PLoS one

Dinucleotide Weight Matrices for Predicting Transcription Factor Binding Sites: Generalizing the Position Weight Matrix

Rahul Siddharthan*

The Institute of Mathematical Sciences, Chennai, Tamil Nadu, India

Abstract

Background: Identifying transcription factor binding sites (TFBS) *in silico* is key in understanding gene regulation. TFBS are string patterns that exhibit some variability, commonly modelled as "position weight matrices" (PWMs). Though convenient, the PWM has significant limitations, in particular the assumed independence of positions within the binding motif; and predictions based on PWMs are usually not very specific to known functional sites. Analysis here on binding sites in yeast suggests that correlation of dinucleotides is not limited to near-neighbours, but can extend over considerable gaps.

Methodology/Principal Findings: I describe a straightforward generalization of the PWM model, that considers frequencies of dinucleotides instead of individual nucleotides. Unlike previous efforts, this method considers *all* dinucleotides within an extended binding region, and does not make an attempt to determine *a priori* the significance of particular dinucleotide correlations. I describe how to use a "dinucleotide weight matrix" (DWM) to predict binding sites, dealing in particular with the complication that its entries are not independent probabilities. Benchmarks show, for many factors, a dramatic improvement over PWMs in precision of predicting known targets. In most cases, significant further improvement arises by extending the commonly defined "core motifs" by about 10bp on either side. Though this flanking sequence shows no strong motif at the nucleotide level, the predictive power of the dinucleotide model suggests that the "signature" in DNA sequence of protein-binding affinity extends beyond the core protein-DNA contact region.

Conclusion/Significance: While computationally more demanding and slower than PWM-based approaches, this dinucleotide method is straightforward, both conceptually and in implementation, and can serve as a basis for future improvements.

Citation: Siddharthan R (2010) Dinucleotide Weight Matrices for Predicting Transcription Factor Binding Sites: Generalizing the Position Weight Matrix. PLoS ONE 5(3): e9722. doi:10.1371/journal.pone.0009722

Editor: Raya Khanin, Memorial Sloan Kettering Cancer Center, United States of America

Received September 14, 2009; **Accepted** February 26, 2010; **Published** March 22, 2010

Funding: This work was funded by the PRISM project at the author's institution, and (particularly for computational support) by the Department of Biotechnology, Government of India (BT/PR7521/BRB/10/482/2006). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The author has declared that no competing interests exist.

* E-mail: rsidd@imsc.res.in

Introduction

Transcription factors (TFs) are proteins that regulate transcription, the process by which messenger RNA is synthesised from a DNA template. TFs facilitate or inhibit recruitment of the RNA polymerase by binding to DNA, usually near the gene that they regulate. Their binding sites are short nucleotide patterns or "motifs". Detection of such motifs in DNA sequence is therefore of great practical importance in the study of gene regulation. These motifs are not exact strings: while most binding sites for a given factor resemble a "consensus string" (for example, ACGCGT, the most common binding sequence for the MBP1 protein in budding yeast), mismatches and variations often occur.

An early study of the variability and statistical properties of binding sites was by Berg and von Hippel [1]. The most popular representation of binding sites is the position weight matrix (PWM) [2,3], which has a convenient visual depiction, the sequence logo [4]. For a motif of length L, a PWM is a $4 \times L$ matrix, $W_{\alpha m}$, where α is A, C, G or T, and m is an integer ranging over the length L of the binding sequence. $W_{\alpha m}$ is the probability of seeing nucleotide α at position m; the sum over α, for each m, is unity. Typically, a PWM is estimated by aligning a large number of known binding sites, and calculating the relative frequencies of each nucleotide at each position. A "pseudocount" is generally added to the raw nucleotide counts, to allow for the limited size of the data. Thus, given N aligned sequences, where the number of nucleotides of type α at column i is $n_{\alpha i}$ (with $\sum_\alpha n_{\alpha m} = N$ for all m), the weight matrix is given by

$$W_{\alpha m} = \frac{n_{\alpha m} + c_{\alpha m}}{N + C_m} \tag{1}$$

where $C_m = \sum_\alpha c_{\alpha m}$. We choose $c_{\alpha m} = 1$, which corresponds to a "uniform prior" or complete lack of prior bias (formally, a pseudocount is equivalent to assuming a Dirichlet prior: see Materials and Methods for further discussion). A sequence logo [4] is a visual representation where the four possible nucleotides are stacked at each position m, one atop the other, with their relative heights proportional to their weights in the m'th PWM column, and the total height proportional to the "information content" of the PWM column, defined as $I_m = 2 + \sum_\alpha w_{\alpha m} \log_2 w_{\alpha m}$.

A PWM assumes independence among different "columns" (values of m). As an extreme example, it cannot describe a case where two successive positions contain the nucleotides AA or TT equally often but not AT or TA: a weight matrix will contain 0.5 for each of A and T at each position, and will imply that all four of AA, AT, TA and TT are equally probable. For the most part, such strong correlations are not observed among different nucleotides in binding sites, but it is known [5–7] that different sites are not completely independent. Nevertheless, Benos *et al.* [8], argued that the independent approximation is a good one in practice.

A related question is whether the binding *energy* can be written as a sum of single-nucleotide binding energies. Djordjevic *et al.* [9] argued that even with the additivity assumption for the binding energy (which they make), the binding probability should be modelled by a Fermi-Dirac function and not a Boltzmann function, while only the latter (which is the rare-binding limit of the former) can justify the PWM model. However, van Nimwegen *et al.* [10] (supporting text) use a simple maximum-entropy argument to show that the additivity assumption on energy does imply the PWM model for binding sites, if one also makes the reasonable assumption that binding sites have a significantly higher expected binding energy than random sites. Therefore, non-independence of nucleotide distributions in different positions probably implies non-additivity of the binding energy.

Several attempts have been made to go beyond PWMs. A biophysical model was presented by Djordjevic *et al.* [9], while several authors have considered purely statistical/bioinformatic approaches that take account of correlations (or other forms of binding-site heterogeneity not describable by PWMs) in various ways [11–14]. Recently, Sharon *et al.* [14] described a "feature-based" model that enhances the PWM picture with representations of other sequence features, including interdependencies in binding site positions). However, none of these approaches has achieved significant popularity, perhaps because they lack the conceptual simplicity of the PWM.

If the independence assumption is adequate, are nearest-neighbour dinucleotides sufficient? Theoretically, the question is made complicated by the effect of sequence on DNA conformation and bendability, which means that the DNA-protein contact interactions (which, one would expect, are reasonably local) are not the only factor at play. O'Flanagan *et al.* [15] observe contributions primarily from nearest-neighbour dinucleotides. However, Faiger *et al.* [16] report that some TATA boxes (binding sites for the TBP) have context-dependent conformations that require one to go beyond nearest-neighbour non-additivity. Sharon *et al.* [14] consider "features" that are much more complicated than nearest-neighbour dinucleotides. Below (see Results), we examine binding sites in yeast for several transcription factors, and conclude that dinucleotide correlations are significant in several cases, and occur with gaps of all lengths in a binding region, not just with nearest-neighbours.

In fact, it has been known for many years that DNA, particularly non-coding DNA, exhibits long-range power-law correlations [17], for reasons that remain unclear. Therefore, such correlations would not be surprising in binding sites.

A notable case where PWMs appear to be severely inadequate is the binding affinity of nucleosomes. Segal *et al.* [18] used dinucleotide matrices to model nucleosome-binding DNA sequences, but their approach differs significantly from what is described below: notably, they confine themselves to nearest-neighbour dinucleotides. I do not address nucleosomes here, but hope to do so at a future date.

Here I describe a straightforward extension of the PWM method, which reduces to the PWM representation for independent positions. Analogous to a position weight matrix $W_{\alpha m}$, which gives the probability of observing each nucleotide α at each position m, let us define $D_{\alpha_1\alpha_2;m_1m_2}$, a *dinucleotide weight matrix* (DWM) that gives the probability of observing each pair of nucleotides α_1 and α_2 at each pair of positions m_1 and m_2 in a binding site. All pairs of positions are considered: recognising that correlations occur at all scales, we are not restricted to nearest-neighbours (as in [18]), and don't explicitly search for correlated pairs or features (as in [14]).

Defining such an object is easy: but the use of $D_{\alpha_1\alpha_2;m_1m_2}$ is not as straightforward as using $W_{\alpha m}$ in predicting binding sites, because dinucleotide probabilities for different pairs of positions are not independent. With PWMs, one is interested in the likelihood $P(S|W)$ of observing the sequence S given a weight matrix model W; or the log-likelihood ratio $\log(P(S|W)/P(S|B))$ of observing the sequence given W, to observing it given a background model B. These numbers are readily calculated given the PWM and a simple background model: for example, if each nucleotide in the background model is represented by its actual genomic frequency (the model that is actually used throughout this work), $P(S|B) = \Pi_m b_{S_m}$ where S_m is the nucleotide at position m in the sequence, and b_α is the background probability of α. Meanwhile, $P(S|W) = \Pi_m W_{S_m m}$, that is, the product of the weight matrix value for each nucleotide at each position in the sequence. Often, instead of a PWM, a log-odds matrix is used whose entries, when summed, directly yield the log-likelihood ratio (the matrices from yeast ChIP data [19,20], that we use below, are in this format).

No such factorisation is possible for $P(S|D)$, the probability of observing a sequence given a dinucleotide model. However, I introduce here a conceptually straightforward approximation. This is a Bayesian estimate of the posterior probability of each nucleotide at each position n, given the neighbouring sequence (ie, all nucleotides within the putative binding region at all positions $m \neq n$. The product of these posterior probabilities, over all nucleotides, is treated as the likelihood of the sequence; and the log-odds is calculated as usual. The formula reduces, as it should, to the PWM value for any position n if nucleotides at other positions are independent of the nucleotide at n. The formula is derived in Materials and Methods.

There are three complications with this approach, which may account for why such unrestricted DWMs have not been previously used: but the first two are answered here, and I argue that the third is an acceptable price to pay for the increased power.

First, there is the question of how to calculate with joint probabilities, or conditional probabilities, that are not independent. This is answered above; the method should in fact be more widely applicable, and this will be explored in the future.

Second, reliable estimation of $D_{\alpha_1\alpha_2;m_1m_2}$ requires availability of many more sequences than estimation of $W_{\alpha m}$, because there are only 4 nucleotides but 16 dinucleotides. But this is increasingly less of a problem, since dozens of known binding sites now exist for several factors across different species. In fact, based on the benchmark results below, I argue that this approach would be particularly useful in analysing binding data from high-throughput experiments (ChIP-chip or ChIP-seq): these yield thousands of putative binding sites, of which hundreds may be sufficiently high-confidence for this purpose. Details on how to estimate the DWM are in Materials and Methods.

Third, a DWM is a much larger object than a PWM: for a binding sequence of length L, a PWM is $4\times L$-dimensional, while a DWM is $16\times\begin{pmatrix}L\\2\end{pmatrix}$ dimensional. The storage required is quadratic in L. This is exacerbated by one of the key observations below: flanking sequence of several nucleotides improves predictions and appears to play a role in determining binding sites, even when only a "core motif" is prominent in a sequence logo. Therefore, though a PWM for eukaryotic factors is typically between 6 and 15 bp long, the DWM here average 30bp in length (the ideal length of the flank is probably factor-specific, and has not been investigated in detail here). A DWM is also harder to visualise: a "sequence logo" cannot capture correlations. While one can consider a representation of "conditional" sequence logos resulting from fixing particular nucleotides, the result would be unwieldy and not very informative. I argue that PWMs and DWMs can live together (just as "consensus" sequence strings continue to be widely used despite the invention of sequence logos). PWMs have their utility as a concise and easy representation of binding motifs, while DWMs offer much better precision in prediction.

Results

Correlations of gapped dinucleotides, and gap distribution

The first question to be answered is whether going beyond PWMs is important enough to justify the additional complexity of DWMs. We examine 40 transcription factors in yeast (that are further studied in the benchmarks below), each of which has at least 32 predicted targets in MacIsaac *et al.* [20]. For each of the predicted target sequences, the PWM supplied by MacIsaac *et al.* was used to predict the best binding sites, plus any additional binding sites with a log-odds of greater than 3.0. For each factor, all pairs of positions within the binding sites were examined for dinucleotide correlations.

Let n_1, n_2 be two positions within the binding motif, with $1\leq n_1<n_2\leq L$, where L is the length of the motif. Let there also be N binding sequences in total. We also construct a position weight matrix W using these N sequences. For each pair of positions, there are 16 possible dinucleotides $\alpha_1\alpha_2$, each of which is examined. If the PWM hypothesis of position-independence holds, the expected number of sequences containing the nucleotides α_1 at n_1 and α_2 at n_2 will be $<n>=pN$, where $p=W_{\alpha_1 n_1}W_{\alpha_2 n_2}$ is the probability of that dinucleotide. The standard deviation will be $\sigma=\sqrt{Np(1-p)}$. Let n be the number of sequences that actually contain this dinucleotide. In the following, we consider $z=|n-<n>|/\sigma\geq 2$ to be evidence of significant dinucleotide correlations. For normally distributed data, fewer than 0.05% of the data points should differ by more than two standard deviations from $<n>$.

It turns out that out of 1,734 dinucleotide-pair positions studied, 322 deviate from the independent-nucleotide assumption by $z\geq 2$. However, a large number of these cases involve extremely small PWM probabilities, and the number of sequences containing these dinucleotides is rather low (but the expected number, and the expected variance, are both close to zero). Therefore we additionally require that $\max(n,<n>)\geq 0.3N$; this yields 87 dinucleotides, still greater than the unrestricted number of correlated sequences expected by chance.

The next question is how the gaps in these dinucleotides are distributed. Figure 1 shows the answer: while nearest-neighbour dinucleotide correlations are the most common, dinucleotide correlations are found at all spacings. Moreover, the dominance of short-ranged correlations is partly explained by the fact tha there are more short-range pair positions (for a motif of length L, there are $L-n$ dinucleotides separated by a "gap" n ($n=n_2-n_1$, above). Correcting for this produces a somewhat more uniform distribution of gaps, up until roughly ı$g=9$, after which occurs a fall-off. This, fall-off, too, is perhaps explained by the fact that there are fewer factors with long binding motifs.

Detailed examination reveals a few other points: in most cases, $n><n>$, that is, there are more dinucleotides seen than would be expected from the PWM values at those positions. Some factors deviate more from PWM values than others, and in many cases, these are the same factors that perform well in the yeast benchmark below, as described there. For details of all factors and deviating column pairs, see Text S1.

Benchmarks for the DWM method

Two sets of benchmarks are described below: a large benchmark on yeast data, using 40 transcription factors, and a

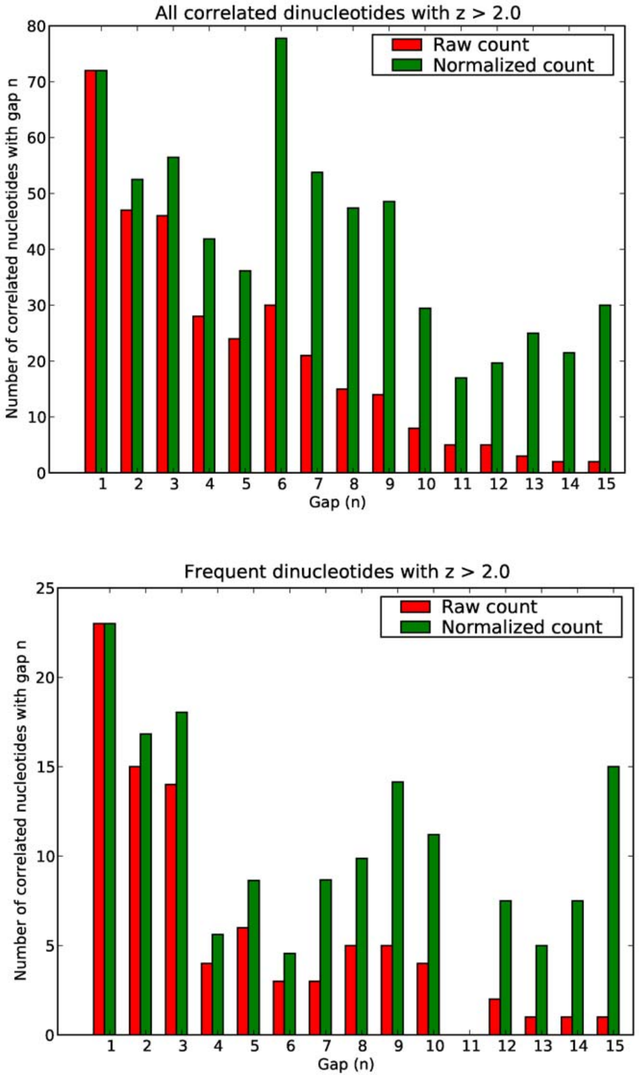

Figure 1. The distribution of gaps in correlated dinucleotide pairs ($z-\text{score}>2$) in yeast TFs, as described in the text. The graph on top shows the full distribution, and the graph below shows only those pairs that are sufficiently abundant (either the predicted or actual number being at least 30% of the total). The green "normalised" bars include a correction for there being fewer possible pairs with larger "gaps". With this correction, the graphs are more uniform.
doi:10.1371/journal.pone.0009722.g001

smaller one on fruitfly data, using the *hunchback* transcription factor. In both cases, predictions from the position or dinucleotide weight matrices that we construct are compared, and compared with previously available ("prior") position weight matrices.

The prior PWMs were used "as is", but in constructing our PWMs and DWMs, sites from the target sequence being benchmarked were excluded. This is important since, when the number of sequences is relatively small, such "self-prediction" can significantly affect the results, especially in the dinucleotide case.

Binding site predictions in yeast

These benchmarks use the genome-wide binding data from ChIP-chip experiments reported by Harbison *et al.* [19] and the revised predicted targets reported by MacIsaac *et al.* [20]. For 40 transcription factors that had at least 32 predicted high-confidence targets, we constructed new PWMs and DWMs, with and without 10bp flanks, as described in Materials and Methods. The matrices were constructed using predicted sites in the targets, but as observed above, "self-prediction" was avoided. Therefore, if there were N targets, $N+1$ matrices were constructed: one that used all targets as data, and one omitting sites from each target by turn, to be used in predicting sites for that target. The prior PWMs, constructed PWMs, and constructed DWMs were used to predict binding sites on all sequences from the original ChIP experiments. The results were compared with the raw binding "p-values" for the same sequences reported by Harbison *et al.*, as well as with the predicted targets from MacIsaac *et al.*

Figure 2 shows the Pearson coefficient of correlation with binding data in Harbison *et al.* [19]. The calculation is described in Materials and Methods. This figure only shows those 26 factors for which predictions correlate with a coefficient of at least 0.3 for at least one of the three methods shown (the original PWM, our DWM without flank, or our DWM with flank). In nearly all of these cases, the dinucleotide matrix, and in particular the DWM that includes flanking sequence, greatly outperforms the PWM. Data for all the factors, and also for our "posterior" PWMs, are portrayed in Figure S1.

One may ask whether the improved coefficient of correlation is merely a consequence of the fewer predictions made by the DWMs. To answer this, Supporting Figure S2 shows (for all 40 factors) the coefficient of correlation for the top P predictions from the prior PWM, where P is the number of predictions made by the DWM with 10bp flank, plus any further predictions with the same logodds as the lowest in this set. In many (but not all) cases, the correlation coefficient is improved; however, in most cases, it remains well below what is achieved by the DWM.

Figure 3 shows the "precision" of predictions for the annotated target genes [20], that is, the fraction of predictions at or above a given logodds cutoff ℓ that are listed as a target, as a function of the sensitivity to known targets, that is, the fraction of listed targets that are found at or above the logodds cutoff ℓ. The prior and new position weight matrices, without flanking sequence, perform very similarly. While either adding flanking sequence alone, or using a dinucleotide matrix alone, cause notable improvements (the dinucleotide WMs without flank are about 50% to 100% more specific than the prior PWMs), DWMs with flank achieve nearly perfect precision over most of the range of sensitivity. Note that the precision here refers to gene target, not to individual binding sites.

To measure sensitivity to individual binding sites, we combined these data with the *Saccharomyces cerevisiae* Promoter Database (SCPD) [21]. 19 of the 40 factors that we consider contain annotated sites in SCPD. Figure 4 plots the fraction of site predictions for these 19 factors that are annotated in SCPD ("precision" to SCPD), as a function of the total number of SCPD sites predicted. Since SCPD is far from an exhaustive database, false positives cannot be counted, but these "precision" curves are hopefully reflective of the true precision if all true binding sites were known.

Finally, we observe some interesting points about specific factors. For each factor, if we look at the number of column pairs that are more than 2 standard deviations away from the PWM expectation (Results, first subsection), and also ask that either the

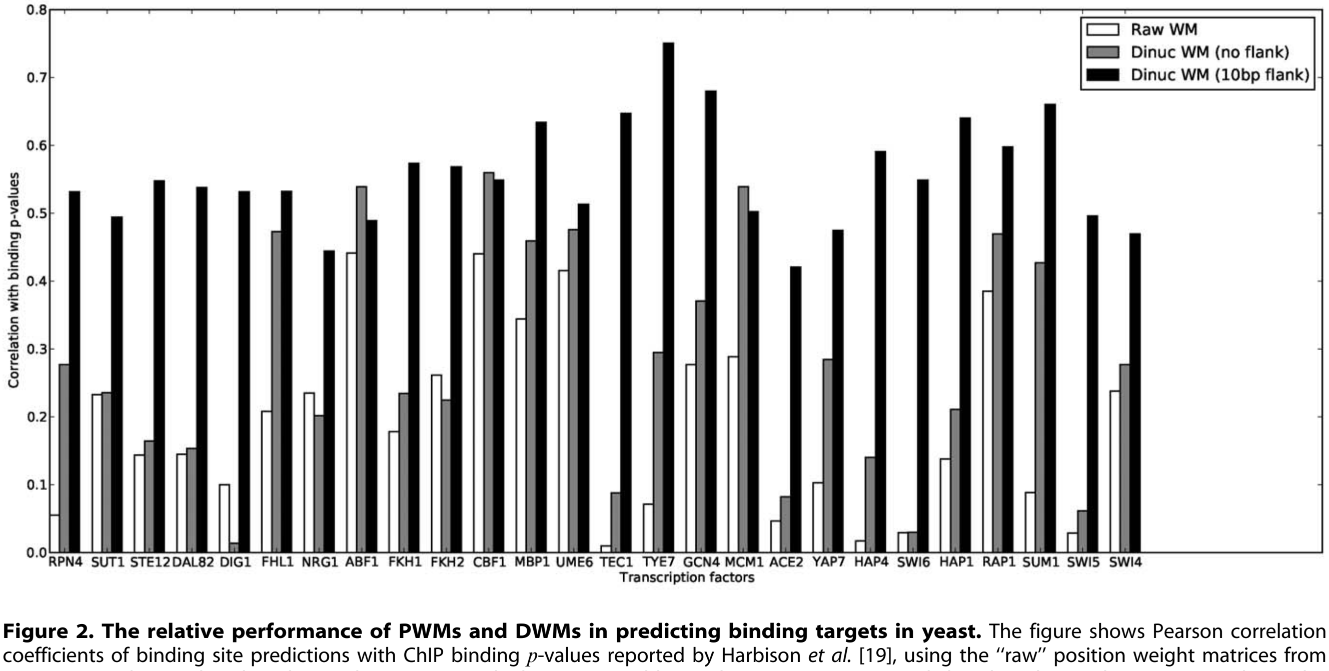

Figure 2. The relative performance of PWMs and DWMs in predicting binding targets in yeast. The figure shows Pearson correlation coefficients of binding site predictions with ChIP binding *p*-values reported by Harbison *et al.* [19], using the "raw" position weight matrices from MacIsaac *et al.* [20], dinucleotide weight matrices with the same "width" as the "raw" matrices, and dinucleotide weight matrices with a 10bp "flanking sequence" on either side of the input matrices. Details are in Materials and Methods.

doi:10.1371/journal.pone.0009722.g002

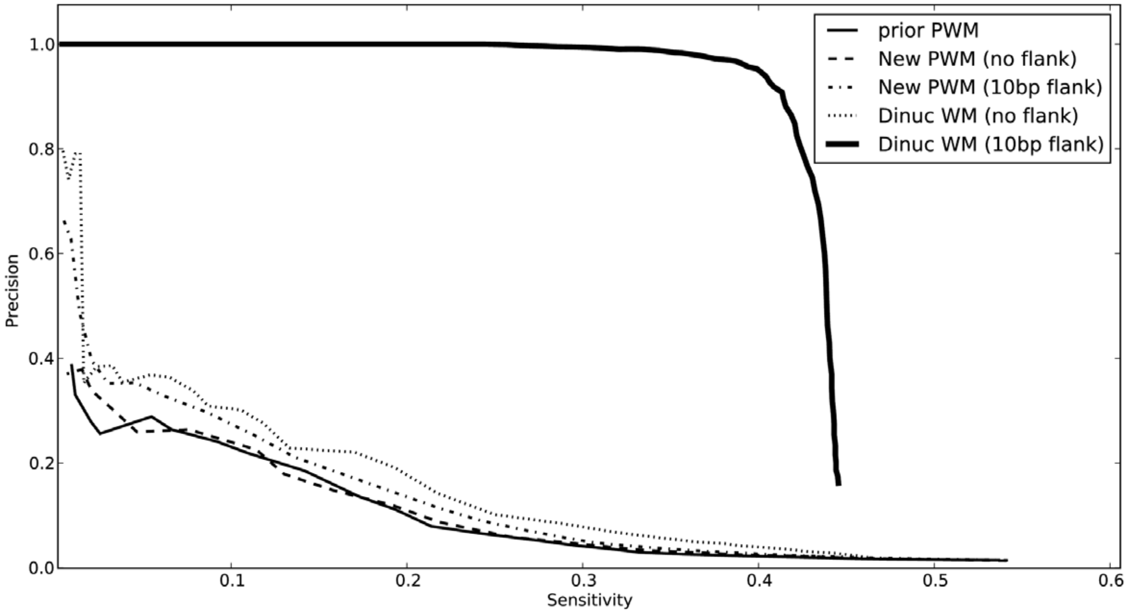

Figure 3. The precision, as a function of sensitivity, of PWMs and DWMs in predicting targets from MacIsaac *et al.* [20]. The precision is the fraction of predictions above a certain logodds cutoff ℓ that correspond to documented target genes. The sensitivity is fraction of known targets that are predicted above that cutoff. These are for the same benchmark data as in Figure 2.
doi:10.1371/journal.pone.0009722.g003

expected or the observed number of dinucleotides is 30% of the number of sequences, we find (as noted earlier) that, across all factors, there are 87 correlated column pairs. Looking at individual factors, we find that there are 10 factors that have 3 or more correlated column pairs, namely RPN4, FKH2, CBF1, ABF1, DIG1, HAP4, TEC1, SUM1, STE12 and MCM1 (which has a remarkable 19 column pairs showing significant correlation). Comparing with Figure 2, we find that for eight of these factors the DWM method greatly outperforms the PWM method: the exceptions are ABF1 and CBF1.

Maerkl and Quake [7] studied the basic helix-loop-helix factors PHO4 and CBF1, together with two human factors, and argued that PWMs are insufficiently able to discriminate while providing many false positives. While PHO4 is not in our list (having only 23 predicted high-confidence targets) and DWMs do not perform notably better than PWMs for CBF1, it is notable that in the case of another HTH factor with a similar binding motif (TCACGTG), TYE7, PWM predictions correlate very poorly with binding data while DWM predictions correlate nearly perfectly. Similarly, ACE2 and SWI5, homologous factors [22] which share similar binding

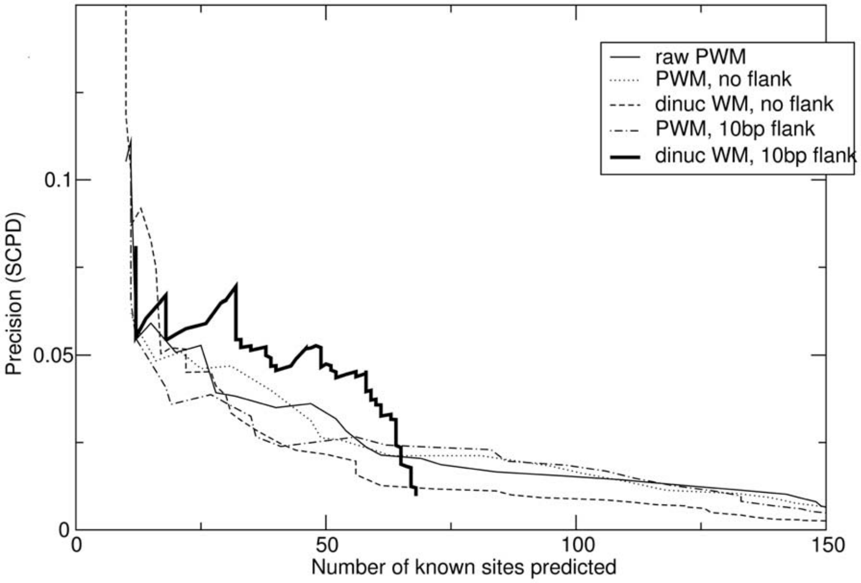

Figure 4. The performance of different methods on individual site predictions in yeast. For the same benchmark as in Figure 2, these are the fraction of site predictions that agree with annotated sites in SCPD, as a function of the total number of SCPD sites predicted.
doi:10.1371/journal.pone.0009722.g004

motifs (GCTGGT), are much better discriminated by DWMs than by PWMs; as, to a lesser extent, are MBP1 and SWI4, which are homologous cell-cycle-related proteins [23]. In many of these cases, including flanking sequence improves the results, suggesting that flanking nucleotides show significant correlations with nucleotides within the core motif, or with one another.

Binding site predictions in fruitfly

The REDfly (formerly FlyReg) database [24,25] contains curated DNAse I footprints of binding sites for several transcription factors in *Drosophila melanogaster*. These form a useful resource for benchmarking, but since a rather small fraction of functional sites are likely to be annotated in this database, the benchmark here uses synthetic sequence that contains embedded REDfly footprints as well as synthetic samples from PWMs, as described in Materials and Methods. The goal was to predict the functional sites, and also to discriminatively predict the functional sites rather than the synthetic samples, using PWMs and DWMs.

Several ChIP-on-chip experiments for transcription factors in *Drosophila melanogaster* have been reported in the literature. Here, data from Li *et al.* [26], who studied six factors, are used. For reasons explained in Materials and Methods, I focussed on the factors *bicoid* (*bcd*), *hunchback* (*hb*), and *kruppel* (*Kr*). Using prior PWMs from B1H data in [27], binding sites were identified in the ChIP peaks and used to construct PWMs and DWMs, with and without a 10bp flank. Peaks that overlap with REDfly footprints were carefully excluded, for the reasons noted earlier.

The results for the *hunchback* factor were impressive. The binding motif for this factor is a weak poly-A pattern that is abundant in the genome; it appears that the dinucleotide method in this case significantly improves the precision of predictions, and, as in the case of the yeast factors, flanking sequence plays a role.

With the other factors (*bicoid* and *kruppel*), the dinucleotide method did not show improvement over the PWM method (data not shown), and in fact, in the case of *bicoid*, the input (B1H-derived) PWM showed significantly better precision in predicting REDfly footprints than even the ChIP-derived PWMs. The reasons are unclear, but a more thorough study of *Drosophila* factors is in progress. Meanwhile, *kruppel* binds to a relatively sharp and well-defined motif, so it is possible that there is no important additional information in dinucleotide correlations.

Figure 5 plots the precision of *hunchback* predictions for real (REDfly) footprints. Figure 6 plots the "discriminative precision". Here the precision is defined as n_{real}/N_p and the discriminative precision is $(n_{\text{real}} - n_{\text{synth}})/N_p$, where N_p is the total number of predictions above a particular logodds cutoff, n_{real} is the number of predictions that overlap real (REDfly) footprints, and $n_{math\text{synth}}$ is the number of predictions that overlap sites that were sampled from the PWM, as a function of the "sensitivity", the fraction of real REDfly footprints that are overlapped by predictions above the same cutoff. Unlike in the yeast SCPD benchmark, these sites are embedded in synthetic sequence; therefore any prediction that is not a REDfly footprint can safely be termed a "false positive". Given the variability of TF binding widths as well as REDfly footprints, and also given the large size of many of the REDfly footprints, predictions whose midpoint lay within 10bp of the REDfly footprint were considered "hits".

The results suggest that the precision of dinucleotide-model predictions is substantially better than PWMs, for a given sensitivity, and for high-confidence predictions DWM predictions are nearly twice as specific to REDfly sites as PWM predictions. But the sensitivity of dinuc WMs is substantially less than PWMs, especially when flanking sequence is included. With the "discriminative precision" the difference is even sharper: PWM predictions mostly have negative discriminative precision, that is, they resemble synthetic samples from themselves more strongly than they resemble actual binding sites; and while the discriminative precision of DWMs gets better for higher-confidence predictions, PWMs actually perform worse in this regard. For *hunchback*, then, DWMs with flanking sequence are clearly better able to distinguish genuine binding sites from similar sequences generated as samples from the respective PWMs.

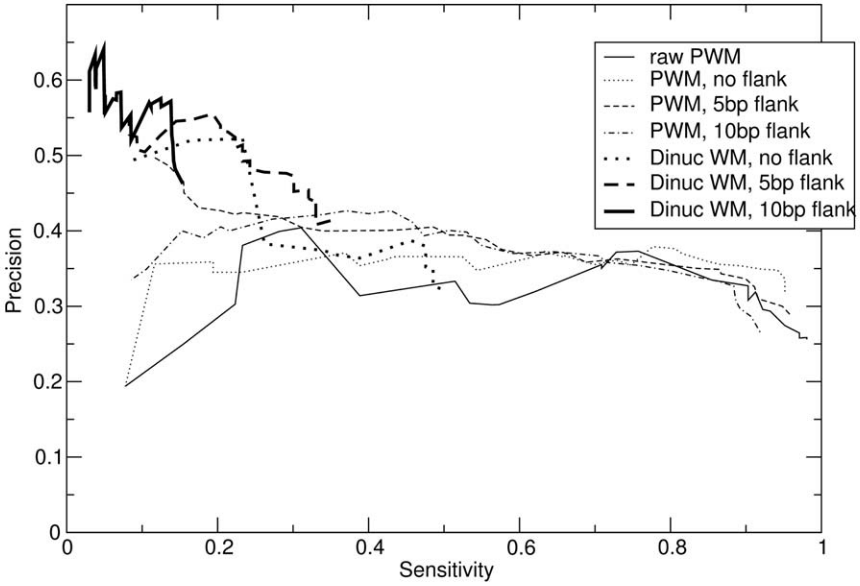

Figure 5. The precision of site predictions in fruitfly. For predictions in synthetic sequence embedding binding site footprints from the REDfly database as well as "fake" sites that are samples of PWMs corresponding to the same factors, this plot shows the precision in predicting REDfly sites, that is, the fraction of predictions that overlap with REDfly footprints, as a function of sensitivity, that is, the fraction of real (REDfly) sites that are predicted. Details of the construction of the synthetic sequence are in Materials and Methods.

doi:10.1371/journal.pone.0009722.g005

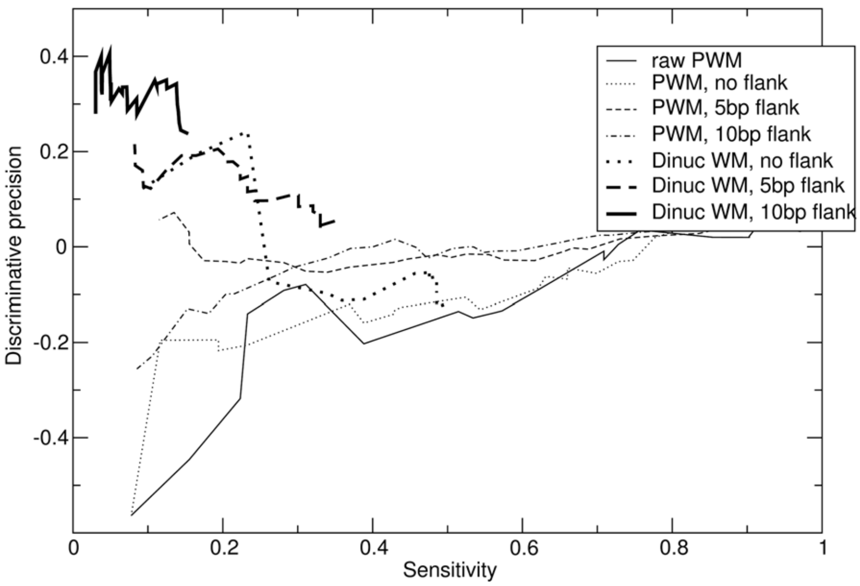

Figure 6. The *discriminative* precision of predictions in fruitfly. For the same predictions as in Figure 5, this plot shows the "discriminative precision" for REDfly sites, that is, difference in the fraction of predictions that overlap with REDfly footprints and the fraction of predictions that overlap with "fake" sites, as a function of sensitivity.
doi:10.1371/journal.pone.0009722.g006

Discussion

The benchmarks on ChIP-characterised factors in yeast and fruitfly suggest strongly that the DWM provides a much-improved representation of binding sites for many transcription factors. In a way, this is unsurprising: DWMs contain, in the worst case (the case of completely uncorrelated positions in a motif), the same information as PWMs, and in other cases much more information; and this must be reflected in their predictive power. The DWMs here were constructed excluding sites from target sequence: therefore, it is reasonable to assume that "complete" DWMs will perform even better. The important points are that, first, the approximation to the likelihood in equation (10) is useful and works well here (its usefulness outside this narrow context remains to be explored, but one can be optimistic); second, a few dozen known binding sites are sufficient to arrive at a reasonably high-quality DWM; third, flanking sequence appears to play a significant role, that is not so strongly apparent when using PWMs.

This approach is thus very promising for the future. While the starting point of an investigation may be a PWM based on a few binding sequences, such a PWM combined with possibly noisy genome-wide binding data may perhaps be used to "bootstrap" a DWM representation. That DWM may in turn be used to predict more binding sites, with much greater confidence than a PWM can ever do.

However, for some factors in yeast, and for *bicoid* and *kruppel* in fruitfly, DWMs seemed to not perform better than PWMs, or even performed worse. The reasons need to be understood, but it may simply be inadequate prior binding data in some cases. Further work is in progress on *Drosophila* factors.

This paper uses a naive method for predicting sites: the log-odds for the binding sequence being explained by a PWM or DWM over a background model. Better methods are commonly implemented with PWMs, for example, using biologically-motivated prior binding probabilities; and taking account of competition between different factors (for example, Stubb [28], a *cis*-regulatory module prediction program). In principle, all the same improvements can be applied to DWM predictions.

It would be of great interest to relate DWMs with a more biophysical binding-energy model of protein-DNA interactions. Just as PWMs can be derived from a simple binding-energy model with some additional assumptions ([10], supporting text), DWMs should be justifiable in terms of protein-DNA binding energetics. As noted earlier, non-independence of nucleotide distributions at different positions probably implies non-addititivity of the binding energy, and this should be taken into account in building improved models.

Ab initio motif-finding and prediction of binding sites using DWMs, and the usage of homologous sequence from related species to improve predictions, are interesting topics that deserves to be addressed in the near future, perhaps as extensions of the PhyloGibbs program [29,30]. Predicting *cis*-regulatory modules using this formalism would also be a useful and interesting exercise.

In summary: The dinucleotide weight matrix described here is easy to calculate, though cumbersome. The method described here of calculating posterior probabilities of binding sites is straightforward, though approximate. When large numbers of binding sites are already known, this formalism should be preferred to PWMs in predicting new sites.

However, it should be emphasised that the DWM formalism presented here is subject to further modification and refinement. In particular, the question of the appropriate "pseudocount" to apply to DWMs is not easy and the answer here is by no means definitive. The appropriate length of flanking sequence is probably highly factor-specific. Lusk and Eisen [31] recently observed that the "cutoff score" used to imply significance for PWM-based binding site predictions is probably variable across factors, and the same will certainly be true for DWM-based predictions. Therefore, while DWMs represent a significant advance over PWMs in predictive power, a "one-size-fit-all" solution to the problem of binding-site prediction is unlikely to exist.

Implementation and availability

All benchmarks listed here were performed using scripts written in Python by the author. These are not user-friendly but are available, with some basic documentation, from the author for interested users. The DWMs generated for the factors discussed in this paper are available as Python "pickle" dumps (which can be loaded and used by other Python programs). A user-friendly, fast implementation of these methods in a compiled language is planned in the future.

Materials and Methods

Constructing PWMs and DWMs from binding-site data

Given N known aligned binding sequences, a PWM can be constructed with normalised base counts in these sequences: in column m, let there be $n_{\alpha m}$ instances of the nucleotide α, with $\sum_\alpha n_{\alpha m} = N$. Then, for N large, $W_{\alpha m} = n_{\alpha m}/N$. Usually N is not terribly large, so one instead uses

$$W_{\alpha m} = \frac{n_{\alpha m} + c_{\alpha m}}{N + C_m} \tag{2}$$

where $c_{\alpha m}$ is a "pseudocount" and $C_m = \sum_\alpha c_{\alpha m}$. Formally, this is the same as assuming a Dirichlet prior on m'th column of the weight matrix, $P(W_{(.)m}) \propto \Pi_\alpha w_{\alpha m}^{c_{\alpha m}-1}$. (See the book by Durbin *et al.* [32] for a discussion.) The special choice $c_{\alpha m} = 1$ expresses complete prior ignorance of W, and is generally appropriate for estimating weight matrices.

$$W_{\alpha m} = \frac{n_{\alpha m} + 1}{N + 4}. \tag{3}$$

We use this choice to construct our PWMs, both for direct benchmarking and for use in the DWM formulas derived below.

If we were completely ignorant of dinucleotide probabilities, we should use the analogous expression to construct DWMs:

$$D_{\alpha\beta;mp} = \frac{n_{\alpha\beta;mp} + 1}{N + 16} \tag{4}$$

where $n_{\alpha\beta;mp}$ is the number of sequences where nucleotide α is found at column m and β at column p. But we know that, in practice, different columns tend to be roughly independent (that is, PWMs generally work well); and for a given N we have a much better estimate of the PWM W than of the DWM D. Therefore, instead of the pseudocount 1 that implies complete ignorance, we use as our prior the product of the corresponding PWM columns, normalised to sum to 16:

$$c_{\alpha\beta;mp} = 16 W_{\alpha m} W_{\beta p}, \tag{5}$$

$$D_{\alpha\beta;mp} = \frac{n_{\alpha\beta;mp} + 16 W_{\alpha m} W_{\beta p}}{N + 16}. \tag{6}$$

Other choices of priors and pseudocounts are, of course, possible, but the choices above are straightforward and work well.

Using the DWM to calculate posterior probabilities

We would like to calculate $P(S|D)$, that is, the probability that a putative binding sequence S is "explained" by a dinucleotide model D. (We can compare it to $P(S|B)$, the probability of it arising from a "background model" B; the ratio of these is the "odds", and the logarithm of this ratio is the "log-odds".)

First we write

$$\begin{aligned} P(S|D) &= \prod_{n=1}^{L} P(S_n|S_1 S_2 \ldots S_{n-1} S_{n+1} \ldots S_L, D) \\ &\equiv \prod_{n=1}^{L} P(S_n|\{S_{m\neq n}\}, D) \end{aligned} \tag{7}$$

that is, the probability of observing the sequence is the product of the probabilities of each nucleotide S_n given all the other nucleotides $S_{m\neq n}$ in the sequence, and given the dinucleotide model D. (This is an approximation: the sum of this over all sequences will not be exactly 1, though the sum of each factor over S_n is 1. However, since we use it essentially as a discrimination score, we ignore this matter). We estimate these individual nucleotide probabilities using the Bayesian expression

$$\begin{aligned} &P(S_n|\{S_{m\neq n}\}, D) = \\ &\frac{P(\{S_{m\neq n}\}|S_n, D) P(S_n)}{\sum_{\alpha = A,C,G,T} P(\{S_{m\neq n}\}|S_n = \alpha, D) P(S_n = \alpha)}. \end{aligned} \tag{8}$$

Here, for the prior probabilities $P(S_n)$ and $P(S_n = \alpha)$ we use the single-nucleotide weight matrix values $W_{S_n n}$ and $W_{\alpha n}$. Finally, we approximate the likelihood of neighbouring sequence given the nucleotide S_n as the product of individual conditional probabilities:

$$\begin{aligned} P(\{S_{m\neq n}\}|S_n, D) &= \prod_{\substack{m=1 \\ m\neq n}}^{L} P(S_m|S_n, D) \\ &= \prod_{\substack{m=1 \\ m\neq n}}^{L} D_{S_m S_n;mn}/P(S_n) \\ &= \prod_{\substack{m=1 \\ m\neq n}}^{L} D_{S_m S_n;mn}/W_{S_n n}. \end{aligned} \tag{9}$$

That is, we write this likelihood as a product of conditional probabilities of the individual nucleotides $S_{m\neq n}$ given S_n; these conditional probabilities are evaluated in the usual way, $P(B|A) = P(B,A)/P(A)$. Putting all of this together, the final expression for $P(S|D)$ is

$$P(S|D) = \prod_{n=1}^{L} \left(\frac{1}{C_n} \left(\prod_{\substack{m=1 \\ m\neq n}}^{L} \frac{D_{S_m S_n;mn}}{W_{S_n n}} \right) W_{S_n n} \right) \tag{10}$$

where C_n is a normalisation constant for the nth factor in the product (equal to the denominator in equation 8). In the case that there are no dinucleotide correlations, we have $D_{\alpha\beta;mn} = W_{\alpha m} W_{\beta n}$ for all α,β,m,n, and the expression reduces to the PWM-based answer, $P(S|W) = \Pi_n W_{S_n n}$.

Yeast binding site prediction benchmarks

Of the factors studied in the ChIP-on-chip benchmarks reported in Harbison *et al.* [19], 40 factors were selected that had at least 32 targets annotated in MacIsaac *et al.* [20], with a *p*-value of 0.001 or better and conservation in 2 species (filename orfs_by_factor_p0.001_cons2.txt), and with a corresponding sequence (or sequences) in the microarray probe file (filename yeast_Young_6k.fsa). Prior matrices were taken from supporting data of MacIsaac *et al.* (filename v1.tamo). Raw *p*-values for binding were taken from supporting data of Harbison *et al.* (filename Harbison_Gordon_yeast_v9.11.csv). All these files were downloaded from the supporting data pages hosted by the authors of those papers. The prior matrices are in logodds format, in most cases using genomic single-nucleotide frequencies for the background model; they were converted to position weight matrices that give the probabilities of individual nucleotides.

The posterior PWMs and DWMs were constructed by the following two-step process: first, for each predicted target, the highest-scoring sites were selected using the prior PWM, with the following criterion: all sites with a logodds of above 3.0 (natural logarithm) were selected. If there were none, but the best site had a logodds of at least 1.5, that site alone was selected. If there were no good matches, the sequence was rejected. These putative binding sequences were aligned (with or without a flanking sequence of 10bp) and used to construct new, interm PWMs and DWMs. These DWMs and PWMs, were in turn used to predict sites in all targets, using the same logodds criteria as earlier (with no additional flanking sequence). The resulting predictions were aligned to construct the final "posterior" PWMs and DWMs, with one difference: in addition to "full" PWMs and DWMs, "partial" PWMs and DWMs were also constructed for each contributing probe by omitting all binding sequences from that probe, and these partial matrices were used for the predictions in that probe described below, in order to ensure that all predictions were based on matrices of completely independent origin.

For constructing the PWMs and DWMs, I chose to use predicted sites, rather than experimentally validated sites, because there are not sufficient numbers of the latter available for most factors. While a genome-wide PWM-based or DWM-based bioinformatic search for binding sites is likely to pick up many false positives, we argue that if we confine the search to regions that are predicted by ChIP experiments to be bound, with high confidence, to the TF in question, and only select the most likely predictions in these regions, these are much less likely to be false positives, while also being much more numerous than experimentally validated sites in databases such as SCPD. Also, while a genome-wide PWM-based search is unlikely to result in positional correlations within predicted sites, such correlations are arguably more likely when only predictions in ChIP-validated regions are considered; and the "bootstrapping" procedure of using the initial DWMs to predict a new set of sites should result in further refinement. These remarks also apply to the methods used for *Drosophila* factors described below.

These PWMs and DWMs, as well as the prior PWMs, were then used to predict sites in every probe sequence in the microarray probe file. To construct Figure 2 and Figure S1, for each method, the total logodds prediction for each probe sequence was calculated (that is, the logodds at each site was summed over all possible sites, with the better of two "orientations" chosen at each site): this was done in order to treat equally the cases of a factor having a few highly specific sites, or several weaker sites. This was cross-correlated with the geometric mean of the "rich medium" *p*-value and the (up to) two best other *p*-values in the file cited above. Up to two other values were averaged because there could be cases where a TF does not bind strongly in the default "rich medium" condition but does bind more strongly under certain other biological conditions, for reasons that cannot be predicted in this sort of bioinformatic analysis. However, if binding was not reported in at least two other conditions, fewer than two other *p*-values were averaged. The Pearson correlation coefficients (with the probe as independent data, and the "total logodds" and "mean *p*-value" as dependent data) are plotted. The calculation is over all probes.

Figure 3 was plotted using the same data, as described in the main text.

The SCPD database includes binding data for 234 factors/complexes in yeast, of which 19 were common to the list of 40 factors studied above. 208 binding regions were annotated for these 19 factors. I extracted these regions, converted them to genomic coordinates, and analysed the precision of the previous site predictions for these factors of the PWM and DWM methods to these sites, as a function of the number of known sites recovered. For this purpose, since the SCPD coordinates are widely variable in size (some "sites" are only one nucleotide long), and the PWMs and dinucleotide WMs are also of different sizes, the following criterion was used: if the midpoint of the annotated SCPD region was within 10bp of the midpoint of the predicted binding site, the region was considered successfully predicted.

Fly binding site prediction benchmarks

The ChIP data used here were taken from Li *et al.* [26]. In this preliminary investigation, to ensure high confidence in predictions, those factors were taken that were bound by 2 antibodies: namely, *hb*, *bcd*, *kni* and *Kr*; and only peak positions overlapped by both sets of antibodies were considered. This yielded 83 peaks for *bcd*, 230 for *hb*, and 818 for *kr*, but only 12 for *kni*. The latter was accordingly dropped and the former three used. Footprints for all of these were obtained from the REDfly database [25], and any footprints that overlapped with the peak list were removed from the peak list. Prior PWMs were obtained from the B1H study of Noyes *et al.* [27], and used to construct posterior PWMs and DWMs in the same two-step manner as described in the yeast benchmark.

The benchmark was on synthetic sequence in which the actual REDfly footprints for each factor, plus 10bp flanking sequence, were embedded. These footprints, with flanking sequences, were separated by 100bp random spacer sequences. In addition, synthetic sequence of the same length, but containing embedded samples of PWMs rather than actual REDfly footprints, was included. The number of copies of synthetic sites was the same as the number of copies of real sites, for each factor. Predictions where the centre of the prediction lay within 10bp of the footprint region were considered "hits".

Supporting Information

Text S1 Details of significant dinucleotides in yeast TFs.
Found at: doi:10.1371/journal.pone.0009722.s001 (0.11 MB TXT)

Figure S1 Details of performance of PWMs and DWMs on yeast TFs. Pearson coefficients of correlation for logodds predictions with published binding p-values for all 40 factors studied, for all matrices used (prior PWM, posterior PWMs and DWMs with and without flanking sequence). Also shown is the correlation for prior PWMs when only the top N are considered, where N is the number of predictions from the DWM with flanking sequence, plus any additional predictions with an equal log-odds score. In addition, sequence logos are shown for the prior PWMs and the posterior PWMs with flanking sequence, in both orientations. In most cases, the logos are extremely similar and

there is little sequence signature in the flanking sequence at the PWM level.
Found at: doi:10.1371/journal.pone.0009722.s002 (1.70 MB PDF)

Acknowledgments

I would like to acknowledge useful discussions, and past and ongoing collaborations, with K VijayRaghavan and K G Guruharsha, which have motivated much of my recent thinking on transcriptional regulation. Some preliminary work along these lines was carried out with the help of P V Kiran Kumar and P Ramesh Nadh in 2008. In that study, the focus was on the *hb* and *bcd* factors in fruitfly only, flanking sequence was not considered, and posteriors were calculated differently; but the results were suggestive.

Author Contributions

Conceived and designed the experiments: RS. Performed the experiments: RS. Analyzed the data: RS. Contributed reagents/materials/analysis tools: RS. Wrote the paper: RS.

References

1. Berg OG, von Hippel PH (1987) Selection of DNA binding sites by regulatory proteins : Statistical-mechanical theory and application to operators and promoters. Journal of Molecular Biology 193: 723–743.
2. Stormo GD, Hartzell GW (1989) Identifying protein-binding sites from unaligned DNA fragments. Proc Natl Acad Sci U S A 86: 1183–1187.
3. Hertz GZ, Hartzell GW, Stormo GD (1990) Identification of consensus patterns in unaligned DNA sequences known to be functionally related. Comput Appl Biosci 6: 81–92.
4. Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: A Sequence Logo Generator. Genome Res 14: 1188–1190.
5. Man T, Stormo GD (2001) Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. Nucl Acids Res 29: 2471–2478.
6. Bulyk ML, Johnson PLF, Church GM (2002) Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. Nucl Acids Res 30: 1255–1261.
7. Maerkl SJ, Quake SR (2007) A systems approach to measuring the binding energy landscapes of transcription factors. Science 315: 233–237.
8. Benos PV, Bulyk ML, Stormo GD (2002) Additivity in protein-DNA interactions: how good an approximation is it? Nucl Acids Res 30: 4442–4451.
9. Djordjevic M, Sengupta AM, Shraiman BI (2003) A biophysical approach to transcription factor binding site discovery. Genome Research 13: 2381–2390.
10. van Nimwegen E, Zavolan M, Rajewsky N, Siggia ED (2002) Probabilistic clustering of sequences: Inferring new bacterial regulons by comparative genomics. Proceedings of the National Academy of Sciences of the United States of America 99: 7323–7328.
11. Barash Y, Elidan G, Friedman N, Kaplan T (2003) Modeling dependencies in Protein-DNA binding sites. In: Vingron M, Istrail S, Pevzner P, Waterman M, eds. Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology.
12. King OD, Roth FP (2003) A non-parametric model for transcription factor binding sites. Nucl Acids Res 31: e116.
13. Zhou Q, Liu JS (2004) Modeling within-motif dependence for transcription factor binding site predictions. Bioinformatics 20: 909–916.
14. Sharon E, Lubliner S, Segal E (2008) A feature-based approach to modeling protein-DNA interactions. PLoS Comput Biol 4: e1000154.
15. O'Flanagan RA, Paillard G, Lavery R, Sengupta AM (2005) Non-additivity in protein-DNA binding. Bioinformatics 21: 2254–2263.
16. Faiger H, Ivanchenko M, Haran TE (2007) Nearest-neighbor non-additivity versus long-range non-additivity in TATA-box structure and its implications for TBP-binding mechanism. Nucl Acids Res 35: 4409–4419.
17. Peng C, Buldyrev SV, Goldberger AL, Havlin S, Sciortino F, et al. (1992) Long-range correlations in nucleotide sequences. Nature 356: 168–170.
18. Segal E, Fondufe-Mittendorf Y, Chen L, Thastrom A, Field Y, et al. (2006) A genomic code for nucleosome positioning. Nature 442: 772–778.
19. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, et al. (2004) Transcriptional regulatory code of a eukaryotic genome. Nature 431: 99–104.
20. MacIsaac K, Wang T, Gordon DB, Gifford D, Stormo G, et al. (2006) An improved map of conserved regulatory sites for Saccharomyces cerevisiae. BMC Bioinformatics 7: 113.
21. Zhu J, Zhang MQ (1999) SCPD: a promoter database of the yeast Saccharomyces cerevisiae. Bioinformatics 15: 607–611.
22. Butler G, Thiele DJ (1991) ACE2, an activator of yeast metallothionein expression which is homologous to SWI5. Molecular and Cellular Biology 11: 476–485.
23. Koch C, Moll T, Neuberg M, Ahorn H, Nasmyth K (1993) A role for the transcription factors Mbp1 and Swi4 in progression from G1 to S phase. Science (New York, NY) 261: 1551–1557.
24. Bergman C, Carlson J, Celniker S (2005) Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, Drosophila melanogaster. Bioinformatics 21: 1747–1749.
25. Halfon MS, Gallo SM, Bergman CM (2008) REDfly 2.0: an integrated database of cis-regulatory modules and transcription factor binding sites in Drosophila. Nucleic Acids Res 36: 594–598.
26. Li X, MacArthur S, Bourgon R, Nix D, Pollard DA, et al. (2008) Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. PLoS Biol 6: e27.
27. Noyes MB, Meng X, Wakabayashi A, Sinha S, Brodsky MH, et al. (2008) A systematic characterization of factors that regulate Drosophila segmentation via a bacterial one-hybrid system. Nucl Acids Res 36: 2547–2560.
28. Sinha S, Liang Y, Siggia E (2006) Stubb: a program for discovery and analysis of cis-regulatory modules. Nucleic Acids Res 34: 555–559.
29. Siddharthan R, Siggia ED, van Nimwegen E (2005) PhyloGibbs: A Gibbs sampling motif finder that incorporates phylogeny. PLoS Comput Biol 1: e67.
30. Siddharthan R (2008) PhyloGibbs-MP: module prediction and discriminative motif-finding by Gibbs sampling. PLoS Comput Biol 4: e1000156.
31. Lusk RW, Eisen MB (2008) Use of an evolutionary model to provide evidence for a wide heterogeneity of required affinities between transcription factors and their binding sites in yeast. In: Pacific Symposium on Biocomputing, 2008. pp 489–500.
32. Durbin R, Eddy S, Krogh G, Mitchison G (1998) Biological Sequence Analysis Cambridge University Press.