A pairwise interaction model provides an accurate description of *in vivo* transcription factor binding sites

Marc Santolini *, **Thierry Mora** * , **Vincent Hakim** *

*Laboratoire de Physique Statistique, CNRS, Université P. et M. Curie, Université D. Diderot, École Normale Supérieure, Paris, France.

Submitted to Proceedings of the National Academy of Sciences of the United States of America

The identification of transcription factor binding sites (TFBSs) on genomic DNA is of crucial importance for understanding and predicting regulatory elements in gene networks. TFBS motifs are commonly described by Position Weight Matrices (PWMs), in which each DNA base pair independently contributes to the transcription factor (TF) binding. Here, we use available fly and mouse ChIPseq data, to precisely test this description for TF binding *in vivo*. We find that the PWM model generally does not reproduce the observed statistics of TFBS. We show that an accurate TFBS description can be obtained by systematically taking into account pairwise correlations between nucleotide at different postions in the TFBS, via the principle of maximum entropy. The resulting pairwise interaction model is formally equivalent to the disordered Potts model of statistical mechanics. It generalizes previous approaches to interdependent positions and is found to outperform models consisting of mixtures of PWMs. The model can account for co-variation of two or more base pairs, as well as secondary motifs. It allows for a general analysis of nucleotide pairwise interactions. The significant pairwise interactions are found to be sparse and dominantly located between consecutive base pairs in the flanking region of TFBS.

Position Weight Matrix | transcription factor | maximum entropy | ChIP-seq | Potts model

Abbreviations: PWM, Position Weight Matrix; TF, Transcription Factor; TFBS, Transcription Factor Binding Site

Gene regulatory networks are at the basis of our understanding of a cell state and of the dynamics of its response to environmental cues. Central effectors of this regulation are Transcription Factors (TF) that bind on short DNA regulatory sequences and interact with the transcription apparatus or with histone-modifying proteins to alter target gene expressions (1). The determination of Transcription Factor Binding Sites (TFBS) on a genome-wide scale is thus of importance and is the focus of many current experiments (2). An important feature of TF in eukaryotes is that their binding specificity is moderate and that a given TF is found to bind a variety of different sequences *in vivo* (3). The collection of binding sequences for a TF-DNA is widely described by a Position Weight Matrix (PWM) which simply gives the probability that a particular base pair stands at a given position in the TFBS. The PWM provides a full statistical description of the TFBS collection when there are no correlations between nucleotides at different positions. Provided that the TF concentration is far from saturation, the PWM description applies exactly at thermodynamic equilibrium in the simple case where the different nucleotides in the TFBS contribute independently to the TF-DNA interaction, such that the total binding energy is the mere sum of the individual contributions (4, 5).

Previous works have reported several cases of correlations between nucleotides at different positions in TFBSs (6, 7, 8, 9). A systematic *in vitro* study of 104 TFs using DNA microarrays revealed a rich picture of binding patterns (10), including the existence of multiple motifs, strong nucleotide position interdependence, and variable spacer motifs, where two small determining regions of the binding site are separated by a variable number of base pairs. Recently, the specificity of several hundred human and mouse DNA-binding domains was investigated using high-throughput SELEX. Correlations between nucleotides were found to be widespread among TFBSs and predominantly located between adjacent flanking bases in the TFBS (9). The relevance of nucleotide correlations remains however debated (11).

On the modeling side, probabilistic models have been proposed to describe these correlations, either by explicitly identifying mutually exclusive groups of co-varying nucleotide positions (7, 12, 13), or by assuming a specific and tractable probabilistic structure such as Bayesian networks or Markov chains (14, 15, 9). However, the extent of nucleotide correlations in TFBSs *in vivo* remains to be assessed, and a systematic and general framework that accounts for the rich landscape of observed TF binding behaviours is yet to be applied in this context. The recent breakthrough in the experimental acquisition of precise, genome-wide TF-bound DNA regions with the ChIPseq technology offers the opportunity to address these two important issues. Using a variety of ChIPseq experiments coming both from fly and mouse, we first show that the independent model generally does not reproduce well the observed TFBS statistics for a majority of TF. This calls for a refinement of the PWM description that accounts for interdependence between nucleotide positions.

The general problem of devising interaction parameters from observed state frequencies has been recently studied in different contexts where large amounts of data have become available. These include describing the probability of coinciding spikes (16, 17) or activation sequences (18, 19) in neural data, the statistics of protein sequences (20, 21), and even the flight directions of birds in large flocks (22). Maximum-entropy models accounting for pairwise correlations in the least constrained way have been found to provide significant improvement over independent models. The PWM description of TF binding is equivalent to the maximum entropy solely constrained by nucleotide frequencies at each position. Thus, we propose, in the present paper, to refine this model by further constraining pairwise correlations between nucleotide positions. This corresponds to including effective pairwise interactions between nucleotides in an equilibrium thermodynamic model of TF-DNA interaction, as already proposed (23). When enough data are available, the TFBS statistics and predictability are found to be significantly improved in this refined model. We consider, for comparison, a model that describes the statistics of TFBSs as a statistical mixture of PWMs (14) and generalizes previous proposals (24, 25). This alternative

Reserved for Publication Footnotes

model can directly capture some higher-order correlations between nucleotides but is found to be outperformed for all considered TF by the pairwise interaction model.

We further show that the pairwise interaction model accounts for the different PWMs appearing in the mixture model by studying its energy landscape: each basin of attraction of a metastable energy minimum in the pairwise interaction model is generally dominantly described by one PWM in the mixture model. Significant pairwise interactions between nucleotides are sparse and found dominantly between consecutive nucleotides, in general qualitative agreement with *in vitro* binding results (9). The proposed model with pairwise interactions only requires a modest computational effort. When enough data are available, it should thus generally prove worth using the refined description of TFBS that it affords.

The PWM model does not reproduce the TFBS statistics

We first tested how well the usual PWM model reproduced the observed TFBS statistics, *i.e.* how well the frequencies of different TFBSs were retrieved by using only single nucleotide frequencies. For this purpose, we used a collection of ChIPseq data available from the literature (26, 27, 28), both from *D. Melanogaster* and from mouse embryonic stem cells (ESC) and a myogenic cell line (C2C12). The TFBSs are short L-mers (we take here $L = 12$), which are determined in each few hundred nucleotides long ChIP-bound region with the help of a model of TF binding. One important consequence and specific features of these data, is that the TFBS collection is not independent of the model used to describe it. Thus, in order to self-consistently determine the collection of binding sites for a given TF from a collection of ChIPseq sequences, we iteratively refined the PWM together with the collection of TFBSs in the ChIPseq data (see Figure S1 and the Supporting text for a detailed description). This process ensured that the frequency of different nucleotides at a given position in the considered ensemble of binding sites was exactly accounted by the PWM. We then enquired whether the probability of the different binding sequences in the collection agreed with that predicted by the PWM, as would be the case if the probabilities of observing nucleotides at different positions were independent. Figure S2 displays the results for three different TFs, one from each of the three considered categories: Twi (Drosophila), Esrrb (mammals, ESC), and MyoD (mammals, C2C12). For each factor, the ten most frequent sequences in the TFBS collection are shown. For comparison, Figure S2 also displays the probabilities for these sequences as predicted by the PWM built from the TFBS collection. The independent PWM model strongly underestimates the probabilities of the most frequent sequences. Moreover, the PWM model does not correctly predict the frequency order of the sequences and attributes comparable probabilities to these different sequences, in contrast to their observed frequencies.

The relative entropy or Kullback-Leibler divergence (DKL) is a general way to measure the difference between two probability distributions (29). In order to better quantify the differences between the observed binding sequence frequencies and the PWM frequencies, we computed the DKL between these distributions for all the considered TF, as shown in Figure S2D. For each transcription factor T, part of the differences comes from the finite number N(T) of its observed binding sites. The results are thus compared for each factor T to DKLs between the PWM probabilities and frequencies obtained for artificial sequence samples of size N(T) generated with the same PWM probabilities. For most TFs (22 out of 28), the difference between the observed binding sequence frequencies and the PWM frequencies is significantly larger than expected from finite size sampling. In the following we focus on these 22 factors for which the PWM description of the TFBSs needs to be be refined. It can be noted that the 6 factors for which the PWM description appears satisfactory are predominantly those for which the smallest number of ChIP sequences is available (see Table S1 and Figure S4).

Pairwise interactions in the binding energy improve the TFBS description

The discrepancy between the observed statistics of TFBSs and the statistics predicted by the PWM model calls for a re-evaluation of the PWM main hypothesis, namely the independence of bound nucleotides. As recalled above, the inverse problem of devising interaction parameters from observed frequencies of "words" has been recently studied in different contexts. It has been proposed to include systematically pairwise correlations between the "letters" comprising the words to refine the independent letter description. In the case of a two-letter alphabet, the obtained model is equivalent to the classical Ising model of statistical mechanics (30). In the present case, the 4-nucleotide alphabet (A,C,G,T) leads to a model equivalent to the so-called inhomogeneous Potts model (30) (hereafter called pairwise interaction model), a generalization of the Ising model to the case where spins assume q values and their fields and interaction parameters depend on the sites considered. In this analogy, nucleotides are spins with $q = 4$ colors.

In practice, the probability of observing a given word $(s_1...s_L)$ in the dataset is expressed as $P[s_1...s_L] = (1/\mathcal{Z}) \exp(-\mathcal{H}[s_1...s_L])$, where $\mathcal{Z}$ is a normalization constant. $\mathcal{H}$ is formally equivalent to a Hamiltonian in the language of statistical mechanics, and reads:

$$\mathcal{H}[s_1...s_L] = -\sum_{i=1}^{L} h_i(s_i) - \sum_{i=1}^{L}\sum_{j<i} J_{i,j}(s_i, s_j), \qquad s_i \in \{A, C, G, T\} \qquad [1]$$

The "magnetic fields" h_i at each site i, along with the interaction parameters J_{ij} between nucleotides at positions i and j, are computed so as to reproduce the frequency of nucleotide usage at each position in the TFBS as well as the pairwise correlations between nucleotides at different positions (see the Supporting text). In principle, the number of parameters in the model is sufficient to reproduce the observed values of all pairwise correlations between nucleotides. This however would result in over-fitting the finite-size data with an irrealistically large number of parameters. Therefore, to obtain the model parameters we instead maximized the likelihood that the data was generated by the model with a penalty proportional to the numbers of parameters involved, as provided by the Bayesian Information Criterion (BIC) (31). Similarly to the procedure followed for the PWM, the pairwise interaction model and the collection of TFBSs for a given factor were iteratively refined together, as schematized in Figure S1.

Figure 1 shows the improvement in the description of TFBS statistics when using the final pairwise interaction model, for the three factors chosen for illustrative purposes. Where the independent model failed at reproducing the strong amplitude and non-linear decrease in the frequencies of the most over-represented TFBSs, the pairwise interaction model provides a substantial improvement in reproducing the observed statistics. The improvement is most apparent when comparing the frequencies of the ten most observed TFBSs between the model and the ChIPseq data (Figure 1 A, C, E), and is further shown by the statistics of the full collection of TFBSs (Figure 1 B, D, F).

The pairwise model ranks binding sites differently from the PWM

Precise predictions of TFBSs are one important output of ChIPseq data. Moreover, they condition further validation experiments such as gel mobility shift assays or mutageneses. We therefore found it worth assessing the difference in TFBS predictions between pairwise and independent models.

First, we compared the set of ChIP sequences retrieved by the independent and pairwise models model at the cutoff of 50% TPR

(True Positive Rate) used in the learning scheme, as shown in Figure S3A. The non overlapping set of ChIPseq sequences (*i.e.* sequences that were picked by one model but not by the other) was found to range from a few percent for TF like Esrrb, up to about 15 % for Twist. Thus, even when stemming from the same ChIPseq data, the two models can be learnt from significantly distinct set of sites.

Second, using the set of ChIPseq peaks on which the pairwise model was learned, we looked for the best predicted sites on each ChIPseq bound fragment using both the pairwise and PWM models (Figure S3B).

The overlap was found to be about 80% on average. The overlap between the sets comprising the two best TFBSs of each ChIPseq was also computed. This resulted in an overlap increase or decrease between the prediction of the two models depending on the average of number of binding sites per retrieved ChIPseq fragment. In a few cases (*e.g* CTCF, Esrrb), the inclusion of the second best TFBS increased the difference between the two models. This generally happened when the ChIPseq fragments were retrieved with typically a single TFBS above threshold (*e.g.* for Essrb the TFBS specificity was fixed to retrieve 50% of 18453 ChIPseq and about 11000 fragments where found by the two models—see Table S1). In these cases, the low specificity TFBSs tended to differ more between the two models than the very specific ones. In several other cases (*e.g* for Fosl1, Max, n-Myc, USF1), the inclusion of the second best predicted binding sites (Figure S3B) greatly increased the overlap between the two model predictions. This corresponded to cases for which the retrieved fragments contained on average two of more TFBSs about the specificity threshold (Table S1). This showed that for these cases the prediction difference between the two models arose predominantly from a different ranking of the best TFBSs.

In conclusion, the TFBS predictions made by the two models can differ significantly both in the rank of ChIPseq fragments and in the rank of binding sites on these fragments.

Comparison with a PWM-mixture model

When described by a PWM, the binding energies of a TF for different nucleotide sequences form a simple energy well with a single minimum at a preferred consensus sequence. Some authors have instead analyzed the binding specificity of transcription factors by introducing multiple preferred sequences (24, 25). A model of this type that naturally generalizes the PWM description consists of using multiple PWMs (14). We found it interesting to investigate this approach based on a mixture of PWMs and compare it with the pairwise interaction model to get some insights into potentially important high-order correlations that would not be captured by the pairwise model. As precisely described in the Supporting text, an initial mixture of K PWMs was generated by grouping into K clusters the TFBS data for a given TF. Similarly to the pairwise interactions, the number of clusters K was constrained, to avoid over-fitting, by penalizing the corresponding model score using the BIC. For a given TF, the PWM mixture and the collection of TFBSs in the ChIPSeq data were refined iteratively until convergence, usually reached after 10 iterations. The results are shown in Figure 2A for the three representative factors, Twi, Esrrb and MyoD.

The best description of Twi ChIPSeq data is, for instance, provided by a mixture of 5 PWMs, which corresponds to 184 independent parameters. The mixture model yields a significant improvement when compared to the single-PWM model for Twi, and milder ones for Essrb and MyoD. In the three cases however, it proves inferior to the pairwise model.

More generally, Figure 2B shows the performances of the different models for all studied TFs using the Kullback-Leibler Divergence or DKL between the data distribution $P(s)$ and the models distributions $P_m(s)$. On the one hand, the mixture model improves the description of the binding data for 12 out of 27 TFs as compared to the single PWM model. The mixture model gives in particular strong improvements in the cases for which the binding sites have a palindromic structure (eg Twi, MyoD, Myog, Max, USF1). This feature often stems from the fact that the TF binds DNA as a dimer, which could give some concreteness to the mixture model: the recruitment of different partners by bHLH factors like MyoD or Myog could indeed lead to a mixture of TFs binding the same sites. On the other hand, the pairwise model clearly outperforms the other models in all cases studied.

As in the PWM case, the finite size of the datasets leads us to expect fluctuations in the estimation of the DKL. In order to assess the magnitude of these finite-size fluctuations, we computed the average DKL between the best-fitting (pairwise) model and a finite-size artificial sample drawn from its own distribution, as shown in Figure 2B. Values of this DKL that are larger than the one obtained with the real dataset are indicative of overfitting, while the opposite case would suggest that the model is incomplete. In all cases, however, the DKL obtained with this control procedure was within error bars of the value computed with respect to the observed sample, with the exception of NRSF, MyoD, and Myog, as seen in Figure 2B. Thus, the pairwise model is generally the best possible model, insofar as the available dataset allows us to probe.

The metastable states of the pairwise interaction model

In order to more directly relate the pairwise interaction and the mixture models, it is useful to consider the energy landscape of the pairwise interaction model in the space of all possible TFBSs. By contrast with the simple, single-minimum energy well of the PWM model, the pairwise interaction model has multiple metastable energy minima. The energy landscape of the pairwise interaction model can thus be seen as a collection of energy wells, each centered on its metastable energy minimum. The span of the different energy wells in sequence space can be precisely defined as the basins of attraction of the different metastable minima in an energy minimizing procedure (see Supporting text). This allows one to associate each observed TFBS to a particular energy minimum. This defines basins of attraction that are used to build representative PWMs for each metastable minimum together with a weight—the number of sequences in the basin of attraction—for this energy minimum. We compared each metastable minimum to the PWMs of the mixture model, by calculating the DKL between the PWM computed from the sequences in its basin of attraction and the PWMs of the mixture model. This gave an effective distance which allowed us to associate each metastable state to the nearest PWM of the mixture model.

Using this procedure, we computed the set of PWMs and weights corresponding to the 27 considered TF pairwise interaction models. The correspondence between the two models is shown in Figure S6 for all TFs for which the mixture model uses more than a single PWM. In the case of Twi, the PWMs of the pairwise model ("metastable PWMs") can be clearly associated to the $K = 5$ PWMs of the mixture model. For MyoD, three of the 5 "metastable PWMs" can be clearly assigned to PWMs of the mixture model. The other two have a more spread out representation. The case of Esrrb is similar with one "metastable PWM" in clear correspondence with one PWM of the mixture model, and the other one less clearly so. This representation allows one to identify some features captured by the pairwise model. For example, in the case of Twist, most of the correlations are coming from the two nucleotides at the center of the motif, which take mainly 3 values among the 16 possible: CA,TG and TA. In the case of MyoD, the representation makes apparent the interdependencies between the two nucleotides following the core E-Box motif, and the restriction to the three main cases of CT, TC and TT.

Properties of the pairwise interactions

The computation of the interaction parameters allows an analysis of some of their properties. In particular, it is interesting to quantify their strengths and measure the typical distance between interacting nucleotides. We address these two questions in turn.

The concept of Direct Information was previously introduced to predict contacts between residues from large-scale correlation data in protein families (32). We used it here to measure the strength of the pairwise interaction between two nucleotides. Using the previously generated interaction parameters from the pairwise model, we built the Normalized Direct Information (NDI), a quantity which varies from 0 for non-existing interactions, to 1 when interactions are so strong that knowing the nucleotide identity at one position entirely determines the nucleotide identity at the other position (see Supporting text). Heatmaps displaying the results for the representative Twist, Esrrb and MyoD factors are shown in Figure 3 and in Figures S7 for the other factors. An important observation is that the direct information between different nucleotides is rather weak—usually smaller than 10%—but substantially larger than the direct interaction between nucleotides in the surrounding background (1-3%, see Figure S8). It is interesting to note that such weak pairwise interactions give rise to a substantial improvement in the description of TFBS statistics, similarly to what was previously found in a different context (16). The pairwise interactions are furthermore observed in Figure 3A to be concentrated on a small subset of all possible interactions. This can be made quantitative by computing the Participation Ratio of the interaction weights, an indicator of the fraction of pairwise interactions that accounts for the observed Direct Information (see Supporting text). Here, typical values of $10-20\%$ were found (Figure 3A and Table S2), showing that the interactions tend to be concentrated on a few nucleotide pairs.

The interaction weights can also be used to measure the typical distance between interacting nucleotides. To that purpose, we computed the relative weight of the Direct Information as a function of the distance between nucleotides (see Supporting text). Figure 3B shows box plots that summarize the results for the considered Drosophilae and mammalian TFs. Both plots show a clear bias towards nearest-neighbor interactions with a strong peak at $d = 1$, and a rapid decrease for $d \geq 2$. Finally, the dominant pair interactions are on average located in the flanking regions of the BS in clear anti-correlation with the most informative nucleotides which are on average in the central region (Figure 3 C). These observations for TF binding *in vivo* agree with similar ones made from a large recent analysis of TF binding *in vitro* (9). The fact that for pair correlations to be important, nucleotide variation at a given location is required, may be one way to rationalize them.

Alternative representation of interactions by Hopfield patterns

Using a simple binary description of neurons, JJ Hopfield suggested, in a classic piece of work (33), that neural memories could be attractors corresponding to patterns arising from pair interactions between neurons. These interaction patterns can be computed in the present case. They offer an alternative way to analyze the patterns of correlation from the pair-interactions between positions, as already proposed in a mean-field context in (34). Because the matrix of interactions J_{ij} is symmetric, it can be diagonalized in an orthonormal basis of eigenvectors ξ^k, the Hopfield patterns in the present case, with corresponding real eigenvalues λ_k. The Potts energy (Eq. (1)) for a binding sequence $s_1 \cdots s_L$ can be rewritten in terms of the Hopfield patterns as (see Supporting text):

$$\mathcal{H} = -\sum_i h_i(s_i) - \frac{1}{2}\sum_{k=1}^{4L} \lambda_k \left(\sum_{i=1}^{L} \xi_i^k(s_i) \right)^2 . \qquad [2]$$

Although here the presence of the diagonal h term prevents the patterns to be metastable energy states, they can still be useful to analyze the interaction matrix. This spectral decomposition of the interaction matrix is also similar in spirit to a principal component analysis, and even equivalent in the case of Gaussian variable. One can thus wonder how many patterns are needed to well approximate the full matrix of interactions J. To address this question, one can rank the eigenvalues λ_k in order of decreasing moduli and note J_p the restriction of the interaction matrix generated by the first p eigenvalues and their associated patterns. The full interaction matrix naturally corresponds to J_{48}. Approximate interaction matrices obtained by keeping different numbers of dominant patterns are shown in Figure S5 for the three considered representative factors. Pairs of successive patterns appear to provide the main interaction domains in this representation, as is particularly clear in the case of MyoD. One can see in Figure S5 that J_6 already closely approximates the full interaction matrix, a reflection in the present representation, that the important interactions are concentrated on a few links between pairs of nucleotides.

Discussion

The availability of ChIPseq data for many TFs has led us to revisit the question of nucleotide correlations in TFBSs. In order to perform a fully consistent analysis of this type of data, we have developed a workflow in which the TFBS collection and the model describing them are simultaneously obtained and refined together. We have found that when a sufficiently large number of TFBSs is available, the PWM description does not account well for their statistics. The general presence of correlations that follows from this finding, agrees with previous reports for particular transcription factors (6, 8, 24) and with the conclusions of large scale *in vitro* TF binding studies (10, 9).

In order to refine the PWM description, we have analyzed a model with pairwise interactions (23), and a PWM mixture model (14). Data overfitting is a concern for multi-parameter models and has been addressed by putting a penalty on the parameter number using the BIC. While the mixture-model improved in some cases the PWM description, especially for palindromic binding sites, a much more significant and general improvement was found with the pairwise interaction model. The success of the pairwise interaction model agrees with the results of its recent application (however, without the BIC) to high-throughput *in vitro* binding data (23). It moreover shows that, at least in the case we considered, pairwise interactions are sufficient to account for higher-order correlations, and that an explicit description like the one provided by the PWM-mixture model is not necessary. For example, for Essrb, metastable states arising from nearest-neighbor interactions reproduce a triplet of flanking nucleotides with a variable spacer from the core motif (Figure S9).

Our detailed analysis of the obtained interaction models for different TFs shows that the weights of pairwise interactions are generally weak. The most important are only about 10 % of the PWM weights, but significantly above the interaction weights in the surrounding background DNA (of the order 1-3% by the same measure). Nonetheless, collectively these interactions significantly improve the model description of the TF binding data as found in other examples (16).

We have here obtained the pairwise interaction models based on the principle of maximum entropy, constrained to account for the pair-correlations measured in the data. This approach has already been followed in a variety of biological contexts, from populations of spiking neurons (16, 17) to protein sequences (20) to bird flocks (22). An interesting feature of these interaction models is their non-convexity, which allows for the existence of many local maxima in the probability distribution of sequences, or local minima of energy. This was noted for repertoires of antibodies in a single individual (21), where many of these local states were observed and suggested as possible signatures of past infections. In a very different context, local probability maxima in the probability distribution of retinal

spiking patterns was reported and linked to error-correcting properties of the visual system (35). In the present case of TFBSs, these local mimima reflect the multiplicity of binding solutions and resemble the individual PWMs of the mixture model. Pairwise interaction models thus somehow incorporate models of multiple PWMs while outperforming them.

The previously considered case of protein sequences shares many similarities to the statistics of TFBSs, since correlations in protein sequences as in TFBSs reflect both structural and functional contraints. In proteins families, correlations are usually interpreted as resulting from the co-evolution of residues interacting with each other in the protein structure. These effects are hard to distinguish from phylogenic correlations or other observational biases. Nonetheless, the inference of interaction models from data was successfully used to predict physical contacts between amino-acids in the tertiary structure (36), and to aid molecular dynamics simulations in predicting protein structure (37, 38, 39). In the case of TFBSs, comparison between *in vitro*(10, 9) and *in vivo* binding data may help to disentangle the different possible origins of the found correlations and seems worth pursuing. It appears similarly interesting to study how much of the found pair correlations can be explained on the basis of structural data. Finally, the role of nucleotide interaction in TFBS evolution (40) should be considered and could improve the reconstruction of TFBSs from multi-species comparison (41, 42, 43).

Independently of these future prospects, we have found that the TFBSs predicted from ChIP-seq data significantly depended on the model used to extract them. Since the pairwise interaction model and the developed workflow significantly improve TFBS description and require a modest computational effort, they should prove worthy tools in future data analyses.

Materials and Methods

Genome-wide data retrieval. We use both ChIP-on-chip data from *Drosophila Melanogaster* and ChIPseq data from *Mus Musculus*. Data was retrieved from the litterature (26, 27) and from ENCODE data accessible through the UCSC website `http://hgdownload.cse.ucsc.edu/goldenPath/mm9/encodeDCC/wgEncodeCaltechTfbs/`, for a total of 27 TFs. Among them, there are 5 developmental Drosophilae TFs: Bap, Bin, Mef2, Tin and Twi, 11 mammalian stem cells TFs: c-Myc, E2f1, Esrrb, Klf4, Nanog, n-Myc, Oct4, Sox2, Stat3, Tcfcp2l1, Zfx, and 11 factors involved in mammalian myogenesis: Cebpb, E2f4, Fosl1, Max, MyoD, Myog, Nrsf, Smad1, Srf, Tcf3, Usf1. Overall, there are between 678 and 38292 ChIP peaks, with average size 280bp. DNA sequences were masked for repeats using RepeatMasker (44).

ACKNOWLEDGMENTS. We wish to thank PY Bourguignon and I Grosse for stimulating discussions at a preliminary stage of this work.

1. Spitz F, Furlong EE (2012) Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* 13:613–626.
2. Stamatoyannopoulos JA (2012) What does our genome encode? *Genome Res.* 22:1602–1611.
3. Wasserman WW, Sandelin A (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.* 5:276–287.
4. Berg OG, von Hippel PH (1987) Selection of DNA binding sites by regulatory proteins. statistical-mechanical theory and application to operators and promoters. *J Mol Biol* 193:723–50.
5. Stormo GD, Fields DS (1998) Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem Sci* 23:109–13.
6. Man TK, Stormo GD (2001) Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucleic Acids Res* 29:2471–8.
7. Benos PV, Bulyk ML, Stormo GD (2002) Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res* 30:4442–51.
8. Bulyk ML, Johnson PLF, Church GM (2002) Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res* 30:1255–61.
9. Jolma A, et al. (2013) DNA-Binding Specificities of Human Transcription Factors. *Cell* 152:327–339.
10. Badis G, et al. (2009) Diversity and complexity in DNA recognition by transcription factors. *Science* 324:1720–3.
11. Zhao Y, Stormo GD (2011) Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nat. Biotechnol.* 29:480–483.
12. Zhou Q, Liu JS (2004) Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics* 20:909–16.
13. Hu M, Yu J, Taylor JMG, Chinnaiyan AM, Qin ZS (2010) On the detection and refinement of transcription factor binding sites using ChIP-Seq data. *Nucleic Acids Res* 38:2154–67.
14. Barash Y, Elidan G, Friedman N, Kaplan T (2003) *Modeling dependencies in protein-DNA binding sites* (ACM), pp 28–37.
15. Sharon E, Lubliner S, Segal E (2008) A feature-based approach to modeling protein-DNA interactions. *PLoS Comput Biol* 4:e1000154.
16. Schneidman E, Berry MJ, Segev R, Bialek W (2006) Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* 440:1007–12.
17. Shlens J, et al. (2006) The structure of multi-neuron firing patterns in primate retina. *J Neurosci* 26:8254–66.
18. Ikegaya Y, et al. (2004) Synfire chains and cortical songs: temporal modules of cortical activity. *Science* 304:559–564.
19. Roxin A, Hakim V, Brunel N (2008) The statistics of repeating patterns of cortical activity can be reproduced by a model network of stochastic binary neurons. *J. Neurosci.* 28:10734–10745.
20. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T (2009) Identification of direct residue contacts in protein-protein interaction by message passing. *Proc Natl Acad Sci USA* 106:67–72.
21. Mora T, Walczak AM, Bialek W, Callan CG (2010) Maximum entropy models for antibody diversity. *Proc Natl Acad Sci USA* 107:5405–10.
22. Bialek W, et al. (2012) Statistical mechanics for natural flocks of birds. *Proc Natl Acad Sci USA* 109:4786–91.
23. Zhao Y, Ruan S, Pandey M, Stormo GD (2012) Improved models for transcription factor binding site identification using nonindependent interactions. *Genetics* 191:781–790.
24. Cao Y, et al. (2010) Genome-wide MyoD binding in skeletal muscle cells: a potential for broad cellular reprogramming. *Dev Cell* 18:662–74.
25. Heinz S, et al. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Mol Cell* 38:576–89.
26. Zinzen R, Girardot C, Gagneur J, Braun M, Furlong E (2009) Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature* 462:65–70.
27. Chen X, et al. (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* 133:1106–17.
28. Dunham I, et al. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57–74.
29. Cover T, Thomas J (2006) *Elements of information theory* (Wiley-interscience).
30. Baxter R (2008) *Exactly solved models in statistical mechanics* (Dover Publications).
31. Bishop C, et al. (2006) *Pattern recognition and machine learning* (Springer New York).
32. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T (2009) Identification of direct residue contacts in protein-protein interaction by message passing. *Proc Natl Acad Sci USA* 106:67–72.
33. Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. U.S.A.* 79:2554–2558.
34. Cocco S, Monasson R, Sessak V (2011) High-dimensional inference with the generalized Hopfield model: principal component analysis and corrections. *Phys Rev E Stat Nonlin Soft Matter Phys* 83:051123.
35. Tkacik G, Schneidman E, II MJB, Bialek W (2006) Ising models for networks of real neurons. *ArXiv* q-bio/0611072v1 4 pages, 3 figures.
36. Morcos F, et al. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci USA* 108:E1293–301.
37. Sułkowska JI, Morcos F, Weigt M, Hwa T, Onuchic JN (2012) Genomics-aided structure prediction. *Proc Natl Acad Sci USA* 109:10340–5.
38. Hopf TA, et al. (2012) Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* 149:1607–21.
39. Marks DS, et al. (2011) Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE* 6:e28766.
40. Lassig M (2007) From biophysics to evolutionary genetics: statistical aspects of gene regulation. *BMC Bioinformatics* 8 Suppl 6:S7.
41. Moses A, Chiang D, Pollard D, Iyer V, Eisen M (2004) MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome biology* 5:R98.
42. Siddharthan R, Siggia E, van Nimwegen E (2005) PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput Biol* 1:e67.
43. Rouault H, Mazouni K, Couturier L, Hakim V, Schweisguth F (2010) Genome-wide identification of cis-regulatory motifs and modules underlying gene coregulation using statistics and phylogeny. *Proc Natl Acad Sci U S A* 107:14615–20.
44. Bao Z, Eddy SR (2002) Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res* 12:1269–76.

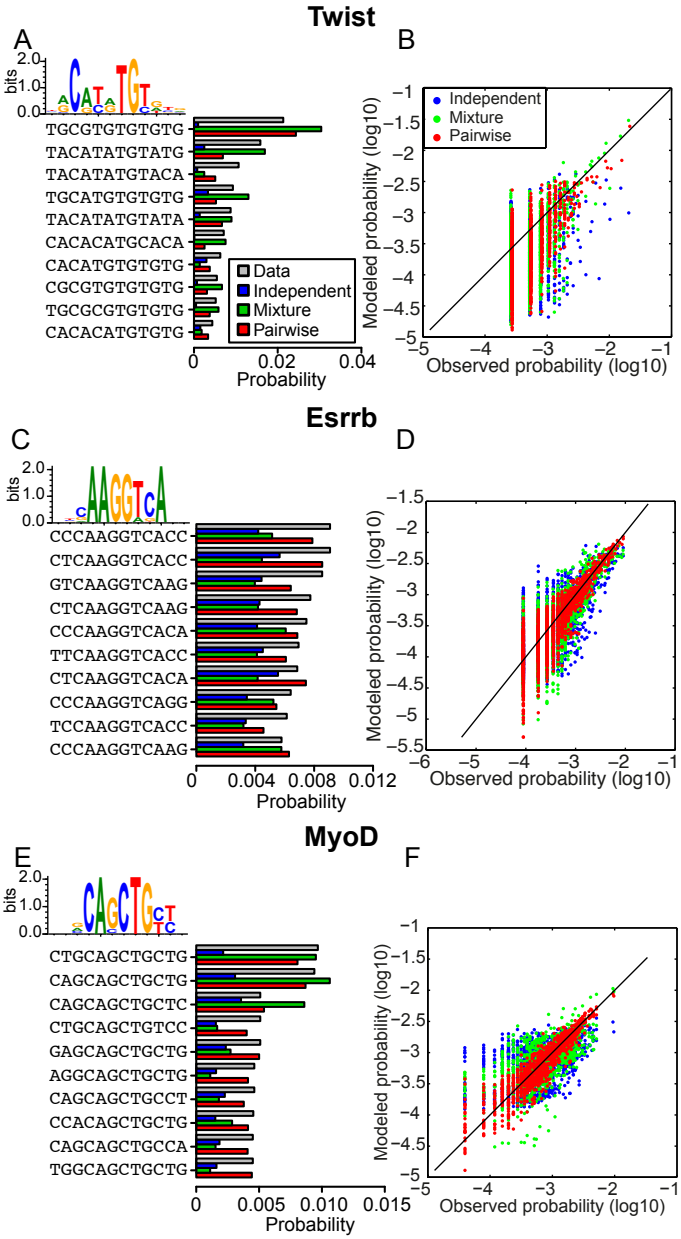

Fig. 1. The observed frequencies (gray bars) of the most represented TFBSs for Twist (A), Esrrb (B) and MyoD (C) TFs, are shown together with the probabilities of these sequences predicted by the independent energy model (blue bars), the pairwise model taking into account interactions between nucleotides (red bars), and the K-means mixture model (green bars). (B,D,F) show the comparison between frequencies for all binding sequences and predicted sequence probabilities for the three models (same color code). The probability predictions of the pairwise model and to a lesser extent of the mixture model are in much better agreement with the observed frequencies than those of the PWM model.

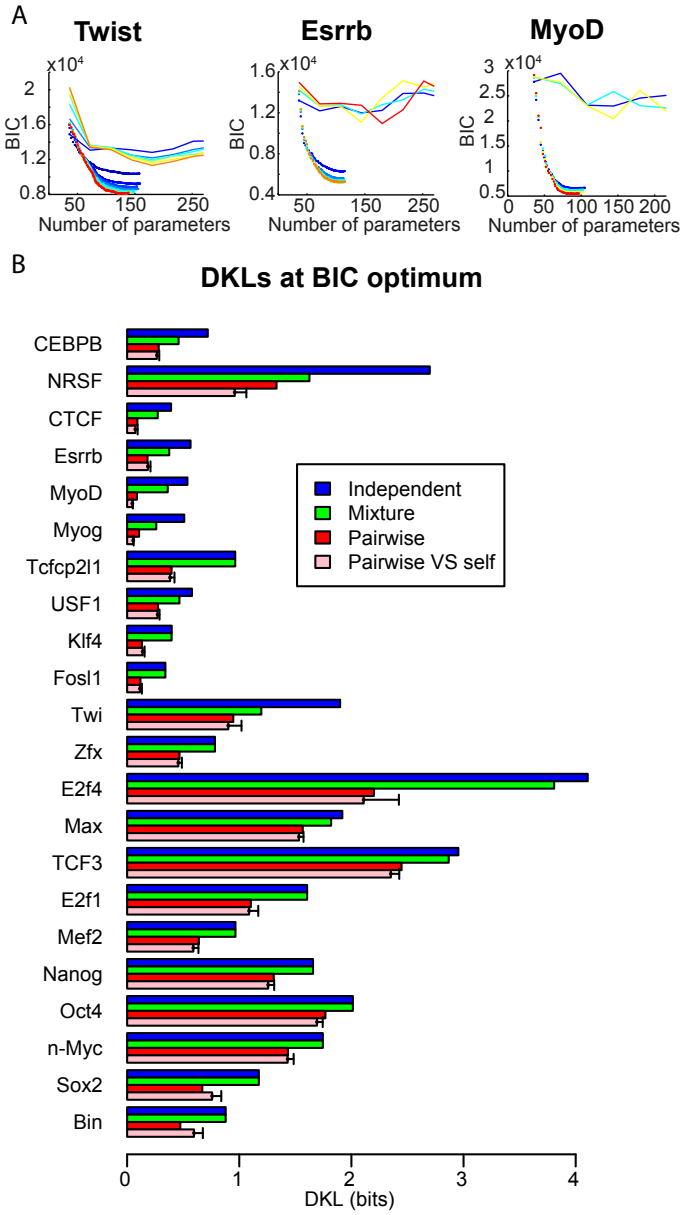

Fig. 2. (A) Minimisation of the Bayesian information criterion (BIC, see Supporting text) is used to select the optimal number of model parameters and avoid over-fitting the training set. The evolution of the BIC is shown for the pairwise model (crosses) and the PWM-mixture model (lines). Colors from dark blue to red indicate the number of interations (see Fig. S1). (B) Kullback-Leibler divergences (DKL) between the independent, K-means and pairwise distributions and the observed distribution for the different TFs, for the BIC optimal parameters. We also show the DKL of the pairwise model with a finite-size sample of sequences drawn from it (pink, see Supporting text). Error bars represent two standard deviations.

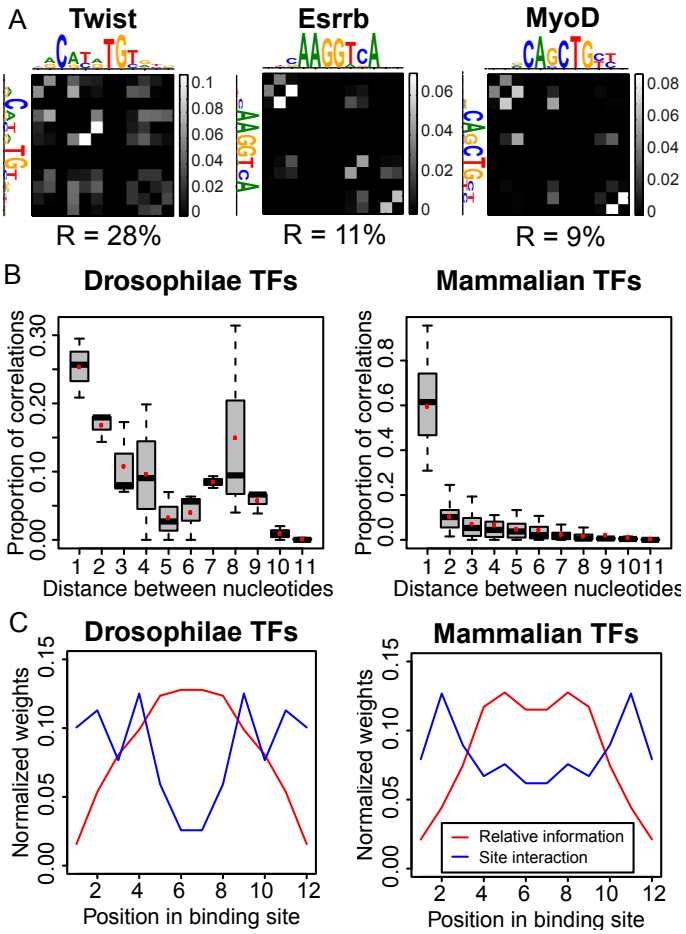

Fig. 3. (A) Heat maps showing the values of the Normalized Direct Information between pairs of nucleotides. The matrix is symmetric by definition. PWMs are shown on the side for better visualization of the interacting nucleotides. The participation ratio R is indicated below each heat map. (B) Distances between interacting nucleotides. The box plots show the relative importance of the Normalized Direct Information as a function of the distance between interacting nucleotides. Red dots denote average values. (C) Sum of normalized direct informations in the TFBSs at a given position, averaged over all considered factors (blue line). The average site information content relative to background as a function of position is also shown (red line). In both quantities, the average over the two TFBS orientations has been taken.