DNA Motif Databases and Their Uses

Gary D. Stormo[1]

[1]Washington University School of Medicine, St. Louis, Missouri

Transcription factors (TFs) recognize and bind to specific DNA sequences. The specificity of a TF is usually represented as a position weight matrix (PWM). Several databases of DNA motifs exist and are used in biological research to address important biological questions. This overview describes PWMs and some of the most commonly used motif databases, as well as a few of their common applications. © 2015 by John Wiley & Sons, Inc.

Keywords: transcription factors • DNA motifs • position weight matrices • binding site predictions

How to cite this article:
Stormo, G.D. 2015. DNA motif databases and their uses. *Curr. Protoc. Bioinform.* 51:2.15.1-2.15.6.
doi: 10.1002/0471250953.bi0215s51

INTRODUCTION

Gene expression is regulated by transcription factors (TFs), many of which are sequence-specific DNA-binding proteins. Many databases exist that contain increasingly large collections of binding-site motifs. This overview includes a brief description of the most common type of specificity representation, the position weight matrix (PWM), and a few of the methods used to determine PWMs. It also points out a few of the largest and most frequently used motif databases that have collections of tools for using those motifs and provides brief descriptions of some of the common ways in which motifs are used in biological research.

DNA MOTIFS

The specificity of TFs is most commonly represented using position weight matrices (PWMs; reviewed in Stormo, 2013), sometimes called position-specific scoring matrices (PSSMs) or simply weight matrices. The literature is inconsistent about names, and sometimes PWM is used for something else, described below, so one should check on the definition being used in any particular paper. A PWM provides a score for all possible bases at each position in a binding site (Fig. 2.15.1 A). In the figure, the PWM has 5 columns for the 5-long binding site and a row for each of the four bases, A, C, G, and T. The score for any 5-long sequence is the sum of the matrix values corresponding to that sequence. In Figure 2.15.1B, a piece of genomic DNA is moved under the matrix so that scores are computed for each of three consecutive positions in the sequence. In this example, the second alignment has the highest score of 5.2, and in fact that is the highest possible score because GTAGG has the highest score in columns 1 to 5 (tied with GTCGG).

Any particular TF will have its own PWM to describe its specificity, and there are various methods for determining the elements of the PWM. Until recently, the most common method was to take an alignment of known binding sites, or sites inferred using a motif-discovery algorithm on some experimental data, such as ChIP-seq experiments [see *UNIT 2.13* (Ji et al., 2011) and *UNIT 2.14* (Feng et al., 2011)]. From the aligned sites, a count matrix is obtained, which simply records the number of each base at each position (this is sometimes referred to as a position frequency matrix, PFM). If one normalizes the count matrix so that each column sums to 1 (or 100 if using percentages), one gets what we have called a PFM (Fig. 2.15.1C), but which is also referred to as a position probability matrix (PPM); in some papers this is called a PWM. The PFM is a probabilistic model of the specificity of the set of binding sites and can be used directly, but that requires multiplying the elements that correspond to the sequence, not adding them.

Current Protocols in Bioinformatics 2.15.1-2.15.6, September 2015
Published online September 2015 in Wiley Online Library (wileyonlinelibrary.com).
doi: 10.1002/0471250953.bi0215s51

A

Base\position	1	2	3	4	5
A	-0.6	-1.1	0.6	-0.4	0.4
C	-3.0	-0.1	0.6	-1.5	-3.0
G	1.3	-0.1	-1.2	1.4	1.2
T	-0.6	0.7	-1.2	-1.5	-1.8

B

.... A **A G T A G** G C .... =- **1.1**

.... A A **G T A G G** C .... = **5.2**

.... A A G **T A G G C** .... = **- 4.5**

C

Base\position	1	2	3	4	5
A	0.10	0.07	0.44	0.11	0.26
C	0.02	0.20	0.44	0.04	0.01
G	0.78	0.20	0.06	0.81	0.70
T	0.10	0.53	0.06	0.04	0.03

Figure 2.15.1 Position Weight Matrix creation and use. (**A**) A PWM for some transcription factor. The binding site for this motif is 5 bases long. (**B**) A short sequence is scanned against the PWM. Shown are three consecutive 5-base segments (in bold); the score associated with the sequence of each segment (aligned to the PWM) is obtained from the sum of the corresponding matrix elements (see also Figure 2.1.1). (**C**). From an alignment of binding sites, a position frequency matrix (PFM) is created that shows the probability of each base at each position in the aligned sites. The PWM of part A was created from this PFM using Equation 1 and setting $P(b) = 0.25$ for all bases.

To get an additive PWM, one can simply take the logarithm of the PFM elements, but more commonly one divides the PFM elements by an expected frequency based on the genomic composition before taking the logarithm, so that the elements of the PWM are defined by Equation 1:

$$\mathrm{PWM}(b, i) = \log \frac{\mathrm{PFM}(b, i)}{P(b)}$$

Equation 1

where PWM(b,i) are the PWM elements for each base, b, at position, i; PFM(b,i) are the frequencies (or probabilities) of each base at each position in the aligned binding sites; and $P(b)$ is the background probability of base b. This is referred to as the log-odds method of determining a PWM from a collection of sites. The information content (IC) for the TF based on the collection of sites is often defined (Stormo, 2013) as Equation:

$$\mathrm{IC} = \sum_{i=1}^{L} \sum_{b=A}^{T} \mathrm{PFM}(b, i)\, \mathrm{PWM}(b, i)$$

Equation 2

where L is the length of the binding site. This is an important relationship to remember—that IC is the average score of all the known sites—because the prediction of binding sites requires a threshold for the score; sites scoring above the threshold would be considered binding sites, and those scoring below would not. Choosing the IC as the threshold would lead to about half of the known sites scoring below the threshold, so most programs use a cutoff less than the IC. One could use the score of the lowest known site, which would be appropriate

if one were confident that all of the "known" sites are truly functional sites, but there may be some false positives in the collection, especially for inferred sites. In addition, some sites may be functional only in a particular context, for instance through cooperative binding with another TF, so that prediction of independent binding sites would require a higher threshold.

In recent years, several high-throughput experimental methods have been developed to determine the specificity of TFs (Stormo and Zhao, 2010; Bussemaker, 2015). These methods generally collect data related to, but not a direct measurement of, binding affinity for thousands, or even millions, of sequences in parallel. The determination of the PWM then requires an algorithm that is more complex than the simple log-odds method. In an extensive comparison of algorithms for the analysis of protein binding microarray (PBM) data, it was shown that the best algorithms for inferring PWMs utilize a biophysical model of the binding process and perform a non-linear regression to find the parameters that provide the best fit to the quantitative data (Weirauch et al., 2013). An interesting result in that paper is that different algorithms applied to the same data could produce PWMs with very different performance; some fit the experimental data much better than others. This result emphasizes that the quality of the PWM depends on both the type of data used and the ability of the algorithm to estimate accurate parameters. Algorithms for other types of high-throughput methods have not been examined in such detail, so it is not clear if optimal algorithms are being used. Many TFs have been studied with multiple methods, and multiple PWMs exist. Usually they are quite similar, but when they differ it is recommended to compare their predictions on test sets to judge their quality.

A limitation of the PWM model for representing specificity is that it assumes the positions in the binding site contribute independently to the binding affinity. This is often a good approximation, but not for every TF (Zhao and Stormo, 2011). There are also examples where the independence assumption is violated and where a TF may bind in different modes to different sites such that a single PWM cannot capture the specificity accurately (Weirauch et al., 2013). Alternative representations have been developed that can allow both variable-length binding sites and combinatorial contributions to affinity (Mathelier and Wasserman, 2013), but so far they are not widely used. One way in which position may contribute non-independently is through variation in structure of the DNA, which depends on nearest-neighbor (or longer) base-pair sequences (Abe et al., 2015). Methods that use structure parameters in combination with PWMs to improve prediction accuracy are being developed (Zhou et al., 2015; also see *http://rohslab.cmb.usc.edu/TFBSshape/*), but are not widely used at this time. Another alternative method of representing specificity is with lists of enriched *k*-mers (DNA sequences of length *k*), where *k* is typically 8. Enriched *k*-mers are generated from PBM data and are available from the UniPROBE database.

DATABASES OF DNA MOTIFS

There are many databases of motifs for TFs, and this section is not intended to be comprehensive; in fact, new ones are regularly created, so even if it were comprehensive now, it would soon be out of date. The databases included here were chosen to meet several criteria: publicly available; containing hundreds of motifs in the form of PWMs; with Web sites containing many useful tools for using the motifs and instructional tutorials on their use.

JASPAR [*http://jaspar.genereg.net/* (Mathelier et al., 2014)] was created in 2004 as a manually curated, publicly accessible database of motifs for eukaryotic organisms. It has undergone several revisions and enhancements since then, with many new features, and is continually updated with new information.

CisBP [*http://cisbp.ccbr.utoronto.ca/* (Weirauch et al., 2014)] is a new database from the Hughes lab at the University of Toronto that is easily the most comprehensive. Currently it includes a list of >160,000 predicted TFs from >300 species. For each TF, it provides a motif if one is known, either from the work of the Hughes lab or obtained from several other databases. It also includes motifs that can be inferred for a TF because it is highly similar to another TF with a known motif. In the paper, the authors describe the criteria for similarity cutoffs, which depend on the TF family, for making motif inferences, as well as the accuracy of such inferences.

UniPROBE [*http://the_brain.bwh.harvard.edu/uniprobe/* (Hume et al., 2015)] is a database from the Bulyk lab at Harvard with motifs obtained using the protein-binding microarray (PBM) technology, either from their own work or from other groups.

The above mentioned databases are all for eukaryotic TFs and motifs. There are also databases specifically for

prokaryotic motifs such as **PRODORIC** (*http://www.prodoric.de/*) and **RegTransBase** (*http://regtransbase.lbl.gov/*).

In addition, there are a number of databases of motifs for particular species. For example **HOCOMOCO** (*http://autosome.ru/HOCOMOCO/*) is a collection of curated motifs from a variety of sources specifically for human TFs.

For *Drosophila*, large collections of motifs can be found in **FlyFactorSurvey** (*http://pgfe.umassmed.edu/ffs/*) and **OnTheFly** (*https://bhapp.c2b2.columbia.edu/OnTheFly/*). For yeast, there are the databases **ScerTF** (*http://ural.wustl.edu/ScerTF*) and **YeTFaSCo** (*http://yetfasco.ccbr.utoronto.ca/*). Until recently, there were relatively few known motifs for TFs in *C. elegans*, but now there are determined or inferred PWMs for about 40% of them (Narasimhan et al., 2015), available in the **CisBP** database. In fact, since CisBP lists TFs and motifs for hundreds of individual species, it is probably the best source for motifs for any specific organism.

The **MEME** suite of tools does not include its own database of motifs, but includes motifs from many different sources (including most of those listed above). It does have tools that can be used to discover motifs de novo in sets of sequences [*UNIT 2.4* (Bailey, 2002); but see new Web site and updated tutorials at *http://meme-suite.org/*], and it also includes tools for comparing motifs to each other and for searching sequences for matches to motifs. MEME is included it in this list because of its links to most of the important motif databases and the many tools that utilize those databases.

USES OF DNA MOTIFS

The following is a collection of typical queries one might make of motif databases. Most the database resources listed in the previous section can accomplish most of these tasks. I pick an example for each query.

Is There a Motif for My TF of Interest?

One could go to any of the databases and ask if a certain TF occurs, and obtain its PWM. But CisBP is probably the best resource for this query because it contains a list of the known and predicted TFs for over 300 eukaryotic species, whether a motif is known or not. If there is a motif, either determined experimentally or inferred by similarity to the motif of another TF, it can be obtained from the database. In fact, all of the motifs, from a variety of sources, are available. If there are multiple PWMs and they are not nearly identical, it is worth looking in JASPAR, where each TF has only one PWM based on a curator's assessment of which is the most accurate.

What TFs Bind to Motifs Similar to This One?

Motif discovery can be applied to sequences without knowing the TF that binds the motif. For example, MEME (*UNIT 2.4*; Bailey, 2002) can find a motif for a set of promoters that are co-regulated, but for which the regulating TF is unknown. The discovered motif can then be compared to databases of motifs to identify ones that are similar. If only one known motif is similar, and the TF for that motif occurs in the species of interest, that result provides a good candidate for the regulating TF. Often there are multiple known motifs with significant similarity to newly discovered one, which may result in a list of candidate TFs. A good tool to compare a motif with a database (or several databases simultaneously) is TOMTOM from the MEME suite of tools (*http://meme-suite.org/*).

Where Are the Likely Binding Sites for My TF?

Given a TF of interest with a known motif, one may ask where it binds in the genome of interest as a way to infer genes that it may regulate. All of the databases allow one to scan a sequence, or set of sequences, for matches to a given motif. For example, the FIMO tool from the MEME suite allows the user to input one or more motifs and one or more sequences to be searched (as in Fig. 2.15.1). A threshold can be set as a *p*-value, which is determined by the motif as the probability of a score occurring by chance. One can also search entire genomes, but in eukaryotic organisms such searches will return a vast number of false-positive predicted sites. This is because most of the genome is inaccessible to TF binding due to chromatin organization. If one first identifies the regions of the genome that are accessible, such as with the use of DNase hypersensitivity, then searching those regions with motifs is much more effective in identifying true binding sites (Neph et al., 2012). The MAST and MCAST programs from the MEME suite allow one to identify which sequences, from a defined set, contain significant matches to a set of motifs (in MCAST, the matches must be clustered).

What TFs Are Likely to Bind to Particular Genome Segments?

Another query of motif databases is to identify all of the motifs with significant matches to a particular sequence, or set of sequences. For example, one may have a promoter, or collection of promoters, that are co-regulated and want to identify candidate TFs that may regulate those genes. The UniPROBE database lets one upload sequences and then searches them for matches to the collection of enriched 8-mers from the database. The CisBP database also has a tool to search a sequence for matches to any motifs in the database. The AME program in the MEME suite allows one to search a set of sequences for significant matches to an entire database (from several different databases).

CONCLUSION

Motifs for thousands of TFs have been determined, and more are being generated at a rapid pace. Several databases have been established that store the motifs, most commonly as PWMs, but some other representations are also available. The databases described in this unit all have tools available online to perform many of the common tasks for which PWMs are used. I have not provided a detailed description of each database, or a comprehensive list of available tools. Both the databases and tool sets are updated regularly, and users are encouraged to read the online manuals and tutorials available at each database to get current information about their contents and capabilities.

Literature Cited

Abe, N., Dror, I., Yang, L., Slattery, M., Zhou, T., Bussemaker, H.J., Rohs, R., and Mann, R.S. 2015. Deconvolving the recognition of DNA shape from sequence. *Cell* 161:307-318. doi: 10.1016/j.cell.2015.02.008.

Bailey, T.L. 2002. Discovering novel sequence motifs with MEME. *Curr. Protoc. Bioinform.* 00:2.4.1-2.4.35.

Bussemaker, H.J. 2015. Recent progress in understanding transcription factor binding specificity. *Brief. Funct. Genomics* 14:1-2. doi: 10.1093/bfgp/elu050.

Feng, J., Liu, T. and Zhang, Y. 2011. Using MACS to identify peaks from ChIP-Seq data. *Curr. Protoc. Bioinform.* 34:2.14.1-2.14.14.

Hume, M.A., Barrera, L.A., Gisselbrecht, S.S., and Bulyk, M.L., 2015. UniPROBE, update 2015: New tools and content for the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res.* 43:D117-D122. doi: 10.1093/nar/gku1045.

Ji, H., Jiang, H., Ma, W. and Wong, W.H. 2011. Using CisGenome to analyze ChIP-chip and ChIP-seq data. *Curr Protoc Bioinform.* 33:2.13.1-2.13.45.

Mathelier, A. and Wasserman, W.W. 2013. The next generation of transcription factor binding site prediction. *PLoS Comput. Biol.* 9:e1003214. doi: 10.1371/journal.pcbi.1003214.

Mathelier, A., Zhao, X., Zhang, A.W., Parcy, F., Worsley-Hunt, R., Arenillas, D.J., Buchman, S., Chen, C.Y., Chou, A., Ienasescu, H., Lim, J., Shyr, C., Tan, G., Zhou, M., Lenhard, B., Sandelin, A., and Wasserman, W.W. 2014. JASPAR 2014: An extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.* 42:D142-D147. doi: 10.1093/nar/gkt997.

Narasimhan, K., Lambert, S.A., Yang, A.W., Riddell, J., Mnaimneh, S., Zheng, H., Albu, M., Najafabadi, H.S., Reece-Hoyes, J.S., Fuxman Bass, J.I., Walhout, A.J., Weirauch, M.T., and Hughes, T.R. 2015. Mapping and analysis of transcription factor sequence specificities. *eLife* 4:e06967. doi: 10.1038/nature11212.

Neph, S., Vierstra, J., Stergachis, A.B., Reynolds, A.P., Haugen, E., Vernot, B., Thurman, R.E., John, S., Sandstrom, R., Johnson, A.K., Maurano, M.T., Humbert, R., Rynes, E., Wang, H., Vong, S., Lee, K., Bates, D., Diegel, M., Roach, V., Dunn, D., Neri, J., Schafer, A., Hansen, R.S., Kutyavin, T., Giste, E., Weaver, M., Canfield, T., Sabo, P., Zhang, M., Balasundaram, G., Byron, R., MacCoss, M.J., Akey, J.M., Bender, M.A., Groudine, M., Kaul, R., and Stamatoyannopoulos, J.A. 2012. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* 489:83-90. doi: 10.1038/nature11212.

Stormo, G.D. 2013. Modeling the specificity of protein-DNA interactions. *Quant. Biol.* 1:115-130. doi: 10.1007/s40484-013-0012-4.

Stormo, G.D. and Zhao, Y. 2010. Determining the specificity of protein-DNA interactions. *Nat. Rev. Genet* 11:751-760.doi: 10.1038/nrm3005.

Weirauch, M.T., Cote, A., Norel, R., Annala, M., Zhao, Y., Riley, T.R., Saez-Rodriguez, J., Cokelaer, T., Vedenko, A., Talukder, S.; DREAM5 Consortium, Bussemaker, H.J., Morris, Q.D., Bulyk, M.L., Stolovitzky, G., and Hughes, T.R. 2013. Evaluation of methods for modeling transcription factor sequence specificity. *Nat. Biotechnol.* 31:126-134. doi: 10.1038/nbt.2486.

Weirauch, M.T., Yang, A., Albu, M., Cote, A.G., Montenegro-Montero, A., Drewe, P., Najafabadi, H.S., Lambert, S.A., Mann, I., Cook, K., Zheng, H., Goity, A., van Bakel, H., Lozano, J.C., Galli, M., Lewsey, M.G., Huang, E., Mukherjee, T., Chen, X., Reece-Hoyes, J.S., Govindarajan, S., Shaulsky, G., Walhout, A.J., Bouget, F.Y., Ratsch, G., Larrondo, L.F., Ecker, J.R., and Hughes, T.R. 2014. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* 158:1431-1443. doi: 10.1016/j.cell.2014.08.009.

Zhao, Y. and Stormo, G.D. 2011. Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nat. Biotechnol.* 29:480-483. doi: 10.1038/nbt.1893.

Zhou, T., Shen, N., Yang, L., Abe, N., Horton, J., Mann, R.S., Bussemaker, H.J., Gordân, R., and Rohs, R. 2015. Quantitative modeling of transcription factor binding specificities using DNA shape. *Proc. Nat. Acad. Sci. U S A* 112:4654-4659. doi: 10.1073/pnas.1422023112.