BMC Bioinformatics

RESEARCH ARTICLE Open Access

Reliable scaling of position weight matrices for binding strength comparisons between transcription factors

Xiaoyan Ma[1,2], Daphne Ezer[1,2], Carmen Navarro[2,3] and Boris Adryan[1,2*]

Abstract

Background: Scoring DNA sequences against Position Weight Matrices (PWMs) is a widely adopted method to identify putative transcription factor binding sites. While common bioinformatics tools produce scores that can reflect the binding strength between a specific transcription factor and the DNA, these scores are not directly comparable between different transcription factors. Other methods, including p-value associated approaches (Touzet H, Varré J-S. Efficient and accurate p-value computation for position weight matrices. Algorithms Mol Biol. 2007;2(1510.1186):1748–7188), provide more rigorous ways to identify potential binding sites, but their results are difficult to interpret in terms of binding energy, which is essential for the modeling of transcription factor binding dynamics and enhancer activities.

Results: Here, we provide two different ways to find the scaling parameter λ that allows us to infer binding energy from a PWM score. The first approach uses a PWM and background genomic sequence as input to estimate λ for a specific transcription factor, which we applied to show that λ distributions for different transcription factor families correspond with their DNA binding properties. Our second method can reliably convert λ between different PWMs of the same transcription factor, which allows us to directly compare PWMs that were generated by different approaches.

Conclusion: These two approaches provide computationally efficient ways to scale PWM scores and estimate the strength of transcription factor binding sites in quantitative studies of binding dynamics. Their results are consistent with each other and previous reports in most of cases.

Keywords: Transcription factor, Position weight matrix (Position-Specific Scoring Matrix), Binding site strength

Background

Sequence-specific transcription factors (TFs) are key elements in the regulation of gene expression. Their binding preferences to DNA have been studied extensively *in vitro*, *in vivo* and using computational methods. *In vitro* methods such as protein binding microarray(PBM) [1], high-throughput SELEX measurements [2] and DNase I-seq [3] have provided fundamental insight into the specificity of TF binding. The systematic compilation of DNA sequences from such experiments (and along with them catalogues such as TRANSFAC [4] or JASPAR [5]) have long suggested that TFs do not just bind to one DNA motif, but can bind to a repertoire of similar sequences. Stacks of such sequences give rise to alignment matrices, in which each column represents the absolute count of A, C, G and T nucleotide occurrences per position along the length of the motif. We use "motif" in this manuscript in reference to the PWM motif for a specific TF. Work by Berg et al. [6] introduced a derivative of the alignment or position frequency matrix (PFM), the position weight matrix (PWM, sometimes also noted as PSSM for position-specific scoring matrices), which takes the log likelihood of observing nucleotides taking their overall frequency into account. Berg et al. [7] later showed that the score obtained by comparing the PWM against a DNA sequence is proportional to the binding energy between this TF and the DNA. In most cases the actual binding energy between the protein and DNA is not known, and the proportionality is scaled with a factor commonly termed λ. Berg et al. originally introduced λ to relate the

*Correspondence: ba255@cam.ac.uk
[1] Department of Genetics, University of Cambridge, Downing Street, CB2 3EH Cambridge, UK
[2] Cambridge Systems Biology Center, University of Cambridge, Tennis Court Road, CB2 1QR Cambridge, UK
Full list of author information is available at the end of the article

population of base-pair choices to binding free energy [6]. It is analogy to the inverse temperature factor in statistical physics to describe the energy distribution and also to serve as a factor in tuning the number of potential binding sites in order to satisfy the constraints on overall energy distribution.

There is no well-characterized and easily computable way to determine the TF binding energy for specific DNA sequences and to compare binding site strength between different types of TFs at large scale. This is problematic when scanning the genome with a library of PWMs, as scoring functions treat each PWM independently, and the absolute score associated with a "good match" to the PWM of one transcription factor might be associated with a mismatch for another factor. A more sophisticated application of binding site strength estimation is, for example, modeling the relationship between enhancer occupancy and gene expression [8, 9]. The experimental PBM approach [1] allows the estimation of the relative binding strength of a protein to "naked" DNA *in vitro*, but the data availability is restricted to a limited number of TFs due to high cost of the technology. In addition, PBMs are also not suitable for TFs with longer motifs, as their accuracy will decrease with the length of the DNA probe [1]. Therefore, PWM-based approaches are used to computationally estimate TF binding affinity to a specific sequence [8, 10].

In the majority of bioinformatic studies, the scaling factor λ is unknown and PWM scores are used at face value as measure of affinity. For example, in our own work [11] we used the PWM score without scaling to compare binding site strength across different TFs in *E. coli*, which might lead to a bias due to the absolute differences between the highest and lowest PWM scores across all TFs of interest. One approach is to scale the PWM score by a p-value for each specific score threshold [12]. This method provides a good way to define putative binding sites by choosing a proper statistical threshold, but it is difficult to correlate these p-values with binding energy estimation, as is required for quantitative studies of enhancer activity [8, 9]. Other work has tried to assess the range of λ on the basis of fitting calculated affinity landscapes to ChIP-seq profiles [13, 14]. However, ChIP data is intrinsically noisy and the height of a ChIP peak may not accurately represent the real binding affinity, undermining the stability and accuracy of λ obtained from these methods. In Roider et al. [13], the estimated λ for the same TF in different conditions diverged greatly in nearly one third of TFs. Furthermore, there is a wide band of possible λ values that optimize the correlation. Aforementioned fitting methods are further reliant on chromatin accessibility data acquired under the same growing conditions or development stages, which is sometimes not available for specific TFs.

We propose a simple approximation to estimate the scaling parameter λ based on existing PWMs, average maximum mismatch energy tolerance estimated by high-throughput binding energy measurements [15] and the distribution of PWM scores across the genome of a specific organism. This method is independent of genome-wide binding and accessibility data. Furthermore, in the cases where there are potentially inconsistent PWMs for a particular TF (*e.g.* derived on the basis of individual binding sites vs. derived from high-throughput efforts), we provide a method to convert the known λ for one PWM of the same TF into another suitable value for a new PWM. This method is based on a computational model of the facilitated diffusion of TFs on the DNA that our group established earlier [16]. We calculate sequence-specific residence times of TFs at the DNA, which is correlated with affinity. We can therefore derive λ for different PWMs of the same TF on the basis of the consistency of simulated residence time. These two strategies (a) calculating λ to scale PWM scores based on the mismatch energy theory using a simple equation and (b) converting the scaling parameter λ between different PWMs of the same TF on the basis of simulated residence time of facilitated diffusion provide simple but useful estimations of binding energy across different TFs using properly scaled PWM scores.

Methods

PWMs of TFs for yeast, fly and vertebrates

Position frequency matrices (PFM) used to construct PWMs were downloaded from the JASPAR database (JASPAR-CORE-2014 non-redundant PFM) [5]. Additional sources of PFMs such as those contained in the BioConductor package *PWMEnrich.Dmelanogaster.background* [17] were used as a source of different matrices for the same TFs. PFMs constructed with less than 30 reference sequences of validated binding sites were removed, as we deemed those insufficient descriptions of binding preference. Given that typical TF binding sites span at least six base pairs, we removed any motifs less than 6 base pairs in length.

A bioinformatics approach was used to derive PWM scores [18] as follows:

$$S_j = \sum_{k=1}^{L} \log_2 \frac{v_{j,k}}{f_{j+k}} \tag{1}$$

where j is the DNA position for the PWM score calculation, L is the length of the motif and k represents k^{th} nucleotide in the PWM motif. In addition, if there is a specific nucleotide in position $(j + k)$ on the DNA, f_{j+k} is the frequency of this nucleotide in the whole genome of a specific organism. Nucleotide frequency used for this study in each organism were as follows: *D. melanogaster*: 0.28 for

A and T, 0.22 for G and C; *S. cerevisiae*: 0.31 for A and T, 0.19 for C and G; vertebrate including human and mouse: 0.29 for A and T, 0.21 for C and G. Please note that the choice of background frequencies can be critical, and that adjustments to local extrema may be necessary. We used a pseudo-count μ to adjust the frequency of nucleotides and obtain $v_{j,k}$ to avoid zero frequency as follows [19]

$$v_{j,k} = \frac{n_{j,k} + f_{j+k} \cdot \mu}{\sum_x n_{x,k} + \mu} \tag{2}$$

where μ is chosen to be 1 [19] and we also show that the choice of the pseudo-count μ does not have significant influence on our results (Additional file 1: Figure S6); $n_{x,k}$ is the frequency of certain nucleotide x in a specific position k of the motif.

Simple equation to calculate λ

λ is the scaling factor that allows for direct comparison of different PWMs in terms of binding energy to DNA. Based on the mismatch energy theory for estimating TF binding strength [7], the mismatch energy at a particular binding site j of TF species i in the genome can be expressed as:

$$E_{mismatch,i,j} = \Delta S_{i,j}/\lambda_i = (S_{max,i} - S_{i,j})/\lambda_i \tag{3}$$

where $S_{i,j}$ stands for the PWM score at position j, $S_{max,i}$ is for the maximum PWM score of TF species i and λ_i is the scaling parameter we want to estimate. Note that the mismatch energy we refer to in the text is derived from information theory, with the unit of bits, which can also be described as "mismatch bits". This is useful in a variety of contexts, such as comparing the binding strength of different TFs. In addition, the expected amount of time that the TF is bound to a particular DNA sequence can be estimated as:

$$\tau_j = \tau_0(\lambda) \cdot e^{-S_j/\lambda} \tag{4}$$

where S_j is the PWM score at position j in the genome, τ_0 is the average residence time calculated as in [16]. This equation is widely used in simulations of TF binding kinetics [20].

Given the utility of the λ for estimating binding strength and occupancy time, it is very important to have a simple strategy for estimating it. We derive our equation based on the following core assumptions: 1) The top 0.1 % of the highest scoring matches of the PWM to intergenic regions are considered to be possible TF binding sites, as suggested by [21]. Their genome-wide study of different eukaryotic TFs revealed an average of 1 binding site in every 1-5 thousand base pairs of intergenic sequence. This top 0.1 % threshold has also been similarly adopted in other studies [10]. In addition, if varying this threshold from top 0.01 % to top $1{\cdot}10^{-4}$ and $1{\cdot}10^{-5}$, the rank of calculated λ still shows good correlation in each group of organisms(Additional file 2: Figure S5). 2) The maximum mismatch energy between the consensus binding motif and specific DNA sequences is proportional to the information content of the PWM matrix of the TF. Note that the mismatch energy we refer to in the text is derived from information theory, with the unit of bits, which can also be described as "mismatch bits". The information content (If) of the PWM matrix is defined below [7],

$$If = \sum_{k=1}^{L} \sum_{i \in A,T,C,G} p_{i,k} \log_2 \frac{p_{i,k}}{f_i} \tag{5}$$

where k is the k^{th} nucleotide in the PWM motif, f_i is the background nucleotide frequency, and $p_{i,k}$ is the adjusted frequency of nucleotide i in position k which is defined as follows,

$$p_{i,k} = \frac{nu_{i,k} + f_i \cdot \mu}{\sum_i nu_{i,k} + \mu}$$

where $nu_{i,k}$ is the frequency of certain nucleotide i in a specific position k of the motif, and f_i is the background nucleotide frequency.

The lower boundary of potential binding sites is approximated by the top 0.1 % of PWM scores following the same reason as mentioned before and corresponds to the maximum mismatch energy tolerance level as follows:

$$E_{maxMismatch,i} = \frac{S_{max,i} - S_{top0.1\ \%,i}}{\lambda_i}$$

where $E_{maxMismatch,i}$ stands for maximum mismatch energy tolerance for TF species i, thus, λ_i can be calculated using:

$$\lambda_i = \frac{S_{max,i} - S_{top0.1\ \%,i}}{E_{maxMismatch,i}} \tag{6}$$

Because different transcription factors have different DNA binding domains, the maximum mismatch energy range can vary from one TF to another. Since there is only data available for 4 individual TFs using microfluidic platform-based binding energy measurements [15], we estimated the maximum mismatch energy for other TFs by using the available data as the average rate and assuming that the mismatch energy tolerance is proportional to the information content of the PWM as follows:

$$E_{maxMismatch,i} =< E_{maxMismatch} > \times \frac{If_i}{< If >} \tag{7}$$

where $< E_{maxMismatch} >$ stands for the average maximum mismatch energy tolerance, which is chosen to be 6 bits as is discussed below from the study of Maerkl et al. [15]; If_i represents the information content of a specific PWM and $< If >$ stands for the average information content corresponding to the average maximum mismatch energy [15], which is 13.2 bits. We reason that if the information content is a good indication of how specific a TF is, the energy drop measured in bits between strong and weak binding sites $(S_{max,i} - S_{top0.1\,\%,i})/\lambda_i$ should have some relationship with the binding specificity of a particular TF. The more specific a TF is, the more significant the energy drop can be. Given limited data in binding energy measurement, we assume that the relationship is simply linear.

We chose an average mismatch energy tolerance of 6 bits based on the study by Maerkl et al. 2007 [15]. They showed by mechanical trapping of molecular interactions a significant decline in binding energy by at most 2 to 3 nucleotide mismatches, and each mismatch nucleotide contributes 2 bits in mismatch energy. Even if more mutations are introduced, the binding energy does not drop further since it has already reached the background non-specific binding energy level.

This experiment was applied only to TFs belonging to the bHLH family. In the absence of more comprehensive data, we must assume that all TFs share this value; although if more general TF in-vitro binding energy measurement results become available, we suggest adjusting the specific top score threshold and corresponding average mismatch energy bits accordingly. Another report featuring TFs from different families including: p53, Max, Glucocorticoid Receptor [22] also provides additional support for 6 bits as average mismatch energy tolerance level since TFs from different families in their study have similar binding kinetics.

In order to control for PWM motif length, in the analysis of λ value comparison across different species and TF families, each λ value was transformed into a Z-score. Specifically, PWM motifs were grouped by motif length, with each group having more than 50 PWM motifs (The groups were: 7-8 bp, 9-10 bp, 11-12 bp, 13-15 bp, >= 16 bp), and the λ values were normalized by the mean and standard deviation within each of these groups (Additional file 3: Table S3 lists the mean and standard deviation value for each group, Additional file 4: Figure S3 depicts the distribution of λ at different motif lengths with color coded points that represent different species).

Estimating λ of a new PWM matrix for the same TF based on the residence time landscape of the facilitated diffusion model

Sometimes there may be more than one PWM available for a specific TF. For instance, different TF motif databases (such as JASPAR [5], SwissRegulon [23], FlyFactorSurvey [24], and HOCOMOCO [25]) may have different versions of PWM motifs for the same TF. In order to directly compare the TF binding energy when using two alternative versions of a PWM, it is important to have a way of scaling the results by λ. λ can be adjusted using the formalism introduced in the previous sections. As a compute-efficient alternative, we developed a more optimal strategy for estimating λ, which does not require the assumption that the PWM information content influences the energy mismatch tolerance. Instead, we base our strategy on the estimation of the sequence specific residence time of a particular TF, which is a biologically meaningful quantity and can be correlated with *in vitro* sequence-dependent sliding measurement of TFs [10]. For the same TF, the distribution of the sequence-specific residence time calculated by Eq. 4 should be as consistent as possible, even when using slightly different PWMs if an appropriate λ is chosen for scaling. Based on this, given a known λ for one PWM, we are able to find another suitable λ for the new PWM.

Note that the stronger the PWM score, the more likely it is that the sequence is bound by a TF and that the residence time of a TF is a biologically meaningful quantity, but there is a much greater number of weak and medium strength binding sites than there are strong sites in the genome. Therefore, if we scored each potential binding site equally, the background of weak and medium strength binding sites would have a greater affect on the estimated λ than the strong binding sites. Therefore, we compare residence times across different quantiles on a logarithmic binding strength scale so that the strongest binding sites have the most influence on our λ estimates.

Specifically, in the following analysis, we take the $-log_{10}$ of the cumulative distribution of PWM scores and select all binding sites with values greater than 3.0 (recall that this corresponds to the 0.1 % percent of binding sites, which were chosen as the lower boundary of weak binding sites). We divide these top-scoring binding sites into bins every 0.1 log-quantile and calculate the average residence time for each of these bins. Our strategy identifies the λ that would produce the most similar residence times for each of these log-quantiles. Assuming that for the first PWM we already have an estimate of λ, by either binding profile fitting or other methods, we can use Eq. 4 to calculate the residence time for each binding strength log-quantile, as described above. In the following analysis of this paper, we borrow the values obtained from Eq. 6 as pre-calculated λ for proof-of-principle, since there are very few well-characterized λ values from profile fitting. Note that τ_0 is calculated via the strategy described in Zabet et al. [16] from all intergenetic regions in the genome, which has a different value for each unique PWM.

Now for the second PWM, we can vary λ between the potential values of 0.1 and 3, which was shown to be a possible λ range [13], and calculate the corresponding residence times at each log-quantile level. We can now compare the reference residence times from the first PWM with the residence times for the second PWM across each binding site strength level, and for each value of λ. The λ that minimizes the mean square error between two sets of calculated residence times is chosen as the suitable λ value for the second PWM matrix. Since outliers can have a big influence on the mean square error, we calculated the sum of the absolute differences for the natural logarithm of residence times between the two PWM matrices for these quartile bins (Eq. 8) to make a comparison with the method that uses mean square error.

$$\sum_q |\ln \tau_{q,\lambda} - \ln \tau_{q,ref}| \tag{8}$$

where q represents each quantile in the quantile series, $\tau_{q,\lambda}$ is the residence time in a specific quantile of a particular λ, $\tau_{q,ref}$ is the residence time in the same quantile of the known λ of the reference PWM matrix. The λ derived by minimizing the mean square error or minimizing the value of the above formula show good consistency with adjusted R^2 of 0.9644 (p=$6.3 \cdot 10^{-9}$). Thus, there should not be significant bias using either of these two methods.

The R scripts for both converting λ between two PWM matrices and estimating λ using Eq. 6 are provided in the following link: *https://github.com/XyMa/PWM_scale.*

Results

Estimating scaling parameter λ for binding site affinity across different species and TF families based on Eq. 6

The λ parameter is the critical link between PWM score, the estimated binding energy and TF residence time. Estimating TF binding site affinity by comparing PWM scores at face value can lead to a large bias, especially when this includes comparisons between many types of TFs, because several properties of the PWM itself can influence the PWM score. For example, the information content of the PWMs is positively correlated to the maximum possible PWM score, as is shown in Additional file 5: Figure S1 with an R^2 value of 0.597. Thus, the absolute value of PWM scores cannot be compared directly across different TFs as an indicator of binding site strength. Proper scaling of PWM score is needed in order to compare binding site affinity across different types of TFs. Based on the methods proposed by Berg et al. [7], the TF binding energy for a specific binding site can be computed by Eq. 3 using the estimated λ.

λ calculated by this method are all within the range suggested by Roider et al. [13], which are listed in Table 1 for different organisms. The values for vertebrate species refer to all available vertebrate TFs obtained from the non-redundant PFM JASPAR database. The upper and lower bound of λ across all organisms are quite similar, in the range of 0.25 to 2.83. This indicates that all eukaryotic TFs, no matter which organisms they belong to, all share energetically similar DNA binding mechanisms, since λ can be interpreted as a metric for the chemical property of stickiness between the TF molecule and DNA. To demonstrate the biological applications of this parameter, Fig. 1 shows an example of the *D. melanogaster Even-skipped stripe 1* enhancer with the comparison between PWM score and the affinity estimation using λ scaling. The usefulness of λ estimates becomes apparent when comparing the first two binding sites indicated by blue arrows in this locus; the second binding site has a higher PWM score, but its binding strength is lower than the first binding site once the λ scaling factor is taken into account. Similar situations also appear in the overlapping binding site of Bicoid and Kruppel indicated by the third arrow. Thus, only comparing the raw value of PWM score [11] may lead to false interpretations of binding site importance. Although there is no current experimental evidence for the relative importance of binding sites for this specific enhancer, this example serves to demonstrate how a different interpretation of the contribution of individual binding sites can lead to alternative testable hypotheses.

Table 1 Maximum, minimum and the mean values of λ in 3 groups of organisms

	S. cerevisiae	*D. melanogaster*	Vertebrates
maximum	2.83	2.72	2.82
minimum	0.26	0.35	0.25
mean	1.25	1.40	1.73

Next, we calculated λ for each TF in *S. cerevisiae*, *D. melanogaster* and available vertebrate TFs in JASPAR [5], which are listed in Additional file 6: Table S1. Figure 2a to 2c show the overall λ distribution in each group of organisms. After controlling for motif length, there is a significant difference between vertebrate and *S. cerevisiae* motifs (Welch t-test p-value = 0.008) (Fig. 2d) and between *D. melanogaster* and vertebrate motifs (p-value = 0.043), but no significant difference between *S. cerevisiae* and *D. melanogaster*. Furthermore, we grouped λ values, normalized by PWM motif length, according to different TF families in JASPAR [5] (Fig. 3). The distribution of raw λ values across different TF-families are depicted in Additional file 7: Figure S2. The basic leucine-zipper family and helix-loop-helix family are two families with the highest average z-score of λ, compared with other groups with Welch t-test p-values equal to $8.9 \cdot 10^{-4}$ and $3.7 \cdot 10^{-5}$ respectively. TF families that belong to the same superfamily show similar λ distribution. For example, β-β-α zinc-finger family and the zinc-finger nuclear receptor family both belong to the zinc-finger TF super family, and no significant difference is detected between these two (Welch t-test p value = 0.35), while both are significantly lower than the aforementioned two families (p-value = 0.012 and $5.0 \cdot 10^{-5}$). In addition, homeobox and forkhead TF families, both of which belong to the helix-turn-helix(HTH) TF super family, show no difference in λ z-score distribution (p value = 0.27), but appear to have lower average λ compared with leucine-zipper, helix-loop-helix family and

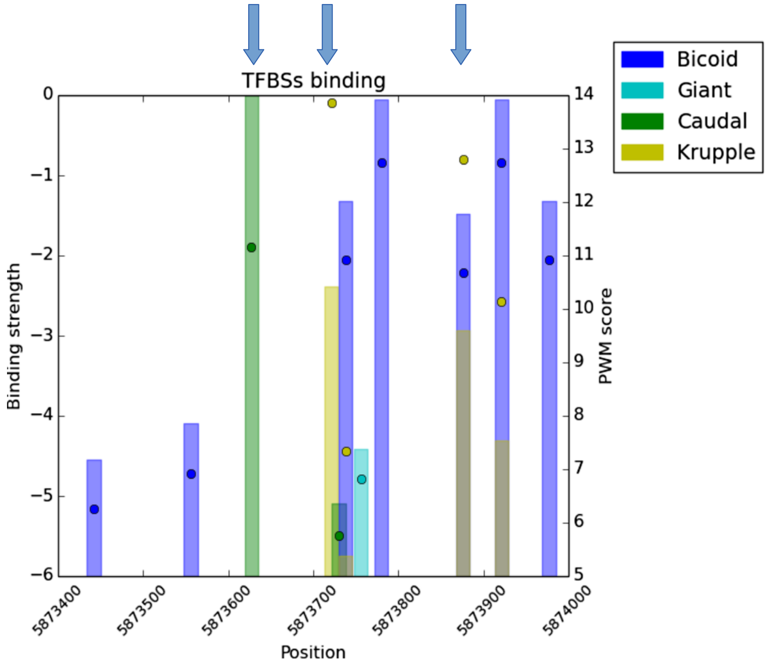


Fig. 1 A comparison between PWM score and binding site strength in the *D. melanogaster even-skipped stripe 1* enhancer. The *even-skipped stripe 1* enhancer on chromosome 2R is dense with binding sites. We compare the raw PWM scores (circles) and the λ-scaled binding strength (height of the bars) for each of these binding sites, colour-coded by the type of TF. Based on raw PWM scores, one might assume that the Caudal site indicated by the first blue arrow would have a lower binding strength than the Kruppel site indicated by the second blue arrow; Eq. 3 instead of Eq. 5, it becomes evident that the opposite is the more likely scenario. The third arrow points to a location where a Kruppel and a Bicoid binding site overlap. Here, the λ adjusted binding strength estimates would suggest that Bicoid binding site is stronger, while a raw PWM score would suggest the opposite. These results illustrate how using raw PWM scores may result in biased interpretation of the relative binding strength of TFs

zinc-finger super family (Welch t-test p-value equals to $5.2 \cdot 10^{-6}$, $1.6 \cdot 10^{-7}$ and $2.2 \cdot 10^{-4}$, respectively).

Since λ is the denominator to the PWM score differences between one binding site and the consensus sequence in Eq. 3, a larger λ indicates lower mismatch energy when ΔS_j is the same. Thus, with the same possible mismatch energy range, if λ is larger, the PWM score can have a greater range from the consensus sequence to the potentially weakest binding site, which indicates that, as suggested by Pabo et al. [26], the binding motif for the TF family has higher flexibility. This is consistent with the fact that the TFs in the zinc-finger super-family, including the

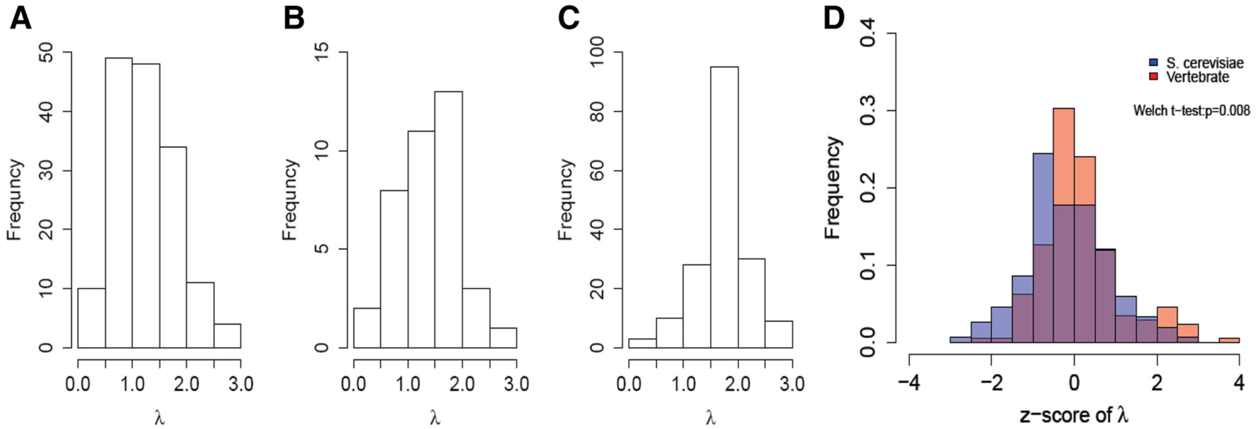


Fig. 2 λ distributions across difference organisms. The histograms depict the λ values estimated from Eq. 6 for the JASPAR non-redundant core motifs in *S. cerevisiae* (**a**), *D. melanogaster* (**b**) and available vertebrates (**c**) [5]. Subfigure D depicts the comparison between z-score distribution of λ for vertebrate and yeast TFs after controlling for motif length

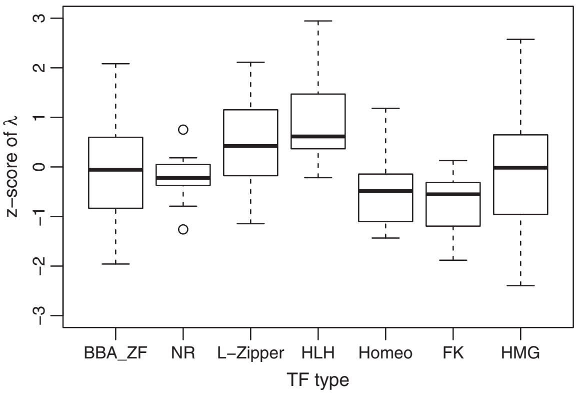


Fig. 3 λ z-score distribution comparison across major TF families. BBA-ZF represents the λ distribution for β-β-α zinc-finger family; NR is zinc-finger nuclear receptor family; L-zipper stands for the basic leucine-zipper family; HLH is helix-loop-helix family; Homeo is homeobox family; FK is fork-head family and HMG is high mobility group family. For each group, λ was calculated by Eq. 6 and z-score is obtained by normalizing λ in each PWM motif length group

nuclear receptor and β-β-α zinc-finger families, are less constrained to a particular motif than HTH super family. Additionally, cross species comparison of λ indicates that from yeast to vertebrate, more flexible TF motifs are used, which is consistent with the result from Itzkovitz et al. [27] that organisms which appeared more recently in evolution tend to use more TFs with motifs of higher flexibility.

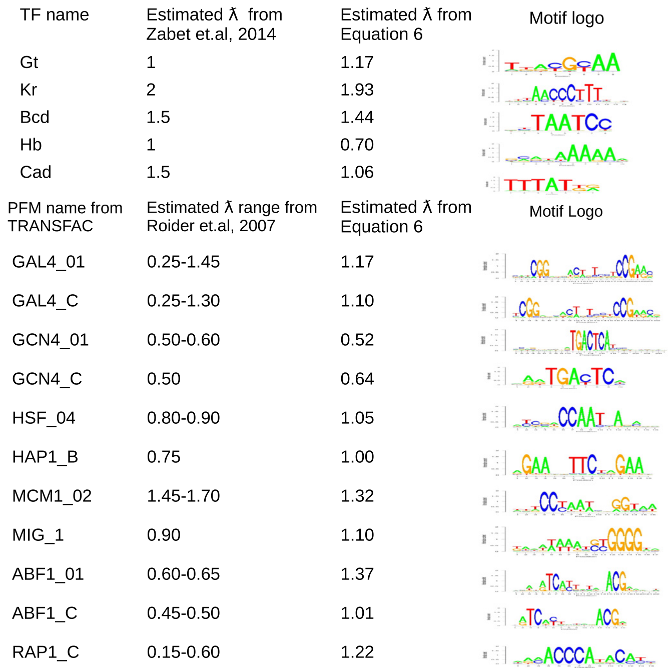
Comparison of λ values estimated with Eq. 6 to λ values derived from fitting ChIP-seq data

We compared our estimated λ values with those estimated from ChIP-seq experiments by Zabet et al. [14] (See Fig. 4). Equation 6 provides a close approximation of all five values estimated in this paper (adjusted $R^2 = 0.64$, p-value = 0.061). We also compare our results with the λ values reported by Roider et al. 2007 [13] for 11 yeast TF motifs from TRANSFAC [4] (See Fig. 4). For each of the 11 TFs, Roider and colleagues fit λ values to ChIP-seq data from cells grown in different growth mediums leading to a range of potential λ values for each TF. However, for each specific cell growth condition, only the most optimal value of λ was selected for each TF, even with circumstances in which there is a plateau in the parameter space with many possible λ values fitting the data nearly equivalently. The range of λ values from their study and the estimated results from Eq. 6 using default parameters are listed in Fig. 4. Our λ value estimations are within, or very close to, their estimated range for 8 out of 11 motifs belonging to 6 out of 8 TFs (absolute differences within 0.25), but another 3 motifs for 2 TFs show poor correlation. It is possible that in some specific cases the assumed default parameters in Eq. 6 could deviate from the real binding properties of these TFs, which can potentially lead to some bias in the estimation of λ. Alternatively, these λ values might lie within the parameter plateau region, and might be a suitable fit for the experimental data.

Converting λ between different PWM matrices of the same TF

In many cases there are two PWMs available for the same TF, and one of these PWMs might already have a reliable estimate of λ from any number of experimental or computational approaches [14]. In such circumstances, we provide a strategy to estimate the unknown λ associated with the alternative PWM. It would be possible to calculate the unknown λ from Eq. 6, but this does not incorporate the additional data available (i.e. the known λ). Our alternative strategy not only incorporates this data, but also loosens the assumption in Eq. 6 that the maximum mismatch energy for DNA binding is proportional to information content.

The procedure to compute a suitable λ is based on the concept of sequence-specific residence time (Eq. 4), as illustrated in Fig. 5. Initially, a well-characterized λ is computed or measured for the first PWM of a particular TF, and then we use this value to derive a λ that is appropriate for the second PWM of the same TF. As part of the calculation of the λ for the second PWM, Fig. 5c shows a heatmap of the estimated residence times for a TF named lame duck (lmd) in a particular binding strength quantile, at different values of λ (ranging from 0.1 to 3.0 as suggested by both [13] and the range of estimated λ using Eq. 6 across different organisms). Both PWMs for the TF come from FlyFactorSurvey database [24], but they are derived from different reports with motif logos shown in Fig. 5b. Blank regions in the heatmap indicate that the choice of λ would generate a residence time outside the range of pre-calculated possible residence times using the first PWM and the existing λ value implying that the λ values for the second PWM are unsuitable. As shown in the heatmap, blank regions often appear in very low values of λ. While if λ is too large, the possible residence time range from weak to strong binding sites is often very restricted, meaning high affinity sites cannot be distinguished from low affinity sites efficiently. λ values with residence times all within the reference range can be further selected, as specified in Methods. Figure 5d–f compares the residence time values between two different PWMs, at different values of λ for the second PWM. We see that the λ in Fig. 5d and 5f would not allow for consistent residence times between the two PWMs, but Fig. 5e does provide consistent results. Therefore, the λ adopted in Fig. 5e is picked up as the suitable value for the second PWM. More examples of residence time heatmaps for converting λ between different PWMs are shown in Additional file 8: Figure S7. In order to evaluate the consistency of λ estimation between the above



TF name	Estimated λ from Zabet et.al, 2014	Estimated λ from Equation 6	Motif logo
Gt	1	1.17	
Kr	2	1.93	
Bcd	1.5	1.44	
Hb	1	0.70	
Cad	1.5	1.06	

PFM name from TRANSFAC	Estimated λ range from Roider et.al, 2007	Estimated λ from Equation 6	Motif Logo
GAL4_01	0.25-1.45	1.17	
GAL4_C	0.25-1.30	1.10	
GCN4_01	0.50-0.60	0.52	
GCN4_C	0.50	0.64	
HSF_04	0.80-0.90	1.05	
HAP1_B	0.75	1.00	
MCM1_02	1.45-1.70	1.32	
MIG_1	0.90	1.10	
ABF1_01	0.60-0.65	1.37	
ABF1_C	0.45-0.50	1.01	
RAP1_C	0.15-0.60	1.22	

Fig. 4 Comparison of λ values estimated with Eq. 6 to λ values derived from fitting ChIP-seq data of Zabet et al. [14] and Roider et al. [13]

method and using Eq. 6, we use the examples of 20 *D.melanogaster* TFs with more than 1 version of PWMs available from different experiments. These PWMs are obtained from the *BioConductor* R package *PWMEnrich.Dmelanogaster.background* [17] and their labels are listed in Additional file 9: Table S2. Since there are only few λ available from binding profile fitting, just for the purpose of illustration, the reference values of λ were precalculated from Eq. 6 instead. New λ values for PWMs obtained from other experiments are computed using both methods and they show good consistency with each other (adjusted $R^2 = 0.88$, Additional file 10: Figure S8). Converting λ between these two PWMs in the opposite direction also show similar results (data not shown). It indicates that both methods provide consistent estimates of λ, even though they have different core assumptions.

Discussion

TF binding site strength estimation using PWM-based methods is essential for modelling TF-DNA interaction in functional genomics; but a proper scaling parameter is needed when using the PWM score to estimate TF binding energy. Therefore, we provide two independent methods for estimating the scaling parameter λ in different conditions. The simple Eq. 6 is widely applicable, since it only requires a PWM as input, which is easy to implement compared to methods using fitting to ChIP-seq [13, 14]. Our second method converts a λ specific to one PWM into λ for a different PWM of the same TF. It is based on the definition of sequence-specific residence time from the facilitated diffusion model of TFs on DNA [16]. This method is particularly useful for converting a previously estimated λ into the one associated with a more up-to-date or otherwise alternative PWMs.

These two methods are consistent with one another (Additional file 11: Figure S4) and with previously established methods. For instance, Eq. 6 can also provide very similar results compared with the estimated λ from ChIP-seq data fitting [13, 14]. Although our estimates of λ are mostly consistent with those estimated by Zabet [14] and Roider [13], it is not possible to robustly compare our λ estimates to experimentally derived values at large scale, as this data is simply unavailable. Having more such data would also enable us to adjust currently fixed parameters in our equation for different TF families, such as the top-scoring threshold, instead of assuming a uniform value across all TFs. The consistent value range of λ in different organisms calculated by this method provides additional support for the applicability of this simple equation. Moreover, the estimated distribution of λ values for different TF families make sense in the light of motif choice for each of the TF families [28]. For example, TFs in the zinc-finger TF super-family, including nuclear receptor zinc-finger and β-β-α zinc-finger families, have more flexible binding motifs, which can suit a wider range of possible binding sites than the helix-turn-helix super-family, which has a more restricted motif consensus sequence [26]. In contrast, some TF families belonging to

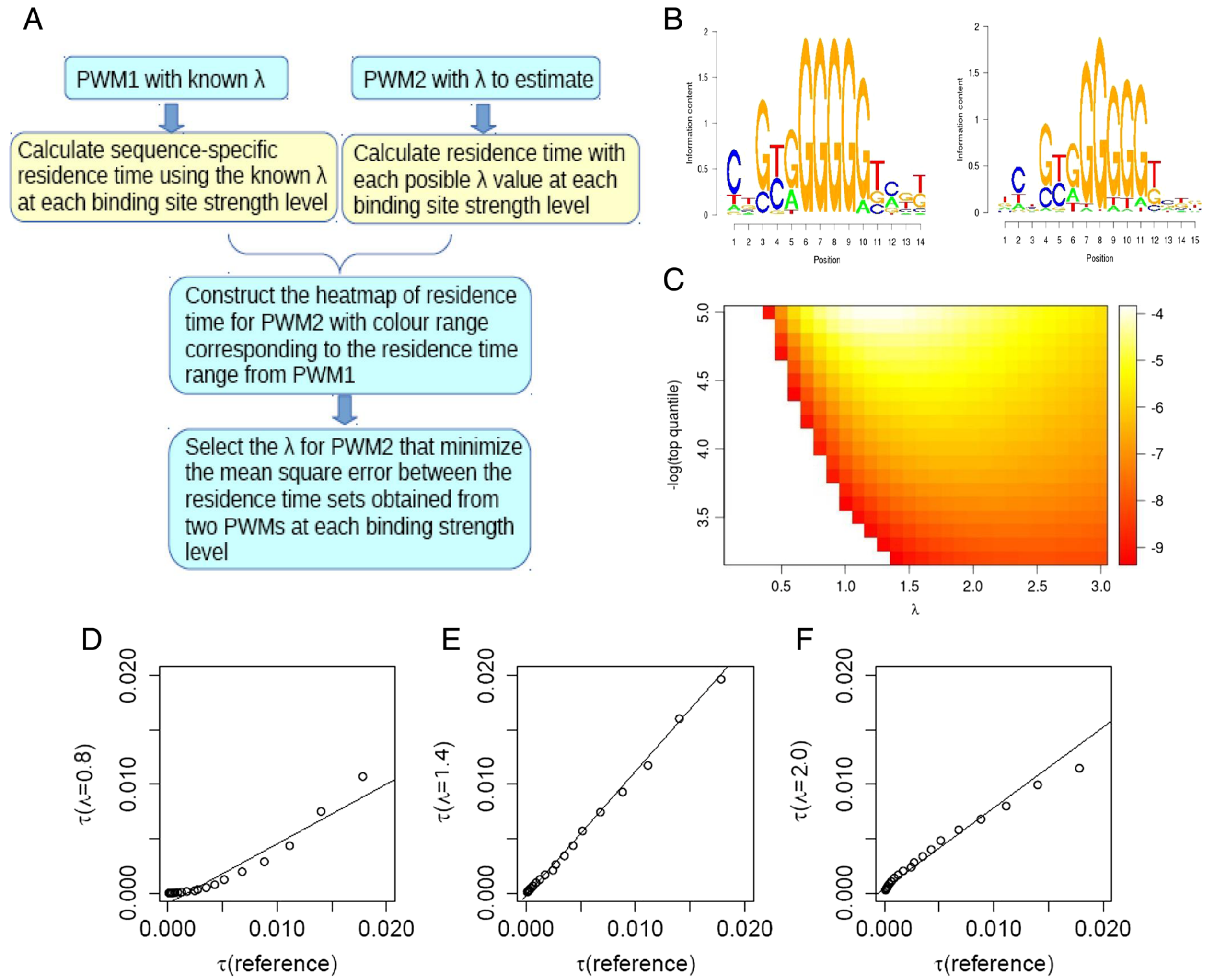


Fig. 5 Conversion of λ between two PWM matrices for the Imd transcription factor. The flow chart shows the procedure to obtain an optimised λ, given two different PWMs and one known and one unknown λ (**a**). Subfigure **b** illustrates the two alternate PWMs for Imd which are available in the FLYFACTORSURVEY database [24]. Equation 6 suggests that PWM1 has a λ of 1.6, and we are trying to find a λ for PWM2. Subfigure **c** is a heatmap of the residence time distribution of PWM2 for different values of λ and different binding site strength level. Each column of the heatmap represents a specific λ value and each row represents a specific binding site strength level measured by the $-log_{10}$ of the corresponding top quantiles from low affinity to high affinity sites. Blank regions in the heatmap indicate λ values which lead to residence time out of the reference scale that is an indication of unsuitable λ values. **d**, **e** and **f** show the correlation of residence time between PWM1 and PWM2 using specific λ values of 0.8, 1.4 and 2.0, respectively. The curve in Subfigure E has the lowest mean square error, and so we assign PWM2 to have a $\lambda = 1.4$

the same super-family and sharing similar binding domain properties can have a strong similarity in λ distribution, *e.g.* homeobox family and forkhead family which both belong to the helix-turn-helix super-family. The two TF families that show the highest average z-score of λ values (namely, basic leucine-zipper and helix-loop-helix families) tend to form homodimers and heterodimers, though some TFs in other TF families also tend to dimerise *e.g.* some members in homeobox family. If PWM motifs for either monomers or dimers are available, the corresponding λ scores can be roughly estimated following the same procedure using Eq. 6, or we can further use the second method mentioned before to convert λ values between different PWMs by the keeping residence time consistent. However, our method only considers TF-DNA interaction, ignoring the effects of TF-TF interactions that could stabilize TF binding.

There are some points that should be noted when using the simple equation method: first, it cannot be applied to very short TF motifs that are less than 6 base pairs in length. Since this method depends on calculating the difference between the PWM score of the 0.1 % highest scoring matches and the maximum score, if the motif is only 5 base pairs in length, the number of possible choices for sequence combination of 5 base pairs is only 1024, then the top 0.1 % highest scoring matches is very likely to be nearly equal to the maximum score. However, most eukaryotic motifs are more than 6 base pairs long. Eukaryotic TFs on average cover 15 bp of DNA with a core motif length of 8-15 bp [8]. Thus, this limitation should not be a problem in the majority of cases. However, if a higher threshold *e.g.* top $1 \cdot 10^{-5}$ is applied with certain adjustment for average mismatch bits in the denominator, it requires the PWM motif to be at least 10 bp long, which will limit the applicability of this simple method. The default cut-off threshold for binding sites is the top 0.1 % of the highest scoring matches, but varying the threshold up to the top 0.001 % does not significantly influence the rank of λ (Additional file 2: Figure S5). Note in Eq. 6, the average mismatch energy bit score in the denominator is the one corresponding to the certain top PWM matches threshold, which means if a new threshold is adopted, the average mismatch energy bit score should be updated accordingly, but given very limited binding energy measurement data, it is difficult to select specific values for each corresponding binding site strength level. Thus, we simply compared the rank correlation of λ, which is not affected by the linear scaling factor of average mismatch energy bits. Although we estimate λ by top scoring genomic sequences, it will not substantially affect the analysis if this is done on random sequences with the same GC content, since given the size of the genome, local binding site patterns will not have much influence on the general distribution of binding site strength. Additional file 1: Figure S6 shows that the number of unique k-mers passing the 0.1 % top scoring matches threshold in genomic sequences correlates well with that in random sequences of the same GC content.

Another assumption in this method is that the mismatch energy tolerance range measured in bits is proportional to the information content of the PWM. This assumption can deal with the bias from the differences in information content of most PWMs; however, it might not hold for PWMs with extremely high information content. For example, the yeast transcription factor IXR1 has an information content of 47 bits according to the PFM from JASPAR [5], which is substantially larger than the average information content of 13.2 bits. In that case, the binding energy will probably be overestimated, which leads to a lower λ, but these cases are very rare and only 7 PWMs in our analysis (less than 1.5 %) have information content greater than 20. Further, we note that the experiment by Maerkl et al. [15] was applied only to TFs belonging to the bHLH family. In the absence of any alternative data, we simply assume that this value is scaled by the information content of the PWM; although if more in-vitro binding energy measurements should become available in the future, we suggest adjusting the specific top score threshold and corresponding average mismatch energy bits accordingly.

There are two limitations of this method, which can potentially lead to some biases between different organisms and different TF families. One limitation is related to the calculation of mismatch energy tolerance in different groups of TF families. We apply a single cut-off threshold of the top 0.1 % highest scoring matches for weak binding sites suggested by Wunderlich et al. [21], but it could be possible that for different TF families, different thresholds should be used due to variations in their DNA binding domains. However, it is difficult to choose specific thresholds for every TF family based on the currently available data. Further, from the definition of information content of the PWM, it sums up information content gain from each nucleotide [20]. It implies that longer motifs including more flanking base pairs will have higher information content compared to the shorter ones with only core motifs, which is an artefact of computation. However, there is no satisfactory way to deal with this problem. One possible solution is using the information content per nucleotide instead of the total information content, but this may be problematic as the information content contributed by flanking sequences constitutes only a very small fraction compared to core motifs. Thus, if dividing total information content by the length of the motif, the dilution of information content could lead to even larger biases. Therefore, instead, in our analysis of comparing λ value distribution across different organisms and TF families, we control for motif length by normalizing it to the mean in each motif length bin. Another potential solution is trying to define a core motif from one PWM, but this requires detailed knowledge about the TF of interest. Additionally, λ will not be a reliable measure of the biochemical stickiness of the TF to the DNA if the PWM itself is not an accurate representation of TF binding. A PWM assumes that each nucleotide position independently contributes to TF binding affinity, which may not be the case [29, 30]. For instance, a study by Storm et al. [31] used both a single nucleotide model and a di-nucleotide model to fit the binding energy measurements [15]. Although they found that the di-nucleotide model provides a better fit to the experimental data, the single nucleotide model could also perform well when non-specific binding energy was taken into account. In addition, the composition of the position frequency matrix of the PWM may contain biases due to the difficulties of attaining an unbiased validated binding site set. Nevertheless, λ can give us insights about DNA binding properties of TFs.

Also, it should be pointed out that residence time in this paper refers to an estimate based on biophysical models [10, 16]. However, other papers report inconsistent scales of residence time according to different experimental approaches. For example, the residence time estimations obtained by Competition-ChIP methods [32] do not share the same order of magnitude compared to the residence times measured by FRAP or single molecular tracking [22, 33, 34], which can probably be an artifact of experimental methods or alternatively, the range of residence time truly varies greatly across different TFs [35]. Because the experimentally determined values are not comparable to each other, we simply adopt bioinformatics-based approaches to compute residence time. Since our method converts λ between different PWMs of the same TF under the concept of residence time, it avoids fitting inconsistent experimental observations and potential variations in DNA-binding kinetics for different TFs.

Although in many cases PWMs are not optimal representations of binding motifs, they have become almost universally adopted to identify potential TF binding sites. It is important to remember that the value of a PWM score is not directly correlated to the binding energy, but rather depends on the scaling parameter λ. Previously, researchers either assumed that λ has similar values across different PWMs or estimated it through computationally intensive binding profile fitting methods [13, 14]. There are several alternative ways to identify potential binding sites based on the p-value of the PWM score [12]. Other studies provide tools to combine more local information *e.g.* DNA sequence conservation and epigenetic marks with PWMs to identify potential binding sites with higher confidence [36]. These methods are useful in defining potential binding sites, but their results are difficult to interpret in terms of TF binding energy which is widely used in modeling TF binding dynamics and enhancer activity [8]. Here we provide two simple strategies for estimating λ, which will let us more clearly link PWM scores with the energetics of TF binding.

Conclusion

Using PWMs as representations of binding motifs have become widely adopted to identify potential TF binding sites. It is important to remember that the value of a PWM score is not directly correlated to the binding energy, but rather depends on the scaling parameter λ. Previously, researchers either assumed that λ has similar values across different PWMs or estimated it through computationally intensive binding profile fitting methods [13, 14]. There are several alternative ways to identify potential binding sites based on the p-value of the PWM score [12]. Other studies provide tools to combine more local information *e.g.* DNA sequence conservation and epigenetic marks with PWMs to identify potential binding sites with higher confidence [36]. These methods are useful in defining potential binding sites, but their results are difficult to interpret in terms of TF binding energy, which is widely used in modeling TF binding dynamics and enhancer activity [8]. Here we provide two simple strategies for estimating λ, which will let us more clearly link PWM scores with the energetics of TF binding. One approach is to simply apply Eq. 6 to estimate λ only based on the given PWM and genome background sequences of a specific organism. It provides results consistent with those estimated by Zabet [14] and Roider [13] in most cases, though there is a small number of exceptions. Further, λ value distribution for different TF families from this method are consistent with DNA binding properties of TF families [26, 28], which further supports the applicability of this simple method. Another approach converts λ between two PWMs of the same TF based on the consistency of residence times. It is useful when we get alternative versions of PWMs from different databases and want to estimate binding site strength in a consistent manner. Both of the approaches we developed are much compute-efficient than previous methods of TF binding profile fitting [13, 14].

Additional files

Additional file 1: Figure S6. Comparison of unique k-mer number passing 0.1 % top PWM score threshold in genomic background versus that in random sequences. For each TF PWM motif, we calculated the logarithm of the number of unique k-mers that passes the threshold in both genomic background and random sequences that have the same GC content and they correlate well with adjusted R^2 equals 0.98, p-value $< 10^{-16}$. (PDF 14.8KB)

Additional file 2: Figure S5. Correlation of λ rank obtained by using different top score thresholds in Eq. 6. We compare the λ rank for different TFs in each group of organisms (subfigure A, B for *S. cerevisiae*, C and D for *D. melanogaster*, E and F for vertebrate PWM motifs) by adopting a different top score threshold of top 0.01 % or 0.001 % instead of the default value of 0.1 % in Eq. 6. The adjusted R^2 for the λ rank correlation between 0.1 % and 0.01 % thresholds for *S. cerevisiae*, *D. melanogaster*, and vertebrate motifs are 0.94, 0.89 and 0.80, respectively, with p-values all less than 10^{-8}. As for the λ rank correlation between 0.1 % and 0.001 % thresholds, the adjusted R^2 are 0.87, 0.92 and 0.74, respectively (p-values all less than 10^{-6}). (PDF 33.2KB)

Additional file 3: Table S3. (TXT 0.232KB)

Additional file 4: Figure S3. Estimated λ distribution in relation to PWM motif length. Each color coded point represents a specific λ value of a TF estimated by Eq. 6 for *S. cerevisiae* (green), *D. melanogaster* (red), and vertebrate (blue).There is a positive correlation between estimated λ value and TF motif length with adjusted R^2 equals 0.33. (PDF 22.6KB)

Additional file 5: Figure S1. The relationship between maximum PWM score and information content of PWMs. Individual dots represents each PWM generated from the non-redundant PFM JASPAR-CORE database [5] after the filtering procedures specified in the Methods section. There is a strong positive correlation between the information content of the PWM and the maximum possible PWM score that could be generated by that PWM, with an adjusted R^2 value of 0.597. (PDF 25.9KB)

Additional file 6: Table S1. (TXT 12.7KB)

Additional file 7: Figure S2. Estimated λ distribution across major TF families. BBA-ZF represents the λ distribution for β-β-α zinc-finger family; NR is zinc-finger nuclear receptor family; L-zipper stands for the basic leucine-zipper family; HLH is helix-loop-helix family; Homeo is homeobox family; FK is fork-head family and HMG is high mobility group family. For each group, λ was calculated by Eq. 6. (PDF 5.33KB)

Additional file 8: Figure S7. Heatmaps for λ conversion between different PWMs. These are additional examples of heatmaps of sequence-specific residence time that are used for λ conversion between different PWMs of the same TF. Alternative versions of PWMs are from *BioConductor* R package of *PWMEnrich.Dmelanogaster.background* [17]. Each column of the heatmaps represents a specific λ value and each row represents a specific binding site strength level. (PDF 25.5KB)

Additional file 9: Table S2. (TXT 1.015KB)

Additional file 10: Figure S8. Consistency of λ estimation between two methods. This figure shows the correlation between λ values obtained from Eq. 6 and from λ conversion using the heatmap of sequence-specific residence time. The adjusted R^2 is 0.88, p-value $= 5.9 \cdot 10^{-5}$. (PDF 11.1KB)

Additional file 11: Figure S4. Comparison of λ values calculated by using different pseudo-count values in PWMs. Subfigure A shows the comparison between the λ values obtained by using PWMs with pseudocounts of 1 and 3 (the adjusted R^2 is 0.973), while subfigure B compares pseudocounts of 1 and 0.3 (the adjusted R^2 is 0.978). Each dot represents a TF from 100 randomly chosen vertebrate TFs in JASPAR database [5]. (PDF 31.2KB)

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
The original study was conceived by XM, BA. The analysis was conducted by XM, DE, CN. The paper was written by XM, DE, BA. All authors approved the final manuscript.

Acknowledgements
We thank Nicolae Radu Zabet for helpful conversations, Robert Foy and Jamie McGinn for editing the manuscript.

Funding
Chinese Scholarship Council (CSC) Scholarship (to X.M.); Marshall Scholarship (to D.E.); TIN2013-41990-R of DGICT, Madrid (to C.N); Royal Society University Research Fellowship (to B.A).

Author details
[1]Department of Genetics, University of Cambridge, Downing Street, CB2 3EH Cambridge, UK. [2]Cambridge Systems Biology Center, University of Cambridge, Tennis Court Road, CB2 1QR Cambridge, UK. [3]Department of Computer Science and Artificial Intelligence, University of Granada, Periodista Daniel Saucedo Aranda, Granada, Spain.

Received: 17 March 2015 Accepted: 8 July 2015
Published online: 20 August 2015

References

1. Mukherjee S, Berger MF, Jona G, Wang XS, Muzzey D, Snyder M, et al. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. Nature Genetics. 2004;36(12):1331–39.
2. Roulet E, Busso S, Camargo AA, Simpson AJ, Mermod N, Bucher P. High-throughput SELEX–SAGE method for quantitative modeling of transcription-factor binding sites. Nature Biotechnology. 2002;20(8): 831–5.
3. Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, et al. An expansive human regulatory lexicon encoded in transcription factor footprints. Nature. 2012;489(7414):83–90.
4. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, et al. TRANSFAC and its module transcompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res. 2006;34(Database issue):D108–10.
5. Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, et al. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Res. 2013;42(Database issue):D142–7.
6. Berg OG, von Hippel PH. Selection of DNA binding sites by regulatory proteins statistical-mechanical theory and application to operators and promoters. J. Mol. Biol. 1987;193:723–50.
7. Berg OG, von Hippel PH. Selection of DNA binding sites by regulatory proteins. Trends in Biochemical Sciences. 1988;13(6):207–11.
8. Kim AR, Martinez C, Ionides J, Ramos AF, Ludwig MZ, Ogawa N, et al. Rearrangements of 2.5 kilobases of noncoding DNA from the drosophila even-skipped locus define predictive rules of genomic cis-regulatory logic. PLoS Genetics. 2013;9(2):1003243.
9. Giorgetti L, Siggers T, Tiana G, Caprara G, Notarbartolo S, Corona T, et al. Noncooperative interactions between transcription factors and clustered dna binding sites enable graded transcriptional responses to environmental inputs. Molecular Cell. 2010;37(3):418–28.
10. Leith JS, Tafvizi A, Huang F, Uspal WE, Doyle PS, Fersht AR, et al. Sequence-dependent sliding kinetics of p53. Proceedings of the National Academy of Sciences. 2012;109(41):16552–57.
11. Ezer D, Zabet NR, Adryan B. Physical constraints determine the logic of bacterial promoter architectures. Nucleic Acids Research. 2014078.
12. Touzet H, Varré J-S. Efficient and accurate p-value computation for position weight matrices. Algorithms Mol Biol. 2007;2(1510.1186): 1748–7188.
13. Roider HG, Kanhere A, Manke T, Vingron M. Predicting transcription factor affinities to DNA from a biophysical model. Bioinformatics. 2007;23(2):134–41.
14. Zabet NR, Adryan B. Estimating binding properties of transcription factors from genome-wide binding profiles. Nucleic Acids Res. 2015; 43(1):84–94.
15. Maerkl SJ, Quake SR. A systems approach to measuring the binding energy landscapes of transcription factors. Science. 2007;315(5809): 233–7.
16. Zabet NR, Adryan B. A comprehensive computational model of facilitated diffusion in prokaryotes. Bioinformatics. 2012;28(11):1517–24.
17. Stojnic R, Diez D. PWMEnrich: PWM Enrichment Analysis. 2014. R Package Version 4.2.0.
18. Stormo GD. DNA binding sites: representation and discovery. Bioinformatics. 2000;16(1):16–23.
19. Wasserman WW, Sandelin A. Applied bioinformatics for the identification of regulatory elements. Nature Reviews Genetics. 2004;5(4):276–87.
20. Stormo GD, Zhao Y. Determining the specificity of protein–DNA interactions. Nature Reviews Genetics. 2010;11(11):751–60.
21. Wunderlich Z, Mirny LA. Different gene regulation strategies revealed by analysis of binding motifs. Trends in Genetics. 2009;25(10):434–40.
22. Mueller F, Wach P, McNally JG. Evidence for a common mode of transcription factor interaction with chromatin as revealed by improved quantitative fluorescence recovery after photobleaching. Biophysical journal. 2008;94(8):3323–39.
23. Pachkov M, Balwierz PJ, Arnold P, Ozonov E, van Nimwegen E. SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates. Nucleic acids research. 2013;41(D1):214–20.
24. Zhu LJ, Christensen RG, Kazemian M, Hull CJ, Enuameh MS, Basciotta MD, et al. FlyFactorSurvey: a database of drosophila transcription factor binding specificities determined using the bacterial one-hybrid system. Nucleic Acids Research. 2011;39(suppl 1):111–7.
25. Kulakovskiy IV, Medvedeva YA, Schaefer U, Kasianov AS, Vorontsov IE, Bajic VB, et al. HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. Nucleic acids research. 2013;41(D1):195–202.
26. Pabo CO, Sauer RT. Transcription factors: structural families and principles of DNA recognition. Annual Review of Biochemistry. 1992; 61(1):1053–95.
27. Itzkovitz S, Tlusty T, Alon U. Coding limits on the number of transcription factors. BMC Genomics. 2006;7(1):239.
28. Luscombe NM, Austin SE, Berman HM, Thornton JM. An overview of the structures of protein-DNA complexes. Genome Biol. 2000;1(1):1–37.
29. Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, et al. DNA-binding specificities of human transcription factors. Cell. 2013;152(1):327–39.
30. Bulyk ML, Johnson PL, Church GM. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. Nucleic Acids Research. 2002;30(5):1255–61.
31. Stormo GD, Zhao Y. Putting numbers on the network connections. BioEssays. 2007;29(8):717–21.
32. Lickwar CR, Mueller F, Hanlon SE, McNally JG, Lieb JD. Genome-wide protein-DNA binding dynamics suggest a molecular clutch for transcription factor function. Nature. 2012;484(7393):251–5.
33. Mueller F, Stasevich TJ, Mazza D, McNally JG. Quantifying transcription factor kinetics: At work or at play? Critical reviews in biochemistry and molecular biology. 2013;48(5):492–514.

34. Chen J, Zhang Z, Li L, Chen BC, Revyakin A, Hajj B, et al. Single-molecule dynamics of enhanceosome assembly in embryonic stem cells. Cell. 2014;156(6):1274–85.
35. Sung MH, Guertin MJ, Baek S, Hager GL. DNase footprint signatures are dictated by factor dynamics and DNA sequence. Molecular Cell. 2014;56(2):275–85.
36. Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome research. 2011;21(3):447–55.