Baryshnikova, A., Costanzo, M., Kim, Y., Ding, H., Koh, J., Toufighi, K., Youn, J.Y., Ou, J., San Luis, B.J., Bandyopadhyay, S., et al. (2010). Nat. Methods *7*, 1017–1024.

Carpenter, A.E., Jones, T.R., Lamprecht, M.R., Clarke, C., Kang, I.H., Friman, O., Guertin, D.A., Chang, J.H., Lindquist, R.A., Moffat, J., et al. (2006). Genome Biol. *7*, R100.

Chong, Y.T., Koh, J.L., Friesen, H., Duffy, S.K., Cox, M.J., Moses, A., Moffat, J., Boone, C., and Andrews, B.J. (2015). Cell *161*, 1413–1424.

González, J.E., Lee, M., Barquinero, J.F., Valente, M., Roch-Lefèvre, S., and García, O. (2012). Anal. Quant. Cytol. Histol. *34*, 66–71.

Herbert, A.D., Carr, A.M., and Hoffmann, E. (2014). PLoS ONE *9*, e114749.

Huh, W.K., Falvo, J.V., Gerke, L.C., Carroll, A.S., Howson, R.W., Weissman, J.S., and O'Shea, E.K. (2003). Nature *425*, 686–691.

Joglekar, A.P., Bouck, D.C., Molk, J.N., Bloom, K.S., and Salmon, E.D. (2006). Nat. Cell Biol. *8*, 581–585.

Lisby, M., Rothstein, R., and Mortensen, U.H. (2001). Proc. Natl. Acad. Sci. USA *98*, 8276–8282.

Styles, E.B., Founk, K.J., Zamparo, L.A., Sing, T.L., Altintas, D., Ribeyre, C., Ribaud, V., Rougemont, J., Mayhew, D., Costanzo, M., et al. (2016). Cell Syst. *3*, this issue, 264–277.

Tkach, J.M., Yimit, A., Lee, A.Y., Riffle, M., Costanzo, M., Jaschob, D., Hendry, J.A., Ou, J., Moffat, J., Boone, C., et al. (2012). Nat. Cell Biol. *14*, 966–976.

Torres, N.P., Ho, B., and Brown, G.W. (2016). Crit. Rev. Biochem. Mol. Biol. *51*, 110–119.

DNA Structure Helps Predict Protein Binding

Gary D. Stormo[1,*] and Basab Roy[1]
[1]Department of Genetics and Center for Genome Science and Systems Biology, Washington University School of Medicine, St. Louis, MO 63108, USA
*Correspondence: stormo@wustl.edu
http://dx.doi.org/10.1016/j.cels.2016.09.004

Incorporating information about DNA structure can increase the reliability of predictions of transcription factor binding sites.

The standard methods for predicting the binding sites of transcription factors (TFs) involve the use of matrix-based scoring systems (Stormo, 2013). While those approaches often work reasonably well, they have two types of limitations. First, they assume the binding sites are all of fixed length and the positions in the binding site contribute independently to the binding affinity. Second, they are agnostic about mechanism; for example, a particular base may be preferred because it interacts directly with the protein, usually through hydrogen bonds or van der Waals contacts, or it may contribute to structural variations in DNA that the protein prefers. In this issue of *Cell Systems*, Mathelier et al. (2016b) address both limitations simultaneously by demonstrating that an approach incorporating information about DNA structure has superior performance to standard matrix-based methods in identifying TF binding sites (Mathelier et al., 2016). Notably, this is the largest and most comprehensive such demonstration making use of an extensive collection of ChIP-seq datasets for many different TFs. Such an improvement in identifying the locations of in vivo binding sites promises to aid in the elucidation of TF-DNA regulatory interactions and in modeling the consequences of sequence variations that effect TF binding and gene expression.

Matrix-based methods of modeling TF binding sites assign a score (which, depending on the method, may be a probability, a log-probability, a log-odds ratio, an energy or other possible interpretations) for the entire site that is the sum (or product if a probability) of the scores for each position in the site (Stormo, 2013). Regardless of the particular method, the models all assume that the positions contribute independently to the total score. Even as much more data have been collected for many more TFs, that approximation has held up surprisingly well (Weirauch et al., 2013). However, there are some TF families where more complex models, allowing for both non-independence between positions and variable lengths of binding sites, have clear advantages (Mathelier and Wasserman, 2013; Weirauch et al., 2013; Slattery et al., 2014).

One of the ways in which the positions in a binding site can contribute non-independently is through variations in the DNA structure, such as minor groove width and various parameters that quantify the roll and twist of the DNA double helix. Those variables depend on local DNA sequence context (Rohs et al., 2009), and incorporating them into an encoding of the sequence captures that context information. Based on extensive computational analysis a database of structural parameters has been derived for all possible DNA pentamers (Zhou et al., 2013). The combined sequence and structure scores for a DNA motif can be used to search for good predicted binding sites in a genomic sequence (Figure 1).

The current work from Mathelier et al. (2016) is more extensive than previous studies that provided evidence that adding structural information to matrix-based methods improves their fit to data from both in vitro and in vivo experiments (Yang and Ramsey, 2015; Zhou et al., 2015). In contrast to previous studies, the authors use a model that weights the importance of structural parameters across the binding site instead of summary statistics for the whole binding site, and they comprehensively compare their

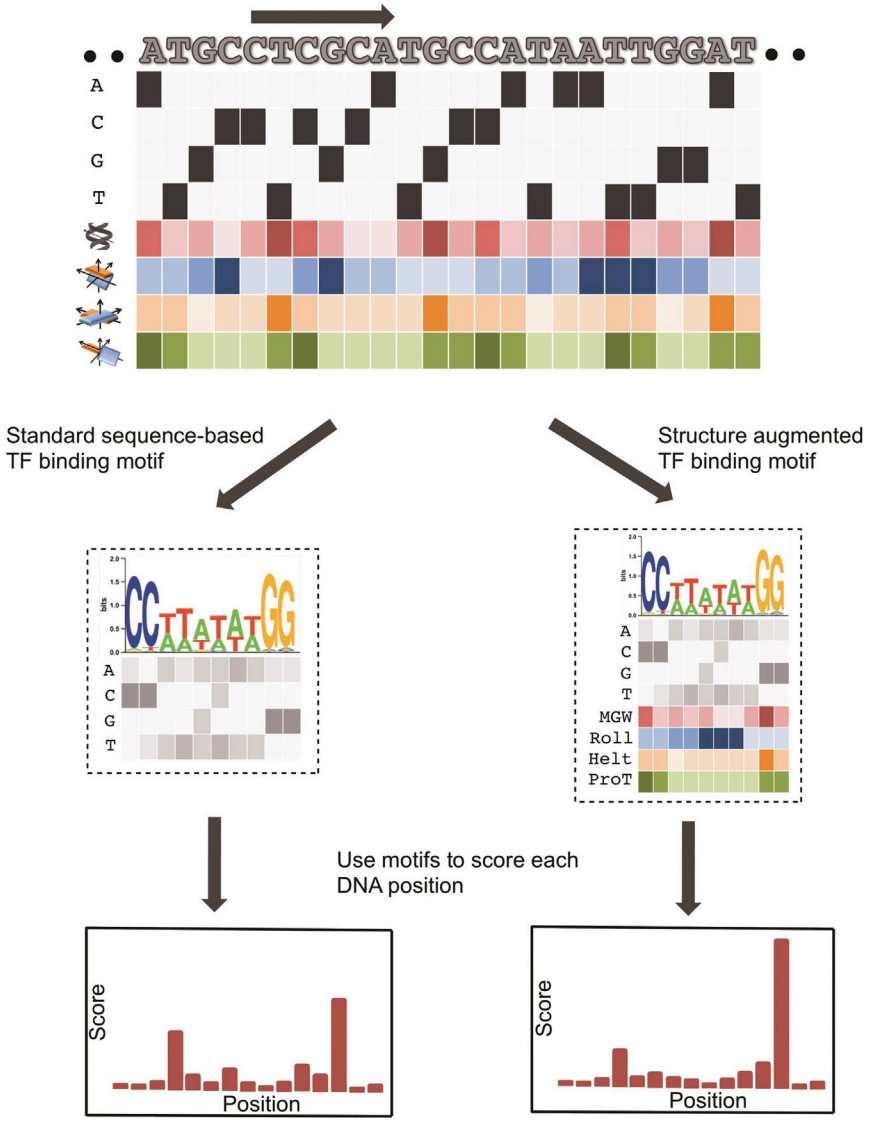

Figure 1. DNA Structure Parameters Improve Motif Modeling

A segment of genomic DNA (upper) is encoded by sequence in the first four rows (A to T from top row to bottom). Black indicates the base that occurs in each position (and is encoded as 1), and white indicates the bases that don't occur (encoded as 0). Below that are four rows of structural parameters—minor groove width (MGW), roll, helical twist (Helt), propeller twist (ProT)—which are real valued numbers (scaled from smallest to largest possible values) shown as different shades of four colors. Mathelier et al. (2016b) include four additional "second order" structural parameters not shown here for simplicity. The sequence-only motif (left) encodes the sequence information in gray scale, indicating the value associated with each possible base at each position (and shown graphically in the logo above). The structure-added model (right) includes parameters in shades of the associated colors, indicating the value associated with that structure parameter. The score assigned to each possible alignment of the motif with the sequence (with the motif scanned along the sequence, black arrow at top) is the sum of the products of the corresponding matrix elements at each alignment position. Note that there is a single highest scoring alignment, which corresponds to the sequence CCATAATTGG that is a good match to the motif. Including structural parameters alters the score at each position and, in this example as evidenced by a single high-scoring position, increases the discrimination for the optimal binding site.

models against several other sequence-only models. The accuracy of each model was assessed by predicting binding sites in ChIP-seq peaks compared to computationally matched random genomic regions from 400 ChIP-seq datasets for 76 TFs containing a diversity of DNA binding domains. To generate the structure-based models, the authors started with sequence-only matrices optimized on the ChIP-seq data and then used support vector machine learning to find weighting parameters for the structure variables that optimized the prediction accuracy.

For most TFs, the increased accuracy is quite modest, but for two TF families in particular, the E2F and MADS-domain families, the improvement is substantial. Because the modeling uses structural parameters explicitly, this also provides a plausible mechanism for differences in TF affinity to different sites and for the enhanced accuracy of binding site prediction. For both the E2F and MADS-domain families, the most important structural parameter was the propeller twist of the base pairs at specific positions of the interactions, consistent with the structures of complexes observed by X-ray crystallography (Mathelier et al., 2016b). Besides being useful for predicting the locations of TF binding sites in genomic DNA, improvements in the motif scoring should also be very useful in predicting the effects of non-coding sequence variants on regulatory interactions that control gene expression.

There remain some open questions and problems. There is currently no standard modeling approach that includes structure and no easy way to search for binding sites with such models. Augmentations to motif databases such as JASPAR (Mathelier et al., 2016a) and Cis-BP (Weirauch et al., 2014) and modified software would enable such enhanced searches (Figure 1), but they don't exist yet.

Another problem arises when trying to infer mechanisms of specificity because the DNA structure depends on the sequence, thus making it difficult to unambiguously separate contributions from sequence and structure. One could determine the structure of the protein-DNA complex at atomic resolution and then observe variations in the DNA structure and direct contacts between the protein and DNA to deduce the mechanism of interaction, but that approach is not amenable to high-throughput analyses. An alternative approach is to build sequence-based motifs that include non-independent contributions between positions in the binding sites, methods which do currently exist (Mathelier and Wasserman, 2013; Stormo 2013). Because the higher-order sequence

contributions can be modeled with independent parameters, eliminating confounding effects, it is possible to identify the combinations of positions that contribute non-independently to binding. That leaves open the problem of inferring mechanism, but those parameters might be used in post-processing steps to infer likely mechanisms that best account for the optimal models, including both direct sequence contacts and structural effects. Advances in inferring structural contributions to binding would also lead to improved recognition models and the design of proteins with desired specificities.

REFERENCES

Mathelier, A., and Wasserman, W.W. (2013). PLoS Comput. Biol. *9*, e1003214.

Mathelier, A., Fornes, O., Arenillas, D.J., Chen, C.Y., Denay, G., Lee, J., Shi, W., Shyr, C., Tan, G., Worsley-Hunt, R., et al. (2016a). Nucleic Acids Res. *44* (D1), D110–D115.

Mathelier, A., Xin, B., Chiu, T.-P., Yang, L., Rohs, R., and Wasserman, W.W. (2016b). Cell Syst. *3*, this issue, 278–286.

Rohs, R., West, S.M., Sosinsky, A., Liu, P., Mann, R.S., and Honig, B. (2009). Nature *461*, 1248–1253.

Slattery, M., Zhou, T., Yang, L., Dantas Machado, A.C., Gordân, R., and Rohs, R. (2014). Trends Biochem. Sci. *39*, 381–399.

Stormo, G.D. (2013). Quant. Biol. *1*, 115–130.

Weirauch, M.T., Cote, A., Norel, R., Annala, M., Zhao, Y., Riley, T.R., Saez-Rodriguez, J., Cokelaer, T., Vedenko, A., Talukder, S., et al.; DREAM5 Consortium (2013). Nat. Biotechnol. *31*, 126–134.

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., et al. (2014). Cell *158*, 1431–1443.

Yang, J., and Ramsey, S.A. (2015). Bioinformatics *31*, 3445–3450.

Zhou, T., Yang, L., Lu, Y., Dror, I., Dantas Machado, A.C., Ghane, T., Di Felice, R., and Rohs, R. (2013). Nucleic Acids Res. *41*, W56–W62.

Zhou, T., Shen, N., Yang, L., Abe, N., Horton, J., Mann, R.S., Bussemaker, H.J., Gordân, R., and Rohs, R. (2015). Proc. Natl. Acad. Sci. USA *112*, 4654–4659.

Deconvolution of the Human Endothelial Transcriptome

Christoph D. Rau,[1] Chen Gao,[1] and Yibin Wang[1,*]
[1]Division of Molecular Medicine, Departments of Anesthesiology, Medicine, and Physiology, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA
*Correspondence: yibinwang@mednet.ucla.edu
http://dx.doi.org/10.1016/j.cels.2016.09.006

A systems approach deconvolutes genes specific to and enriched in endothelium from whole-organ transcriptome data, with applications to other cell types and tissues.

Each human organ consists of many different types of cells working together to carry out complex physiological functions, but understanding cell-type-specific gene expression (Otsuki et al., 2014) is challenging because most published human transcriptome datasets are composite gene-expression profiles generated from mixed cell populations. In this issue of *Cell Systems*, Butler et al. (2016) report a relatively simple but potentially effective way to identify endothelial-cell-enriched genes based on the straightforward correlation of three endothelial-cell-specific reference genes with transcriptome datasets generated from unfractionated human tissues (Butler et al., 2016). They demonstrate that such an approach detects many known and novel endothelial-cell-enriched mRNA species among different human tissue samples. If this concept is extendable to other cell types, this informatics trick could facilitate a rapid, cost-effective deconvolution of whole-tissue gene-expression profiles in order to reveal cell-type-specific features in the otherwise convoluted human transcriptomes.

Current state-of-the-art approaches to identify and analyze cell-type-specific gene expression within a tissue or organ are mechanical in nature, requiring either isolation of a target cell population via cell fractionation or laser capture microdissection prior to RNA analyses (Datta et al., 2015), or employing single-cell RNA sequencing (RNA-seq) on a dissociated cell mixture (Kolodziejczyk et al., 2015). Both methods are resource intensive and technically challenging (Stegle et al., 2015), thus limiting their widespread applications. Furthermore, these methods are not useful for the analysis of datasets acquired from prior studies.

The study reported by Butler et al. (2016) builds on the Human Protein Atlas Project (HPA: http://www.proteinatlas.org) (Uhlén et al., 2015), in which 124 human samples from 32 organs were analyzed by histology and RNA-seq. The authors selected three genes (*CLEC14A*, *vWF*, and *CD34*) as highly reliable endothelial cell reference markers because their mRNA levels were highly correlated with the degree of vascularity across different tissue beds, an indicator of the number of endothelial cells in a tissue, and because their expressions were highly correlated to one another.

To demonstrate that these three genes really are endothelial cell specific and can provide adequate sensitivity to detect additional endothelial cell mRNAs based on their combined correlation coefficients, Butler et al. (2016) performed several tests using previously established