known TFs or TFs in organisms for which ChIP-seq data are unavailable. The sequence analysis often involves multiple steps such as TF model collection, promoter sequence retrieval, and visualization; thus, several different tools are required. We have developed a novel integrated web tool named LASAGNA-Search that allows users to perform TFBS searches without leaving the web site. LASAGNA-Search uses the LASAGNA (Length-Aware Site Alignment Guided by Nucleotide Association) algorithm for TFBS alignment. Important features of LASAGNA-Search include (i) acceptance of unaligned variable-length TFBSs, (ii) a collection of 1726 TF models, (iii) automatic promoter sequence retrieval, (iv) visualization in the UCSC Genome Browser, and (v) gene regulatory network inference and visualization based on binding specificities. LASAGNA-Search is freely available at http://biogrid.engr.uconn.edu/lasagna_search/.

Transcription factors (TF) regulate their target genes by physically binding to the gene regulatory regions. TF binding to DNA is sequence specific; thus binding sites for a specific TF share common sequence patterns called motifs. Given a set of known binding sites for a particular TF, these motifs can be used to search unannotated promoters to identify putative transcription factor binding sites (TFBSs), providing an inexpensive alternative to experimental determination of TFBSs in a wet lab (1–4).

Various methods have been proposed for motif modeling. A consensus model summarizes binding sites by the consensus sequence (5), while a position-specific weight matrix (PWM) model summarizes TFBSs by a scoring matrix (6). Some extensions of these simple methods do not rely on the assumption of position independence and instead score nucleotide pairs (7,8). Other methods model position dependence by first-order Markov chains (9), profile hidden Markov models (10), or principal components (11). SiTaR(12), on the other hand, does not summarize TFBSs but instead uses input motifs to identify TFBSs in the query data set.

Despite the more sophisticated models proposed over the past decades, the PWM method remains a simple and widely-used TFBS search approach. It represents a motif by a $4 \times l$ matrix, where $l$ is the length of binding sites and each row corresponds to one of the four nucleotide bases. Column $i$ of the matrix keeps the scores of matching the $i$th letter of a length-$l$ sequence (an $l$-mer) to nucleotides A, C, G, and T, respectively. To score an $l$-mer, the PWM method sums up the scores of the $l$ individual letters. Databases such as JASPAR (13), TRANSFAC (14), and UniPROBE (15) store matrices of TFs that can be easily used by web tools implementing the PWM method or its variants (16–20). A
stored matrix usually consists of counts or probabilities and can be easily converted to be compatible with different scoring schemes.

A PWM is usually built from experimentally validated binding sites – DNA segments that can be physically bound by a TF but may or may not be functional. These binding sites often vary in length and are not aligned. In cases where a transcription factor does not have available PWMs, researchers must resort to studying its binding sites. Nearly 38.1% of the TFs we found in the TRANSFAC Public database do not have PWMs available, so their binding sites have to be aligned. Since TRANSFAC may build more than one PWM for a TF, a lack of matrices for TFs that recognize more than one motif is unlikely to occur (21). Open-annotation databases such as ORegAnno (22) and PAZAR (23) contain valuable user-curated TFBSs. To utilize binding sites in the ORegAnno database, one has to align them before building PWMs. The PAZAR database, on the other hand, is another important resource because it dynamically creates PWMs for users using MEME (24). Although PWMs represent motifs in a compact form, information about position dependence is lost when converting TFBS alignments to PWMs. It has been shown in many studies (7–9,11), that position dependence significantly improves the search performance of a method. For this reason, TFBS alignments may be preferred compared to PWMs.

A typical TFBS search web tool takes a PWM and promoter sequence as inputs and returns putative binding sites. Many web tools include useful features in addition to the basic search function. Some accept variable-length binding sites (9,12,25), offer precomputed models built from PWMs or TFBSs (10,13,17,26–28), adopt a TFBS search method that exploits position dependence (9,10), offer promoter sequence retrieval, or integrate a sequence retrieval tool (10,18,25,26,27). The MAPPER2 database (10) supports visualization of hits in the UCSC Genome Browser (29) for three organisms. Another useful function is visual representation of predicted binding specificities as a gene regulatory network (GRN) (28). Until now, there was no single web tool that incorporated all the aforementioned features.

We created a web tool for TFBS search and visualization that we call LASAGNA-Search. LASAGNA-Search accepts variable-length TFBSs in addition to PWMs. It offers 1726 precomputed models based on TFBSs and PWMs collected from the TRANSFAC Public, JASPAR, ORegAnno and UniPROBE databases. Its search module exploits position dependence for a TFBS-based model whenever performance gain is indicated by cross-validation. Automatic promoter sequence retrieval is supported for seven organisms at LASAGNA-Search, which enables visualization of search results in the UCSC Genome Browser. Search results can also be visualized along promoter sequences locally at LASAGNA-Search for any organism. Finally, a GRN can be constructed from search results and visualized locally with various options.

Materials and methods

Figure 1 shows the architecture of LASAGNA-Search. We introduce the major components in the following sections.

Alignment module

The alignment module aligns variable-length TFBSs to build a TF model. This module has been extensively compared to ClustalW2 (30) and MEME (24) with favorable outcomes (see Supplementary Figure S2). These two methods were chosen because they are widely-used representatives of two different types of TFBS identification methods. ClustalW2, which is based on pairwise sequence similarity, was used for TFBS alignment in several studies (9,11,31). MEME is a commonly used de novo motif discovery tool (24). We describe
the key ideas behind the algorithm and include the technical details in the Supplementary Methods.

A binding site includes a core region – a short stretch of DNA to which the TF actually binds – flanked by a few bases on each side. The core region may not be determined accurately if the resolution of the binding site calling technique is not sufficient. When progressively aligning binding sites, the order in which the sites are aligned is important. We observed that aligning binding sites from the shortest to the longest generally yields better alignments. Shorter binding sites tend to contain fewer non-informative bases flanking the core region. Therefore, we use TFBS length to guide the alignment process. Extensive experiments have validated the efficacy of this algorithm.

### Search module

The search module uses a TF model and promoter sequence as inputs. The TF model is a PWM variant, which scores sequences of length \( l \). Depending on the TF, scores of nucleotide pairs may contribute to the score of a sequence, controlled by the parameter of \( K \geq 0 \), the maximal distance between a nucleotide pair. The value of \( K \) is TF-dependent and

**Figure 2.** An inferred gene regulatory network consisting of the human genes TP53 and MYB. A small gene regulatory network made up of human genes TP53 and MYB inferred from scanning the promoters (950bp upstream to 50bp downstream) using two TP53 TF models and two MYB TF models. Genes are denoted by green ellipses and TF models are represented by red octagons. (A) The inferred network containing the 2 genes and the 4 TF models. (B) The dotted edges from the 2 coding genes to the 4 TF models are established in this network. (C) The simplified network after removing 4 nodes. These nodes are removed because the two TP53 TF models are coded by the TP53 gene and the two MYB TF models are coded by the MYB gene.

**Table 1. Summary of TF Model Collections.**

<table>
<thead>
<tr>
<th>Database</th>
<th>Type</th>
<th>Models</th>
<th>Mapped Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRANSFAC</td>
<td>TFBS</td>
<td>189</td>
<td>188</td>
</tr>
<tr>
<td>ORegAnno</td>
<td>TFBS</td>
<td>133</td>
<td>132</td>
</tr>
<tr>
<td>TRANSFAC</td>
<td>PWM</td>
<td>398</td>
<td>366</td>
</tr>
<tr>
<td>JASPAR CORE</td>
<td>PWM</td>
<td>476</td>
<td>457</td>
</tr>
<tr>
<td>UniPROBE</td>
<td>PWM</td>
<td>530</td>
<td>524</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>1726</td>
<td>1667</td>
</tr>
</tbody>
</table>

1. Models of TFs whose coding genes were found.
is determined by cross-validation. Hence, $K$ is greater than zero only if nucleotide pairs improve the search performance for a TF. We refer readers to the Supplementary Methods for additional technical details.

It is commonly assumed that the first letter of an $l$-mer sequence is aligned with the first position of a TF model binding site and the $l$-mer is scored accordingly. Unlike more conventional approaches, we align an $l$-mer with a TF model by sliding an $l$-mer and its reverse-complement through the model such that the overlap between the two is at least one nucleotide using the framework described in the section.

**Evaluation of Precomputed TF Models.** We found that this is significantly better than the conventional approach for locating TFBSs (see Supplementary Figure S3). Moreover, this approach allows easy scoring of an $l$-mer by a cluster of TF models of different widths. Scoring with a cluster of TF models has been shown to outperform using only the best model in the cluster (32) and hence is a feature to be added to LASAGNA-Search in the near future.

For each putative binding site hit, the search module computes the score and the $p$-value indicating the probability of observing a score equal to or higher than the score by chance. We describe the $p$-value computation in the Supplementary Methods. While $p$-values are not corrected for multiple testing, they are useful for ordering hits found by different TF models. To take into account the length of the promoter sequence in which a hit is found, an $E$-value is computed for the hit. The $E$-value gives the expected number of times a hit of the same or higher score is found in the promoter sequence by chance. If $L$ is the length of the promoter sequence and $l$ is the length of the putative binding site, then $E$-value $= p$-value $\times (L - l + 1)$, which is approximately $p$-value $\times L$ when $L \gg l$.

**Promoter retrieval module.** Currently, LASAGNA-Search supports retrieving promoter sequences for seven species: *Homo sapiens, Mus musculus, Rattus norvegicus, Drosophila melanogaster, Saccharomyces cerevisiae, Caenorhabditis elegans*, and *Caenorhabditis briggsae*. Users may enter the NCBI Gene identifier (ID), the official gene symbol or an mRNA accession number of a gene to retrieve its upstream promoter region. The upstream region of a gene is specified by positions relative to the transcription start site (TSS) obtained from the UCSC Genome Browser (29). Information in the NCBI Gene database is used for conversion between Gene IDs and symbols.

![Figure 3. LASAGNA-Search input page user interface.](image-url)
GRN inference

LASAGNA-Search automatically constructs a network based on search results. A directed edge from a TF model to a gene is established if at least one significant hit is found in the promoter region of the gene by the TF model. The lowest p-value of these hits is used to compute the weight on this edge. That is, the thickness of the edge is proportional to -log p-value. In cases where the coding genes of a TF model are known, these genes may be added to the network with dotted arrows from the genes to the TF model. To simplify the network, the node for a TF model may be removed, leaving only its coding genes in the network. Figure 2 shows an example network of human genes TP53 and MYB. Visualization of GRNs at LASAGNA-Search is enabled by Cytoscape Web (33). We describe how the networks in Figure 2 were generated in the section titled User Interface.

TF model collections

LASAGNA-Search currently offers five precomputed TF model collections. The collections are categorized by the type of data used to build a model. Table 1 lists the type and number of models for each collection. To facilitate GRN visualization, we mapped TF models to genes coding for the TFs. The number of models that can be mapped for each collection is also listed in Table 1. Models in the TFBS-based collections were built from unaligned TFBSs, while models in the PWM-based collections were built from PWMs. We describe these two categories in the following sections.

TFBS-based collections

We collected experimentally validated transcription factor binding sites from the TRANSFAC Public database and the ORegAnno database. In these two collections, binding sites of a TF were not collected across organisms. TF models are non-redundant in the sense that a TF of a species has only one model based on all the available binding sites in a database. The binding sites of a TF were aligned to build a model. We built one model for each TF because, for most TFs, the binding affinity can be explained by only one model (34). In case a TF recognizes more than one motif (21), we rely on database curators to distinguish binding sites belonging to distinct motifs. Moreover, the TFBS-based collections are complemented by our PWM-based collections, which offer more than one model for some TFs.

Binding sites for five organisms were collected from the TRANSFAC Public database (release 7.0) (14), including Homo sapiens, Mus musculus, Rattus norvegicus, Drosophila melanogaster and Saccharomyces cerevisiae. For each organism, a TF was included in our collection if it contained at least 10 binding sites. Binding sites for 189 TFs across the five species were collected. Although TRANSFAC builds PWMs, 72 (38.1%) of the TFs did not have PWMs in TRANSFAC.

In addition to the five organisms

Binding sites for five organisms were collected from the TRANSFAC Public database (release 7.0) (14), including Homo sapiens, Mus musculus, Rattus norvegicus, Drosophila melanogaster and Saccharomyces cerevisiae. For each organism, a TF was included in our collection if it contained at least 10 binding sites. Binding sites for 189 TFs across the five species were collected. Although TRANSFAC builds PWMs, 72 (38.1%) of the TFs did not have PWMs in TRANSFAC.

In addition to the five organisms

Binding sites for five organisms were collected from the TRANSFAC Public database (release 7.0) (14), including Homo sapiens, Mus musculus, Rattus norvegicus, Drosophila melanogaster and Saccharomyces cerevisiae. For each organism, a TF was included in our collection if it contained at least 10 binding sites. Binding sites for 189 TFs across the five species were collected. Although TRANSFAC builds PWMs, 72 (38.1%) of the TFs did not have PWMs in TRANSFAC.

In addition to the five organisms

Binding sites for five organisms were collected from the TRANSFAC Public database (release 7.0) (14), including Homo sapiens, Mus musculus, Rattus norvegicus, Drosophila melanogaster and Saccharomyces cerevisiae. For each organism, a TF was included in our collection if it contained at least 10 binding sites. Binding sites for 189 TFs across the five species were collected. Although TRANSFAC builds PWMs, 72 (38.1%) of the TFs did not have PWMs in TRANSFAC.

In addition to the five organisms

Binding sites for five organisms were collected from the TRANSFAC Public database (release 7.0) (14), including Homo sapiens, Mus musculus, Rattus norvegicus, Drosophila melanogaster and Saccharomyces cerevisiae. For each organism, a TF was included in our collection if it contained at least 10 binding sites. Binding sites for 189 TFs across the five species were collected. Although TRANSFAC builds PWMs, 72 (38.1%) of the TFs did not have PWMs in TRANSFAC.

In addition to the five organisms

Binding sites for five organisms were collected from the TRANSFAC Public database (release 7.0) (14), including Homo sapiens, Mus musculus, Rattus norvegicus, Drosophila melanogaster and Saccharomyces cerevisiae. For each organism, a TF was included in our collection if it contained at least 10 binding sites. Binding sites for 189 TFs across the five species were collected. Although TRANSFAC builds PWMs, 72 (38.1%) of the TFs did not have PWMs in TRANSFAC.

In addition to the five organisms

Binding sites for five organisms were collected from the TRANSFAC Public database (release 7.0) (14), including Homo sapiens, Mus musculus, Rattus norvegicus, Drosophila melanogaster and Saccharomyces cerevisiae. For each organism, a TF was included in our collection if it contained at least 10 binding sites. Binding sites for 189 TFs across the five species were collected. Although TRANSFAC builds PWMs, 72 (38.1%) of the TFs did not have PWMs in TRANSFAC.

In addition to the five organisms

Binding sites for five organisms were collected from the TRANSFAC Public database (release 7.0) (14), including Homo sapiens, Mus musculus, Rattus norvegicus, Drosophila melanogaster and Saccharomyces cerevisiae. For each organism, a TF was included in our collection if it contained at least 10 binding sites. Binding sites for 189 TFs across the five species were collected. Although TRANSFAC builds PWMs, 72 (38.1%) of the TFs did not have PWMs in TRANSFAC.

In addition to the five organisms

Binding sites for five organisms were collected from the TRANSFAC Public database (release 7.0) (14), including Homo sapiens, Mus musculus, Rattus norvegicus, Drosophila melanogaster and Saccharomyces cerevisiae. For each organism, a TF was included in our collection if it contained at least 10 binding sites. Binding sites for 189 TFs across the five species were collected. Although TRANSFAC builds PWMs, 72 (38.1%) of the TFs did not have PWMs in TRANSFAC.

In addition to the five organisms

Binding sites for five organisms were collected from the TRANSFAC Public database (release 7.0) (14), including Homo sapiens, Mus musculus, Rattus norvegicus, Drosophila melanogaster and Saccharomyces cerevisiae. For each organism, a TF was included in our collection if it contained at least 10 binding sites. Binding sites for 189 TFs across the five species were collected. Although TRANSFAC builds PWMs, 72 (38.1%) of the TFs did not have PWMs in TRANSFAC.
As seen in Table 1, nearly all of the TF models in the two collections were mapped to TF coding genes. Only one model in each collection remained unmapped due to lack of information in the source databases: ETF (T00270) in TRANSFAC and MYF in ORegAnno.

**PWM-based collections**

In addition to binding sites, we also collected position-specific weight matrices (PWMs) from the TRANSFAC Public database, the JASPAR CORE database (13) and the UniPROBE database (15). A PWM is a $4 \times l$ matrix, where $l$ is the length of the binding sites. Each element in column $i$ of a PWM usually contains the count or probability of a nucleotide at position $i$. PWMs are valuable resources for a number of reasons. Most PWMs in TRANSFAC and JASPAR were built by domain experts. For instance, some PWMs in TRANSFAC and JASPAR were collected from the TRANSFAC Public database (e.g., TR ANSFAC matrix M00152). Moreover, a PWM in TRANSFAC may be based on binding sites of two or more TFs having similar binding specificities (e.g., TRANSFAC matrix M00158). Another reason that PWMs are particularly valuable is that some techniques produce only matrix data. The UniPROBE database, for example, stores data from protein binding microarray (PBM) experiments (35). The PBM technique assigns a binding specificity score to each 10-mer sequence variant. Berger and Bulyk (35), however, do not suggest setting a specificity cut-off threshold to report binding sites. Instead, PWMs are produced by the Seed-and-Wobble algorithm.

From the UniPROBE database, we collected 530 PWMs from six species: Homo sapiens, Mus musculus, Saccharomyces cerevisiae, Caenorhabditis elegans, Plasmodium falciparum and Cryptosporidium parvum. These 530 PWMs correspond to 414 non-redundant TFs (proteins or protein complexes). We collected 476 PWMs from the JASPAR CORE database, where the PWMs were categorized into six species groups: vertebrates, insects, plants, fungi, nematodes and urochordates. Finally, 398 PWMs were collected from the TRANSFAC Public database and grouped into the following categories: vertebrates, insects, plants, fungi, nematodes, and Bacteria.

According to Table 1, the PWM-based collections contain more unmapped TF models than the TFBS-based collections because some source databases lack information. Matrices such as MA0102.1 and MA0061.1 in the JASPAR CORE database were built from TFBSs of more than one organism, but accession numbers for the homologous proteins are not available. Some matrices in the TRANSFAC and JASPAR CORE databases have protein accession numbers, but records of the corresponding coding genes cannot be found in the NCBI Gene database. These proteins often belong to species such as *Pisum sativum* and *Triticum aestivum*, which are not as well-studied as model organisms.

**Results and discussion**

**Input page**

The LASAGNA-Search input page is divided into three parts: TF model input, promoter sequence input, and result filtering. Figure 3A shows a screenshot of the input page. Two options are available for result filtering. One is to set a p-value threshold so that only hits with equal or lower p-values will be reported. The other is to set k so that only the k hits with the highest scores will be reported.

For TF model input, LASAGNA-Search accepts variable-length TFBSs for model building. Users may input TFBSs in the FASTA format. Clicking the “Start Searching” button aligns the TFBSs. The PWM and sequence logo (36) of the automatically trimmed alignment will be displayed. Users may choose to further trim the alignment or recover previously trimmed columns. Figure 3C shows the user interface for TFBS alignment trimming. In addition to TFBSs, users may input a PWM for model building. LASAGNA-Search recognizes formats used by JASPAR, TRANSFAC and UniPROBE.

LASAGNA-Search currently offers two ways of selecting models in the TFBS-based and PWM-based collec-
Promoter sequences may be input in the FASTA format. Users may also retrieve promoter sequences by NCBI Gene IDs, gene symbols, or mRNA accession numbers. By clicking the “Search” button, LASAGNA-Search will display the matching promoters. Figure 3D shows the promoters found using keywords CCND1 and MYB. Users may choose to examine only promoters of a particular organism. In figure 3D, only the matching human promoters are listed after applying the filter. Promoters are selected in a manner similar to selecting TF models. Finally, users may also select from a list of randomly sampled promoters from a chosen organism.

Results page
The results page is organized into five tabs. The first tab displays hits on all the promoter sequences; the second tab displays hits pertaining to one promoter sequence at a time; the third tab shows the GRN inferred from the search results; the fourth tab allows for importing previous search results to be merged with the current search results; the last tab contains the inputs, including the selected TF models, the selected promoters, and the search parameters. Figure 4 shows an example results page with the third tab named “Promoter view” showing.

Only hits meeting the specified criterion are reported in the first and second tabs. For each hit, the model name, sequence, zero-based position, strand, score, \(p\)-value, and \(E\)-value are reported. Hits found in the same promoter sequence can be sorted by model name, sequence, position, strand, \(p\)-value, and \(E\)-value by clicking the respective column header. By default, the hits are displayed in an HTML table. Users may click a button on the results page to obtain the hits in a tab-delimited format. These hits can be easily imported into a new search session. This is particularly useful when additional TFs of interests are identified after an initial search.

Table 2. Comparing LASAGNA-Search to Other TFBS Search Web Tools.

<table>
<thead>
<tr>
<th>TF Model</th>
<th>LASAGNA-Search</th>
<th>matrix-scan</th>
<th>MAPPER2</th>
</tr>
</thead>
<tbody>
<tr>
<td>User PWM</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>User TFBS</td>
<td>Yes, unaligned, variable-length TFBSs.</td>
<td>No</td>
<td>Yes, aligned TFBSs.</td>
</tr>
<tr>
<td>Model collection</td>
<td>Yes, 1726 TFBS-based and PWM-based models.</td>
<td>Not available</td>
<td>Yes, 1017 TFBS-based models.</td>
</tr>
</tbody>
</table>

| Promoter Sequence | | |
|-------------------|-----------------|-------------|---------|
| Format | FASTA | FASTA and 5 other formats. | FASTA |
| Retrieval tool | Yes, built-in for 7 species. | Yes. Retrieve sequence and retrieve EnsEMBL sequence on the same web site. | Yes, built-in for 3 species. |
| Filtering | \(p\)-value | \(p\)-value | \(E\)-value |
| Local visualization | Yes | Yes | Yes |
| UCSC Genome Browser Visualization | Yes, supports 7 species. | Limited. Build, coordinates and orientation must be specified in FASTA sequence header. | Yes, supports 3 species. |
| GRN inference | Yes | No | No |

Figure 6. Comparison of precomputed TF models.
Plots of the performance of MAPPER2 compared to that of LASAGNA-Search for (A) human TFs, (B) mouse TFs and (C) all the TFs. Each point in a plot corresponds to a TF whose binding sites can be predicted by more than one model. Each model is scored by average precision and the average score across the models is used to plot the point. The outlier, Mafk, in (B) is marked in red. The number of TFs and the Wilcoxon signed-rank test \(p\)-value are shown for each plot.
Users may display the hits along the promoter sequence, where the \(-\log p\)-value of each hit is used as the height to plot a box. This allows easy visualization of the predicted binding sites for a model in the context of other models. Finally, the hits can be saved in GFF (general feature format) or bedGraph format for visualization in the UCSC Genome Browser (29). Links are provided for each promoter sequence to automatically create a custom track that redirect users to the UCSC Genome Browser. Figure 5 shows a custom track of putative binding sites predicted by LASAGNA-Search in the context of four other relevant tracks.

The automatically inferred GRN can be displayed and manipulated by clicking the tab named “Gene regulatory network.” To produce a sparser network, users may set a more stringent p-value than the one used to filter hits. Users may show only nodes belonging to one or more species listed under “Filter by species.” Figure 2A shows the network after restricting the species to Homo sapiens. Users may choose to display the TF coding genes by checking “Map TFs to coding genes.” Figure 2B shows the resulting network. While six nodes are present in the GRN in Figure 2B, there are essentially only two genes and their products in the network. When a GRN involves more genes, it may be desirable to simplify the GRN, replacing the TF models with their respective coding genes. Figure 2C displays the simplified two node GRN generated by checking “Simple network.” We note that a GRN can be simplified only after the TF models with their respective coding genes have been acquired. Users may input a PWM or unaligned TFBSs to LASAGNA-Search for model building. Although matrix-scan also accepts PWMs, both matrix-scan and MAPPER2 do not accept unaligned TFBSs. All three tools accept promoter sequences from FASTA, while matrix-scan handles sequences in five additional formats. Automatic sequence retrieval for matrix-scan is accomplished by interfacing with two tools, “retrieve sequence” and “retrieve Ensembl sequence,” on the same web site. These two tools are capable of retrieving sequences in a wide range of species and can be used with any TFBS search tools. LASAGNA-Search and MAPPER2 offer integrated promoter retrieval tools supporting seven and three organisms, respectively.

Visualization of predicted binding sites is usually tightly connected with promoter sequence retrieval. This is because creating a custom track in the UCSC Genome Browser requires knowledge of the genome build (release version) and the genome coordinates for the promoter sequence must also be known. For LASAGNA-Search and MAPPER2, hits found on any promoter sequences retrieved by the provided tool can be visualized with ease in the UCSC Genome Browser. Visualizing hits found by matrix-scan in the UCSC Genome Browser is possible only when the genome build and coordinates are specified in the FASTA header of the promoter sequence. Headers of sequences retrieved by the aforementioned two tools, however, do not contain the required information for enabling visualization of hits in the UCSC Genome Browser.

### Comparison of features to existing web tools

LASAGNA-Search allows users to scan promoter sequences for TFBSs without leaving the LASAGNA-Search page. Many features of LASAGNA-Search were designed to be user-friendly. Hence, even without the knowledge of PWM or TFBS databases and promoter sequence retrieval tools, users can search for binding sites in a promoter sequence and visualize the hits in the UCSC Genome Browser immediately. There are several integrative TFBS search web tools available. By comparing it to existing web tools, we can better understand the advantages and disadvantages of LASAGNA-Search and suggest future improvements. Table 2 summarizes the comparison of LASAGNA-Search to matrix-scan and the search engine of MAPPER2 database for identifying TFBSs.

LASAGNA-Search and MAPPER2 have large libraries of TF models, while users need to collect PWMs before using matrix-scan. Users may input a PWM or unaligned TFBSs to LASAGNA-Search for model building. Although matrix-scan also accepts PWMs, both matrix-scan and MAPPER2 do not accept unaligned TFBSs. All three tools accept promoter sequences from FASTA, while matrix-scan handles sequences in five additional formats. Automatic sequence retrieval for matrix-scan is accomplished by interfacing with two tools, “retrieve sequence” and “retrieve Ensembl sequence,” on the same web site. These two tools are capable of retrieving sequences in a wide range of species and can be used with any TFBS search tools. LASAGNA-Search and MAPPER2 offer integrated promoter retrieval tools supporting seven and three organisms, respectively.

### Evaluation of precomputed TF models

Since MAPPER2 is the web tool most similar to LASAGNA-Search, we compare the TF model collections offered by these two tools on a whole-genome basis. The MAPPER2 database stores hits from the 10Kbp upstream region of each transcript for each TF model, so we scanned the same sequences using TF models offered by LASAGNA-Search. We have no access to the profile hidden Markov models (41) used by MAPPER2 and the dynamic scanning interface offered by MAPPER2 was not functioning at the time of writing. Fortunately, MAPPER2 allows users to download the top 1000 hits for each model. We therefore limited our comparison to the top 1000 hits produced by each TF model.

To evaluate model performance, human and mouse ChIP-seq data from the ENCODE project (42) were used as the gold-standard. We compared the results for all validated TFs on a per-TF basis. Supplementary Tables S1 and S2 list the ChIP-seq tracks (experiments) by TF for human and mouse, respectively. We associated each TF with models that can be used to predict its binding sites. Each of the 1000 hits produced by a model was checked against the ChIP-seq peaks for the TF. A hit was marked as a true positive if it was completely covered by a peak in at least one experiment. A ChIP-seq peak is much longer than TFBS. Otherwise, the hit was marked a false positive.

Evaluating a model based on the top 1000 hits is analogous to evaluating a search engine based on the top 1000 documents. Therefore, we used average precision (43) to score models. This performance measure is widely used in web interfaces to four TFBS search tools with access to whole-genome promoter sequences. However, these tools have no access to the PWM database in the suite, nor do they scan promoters of specific genes or offer visualization of hits. Two tools motivated by evolutionary conservation are COTRASIF (26) and ReXSpecies2 (27). COTRASIF collects 138 JASPAR and 398 TRANSFAC PWMs and offers whole-genome Ensembl promoter sequences. However, it does not offer selection of gene-specific promoter sequences nor does it offer visualization. ReXSpecies2, on the other hand, sources PWMs from JASPAR, scans promoters of specific genes, and allows visualization in the UCSC Genome Browser, but it focuses only on human and mouse sequences, and selecting individual PWMs requires use of regular expression-like syntax.
the information retrieval community and is defined as:

\[
\sum_{k=1}^{1000} P(k) \times tp(k) / c
\]

[Eq. 1]

where \( P(k) \) gives the precision based on the fraction of the top \( k \) hits that are true positives. Indicator \( tp(k) \) is 1 if hit \( k \) is a true positive. Otherwise, \( tp(k) \) is 0. The denominator \( c \) is the portion of bases in upstream regions that are covered by peaks and was computed based on all ChIP-seq experiments used to validate the model. We also scored each model by accuracy, which is equivalent to \( P(1000) \).

The performance of LASAGNA-Search and MAPPER2 for a TF was measured by the average score of the associated models. Results for LASAGNA-Search are listed in Supplementary Tables S3 and S4, while results for MAPPER2 are listed in Supplementary Tables S5 and S6. Average precision and accuracy are given in individual columns. Each row presents the performance of model predictions of the binding sites of a TF. Figure 6 shows the comparison between LASAGNA-Search and MAPPER2 in terms of average precision. A similar comparison in terms of accuracy is shown in Supplementary Figure S4.

An outlier corresponding to Mafk is seen in Figures 6 and S4. Four models in LASAGNA-Search and one MAPPER2 model were used to predict Mafk binding sites (see Tables S4 and S6). Interestingly, the best model of each tool is based on the same TRANSFAC matrix M00037. The LASAGNA-Search model is a PWM model that has no position dependence information. The MAPPER2 model, however, uses a hidden Markov model that considers position dependence. The use of position dependence gave the MAPPER2 model an edge over the LASAGNA-Sarch model. The other three LASAGNA-Search models performed much worse than the one based on the M00037 matrix, resulting in poor average performance on Mafk. While it is difficult to draw conclusions based on only 13 mouse TFs, the results from human TFs indicate that LASAGNA-Search models are significantly better. Overall, we observe that LASAGNA-Search significantly outperforms MAPPER2, indicating that the models used in LASAGNA-Search more accurately predict TFBSs.

We plan to improve LASAGNA-Search by expanding the content and incorporating useful features. Additional organisms will be supported in automatic promoter retrieval and visualization in the UCSC Genome browser. To expand our TF model collections, more sources of TFBSs and PWMs such as the PAZAR database (23) and ChIP-seq data will be considered. The general binding preference (GBP) score (39) is based on multiple evidence sources including evolutionary conservation and has been shown to improve prediction of binding sites. Integrating the GBP scores with the search module will be investigated.

In a recent report, using a cluster of TF models to scan a sequence for binding sites has outperformed the best model in the cluster (32). This strategy will benefit from our large collections of TF models and improve the TFBS search performance of LASAGNA-Search. Finally, we will enable the search for two-block motifs (44,45), which are binding sites composed of two half sites separated by variable-length gaps. While plenty of work has been devoted to de novo two-block motif discovery (44,46,49),

What do published researchers across the country have in common?

- measuring plant metabolism
- petrochemical cooling
- temperature gradient bar
- liquid sample stirring & cooling
- medication storage
- direct contact cooling to -73°C
- thermal nociception testing
- dual temperature zones

They use TECA Cold/Hot Plates in their research studies.

Contact TECA Corporation today to learn more! sales@thermoelectric.com 773.342.4900 or 888.832.2872 www.thermoelectric.com
searching for two-block motif instances is more straight-forward. Using two TF models with or without a gap penalty (44) will be investigated.

Acknowledgments
We are indebted to Prof. Daniel Schwartz and two anonymous reviewers for their comments, which greatly improved LASAGNA-Search and this paper. This work was supported in part by the National Science Foundation grant numbers CCF-0755373 and OCI-1156837.

Competing interests

The authors declare no competing interests.

References


3. Fiore, R., S. Khudayberdiev, M. Christensen, G. Siegel, S.W. Flavell, T.K. Kim, M.E. Greenberg, and G. Schratt. 2009. Mef2-mediated transcription of the miR379-410 cluster regulates activity-dependent dendritogenesis by fine-tuning Pumilio2 protein levels. EMBO J. 28:697-710.


32. Oh, Y.M., J.K. Kim, S. Choi, and J.Y. Yoo. 2012. Identification of co-occurring transcription factor binding sites from DNA

Received: 17 October 2012; accepted: 25 January 2013

Address correspondence to Chun-Hsi Huang, Department of Computer Science and Engineering, 371 Fairfield Way, Unit 4155, University of Connecticut. Storrs, CT, USA. Email: huang@engr.uconn.edu

To purchase reprints of this article, contact: biotechniques@fosterprinting.com

**PXi Power**

- **Imaging made easy whatever your application**
- **Automated imaging for chemiluminescence, fluorescence, IR and fluorescence multiplexing**
- **Super sensitivity and superb images**
- **Powered by GeneSys, the acclaimed acquisition and control software from Syngene**

If you only ever invest in one imaging system for gels and blots for your laboratory, then PXi is the smart choice.

[Image of PXi biotech advertisement]