Supporting Text

Additional Details of Methods, Experiments and Results

Parameter Estimation. We performed 50 random restarts for the quasi-Newton algorithm for parameter estimation and picked the solutions corresponding to the highest log-likelihood achieved at convergence. Figure 7 illustrates the parameters of the four resulting profile models pictorially. We have not attempted to interpret these numerical representations of each profile model in terms of their biological implications. But based on a rough inspection, it is not difficult to read off some interesting high-level biological characteristics. For example, for the *basic-domain* profile model, the transition probabilities between the four conserved nucleotidet-distribution prototypes [1] (the first four mixture components of the Dirichlet mixture) appear to be rather high (evident from the bright diagonal block at the upper left corner of the B matrix), as are the self-transition probabilities of all of the four nonconserved Dirichlet components (evident from the bright diagonal stripe at the lower right corner of the B matrix). The transition probabilities between the conserved and nonconserved Dirichlet components are relatively low (dark off-diagonal areas in B). Furthermore, it appears that the initial probability is high for the sixth Dirichlet component, a fairly nonconserved one. This suggests a general meta-sequence feature, which implies that motifs of the basic-domain family are likely to begin with a consecutive run of mostly non-conserved positions, followed by a consecutive stretch of mostly conserved positions, and possibly followed by another consecutive run of mostly nonconserved positions, reminiscent of the bell-shaped signature in Figure 2. Although it is possible to find many other similar high-level characteristics, some of which may even reveal previously unnoticed biological features (*e.g.*, characteristic position-specific multinomial distribution (PSMD) prototypes of motif families), in this article we refrain from such elaborations, but simply maintain that MotifPrototyper is a formal mathematical abstraction of the meta-sequence properties intrinsic to a motif profile represented by the training examples.

Motif Classification. We estimate the class label of a given set of aligned motif instances based on the posterior probability of each possible assignment of class membership to the motif alignment under test. Let z denote the family membership indicator, the posterior probability of $z = k$ is proportional to the magnitude of the conditional likelihood under the "k-th" MotifPrototyper multiplied by the prior probabilities of $z = k$:

$$p(z = k|\mathbf{A}) \quad \propto \quad p(z = k)p(\mathbf{A}|\{\alpha, \pi, B\}_k)$$

To examine the generalizability of MotifPrototyper to newly encountered motif patterns, we performed a 10-fold cross-validation (CV) test for motif classification, in which we learn the profile models from 90% of the training motif matrices and evaluate their classification performance on the remaining 10% of the motif matrices. We do so 10 times so that each motif pattern corresponding to a particular TF will be classified exactly once as a test case.

PWM Estimation and Motif Scoring. If the family membership is not known *a priori* for a

[1] Note that the parameter vector of a Dirichlet component can be regarded as a vector of pseudo-counts of the nucleotides. Thus, a Dirichlet parameter vector with a dominant element implies a conserved nt-distribution prototype, whereas a Dirichlet parameter vector without a dominant elements implies a heterogeneous, or non-conserved nt-distribution prototype.

given set of motif instances, we can assume that the PWM (to be estimated) admits a mixture of MotifPrototyper models. The posterior distribution of a PWM under such a mixture prior is slightly more complex:

$$\begin{aligned} p(\theta|\mathbf{A}, \{\alpha, \pi, B\}_{k=1}^{K}) &= \sum_k p(\theta|\mathbf{A}, \{\alpha, \pi, B\}_k, z = k) p(z = k|\mathbf{A}, \{\alpha, \pi, B\}_{k=1}^{K}) \\ &\propto \sum_k p(\theta|\mathbf{A}, \{\alpha, \pi, B\}_k) p(\mathbf{A}|\{\alpha, \pi, B\}_k) p(z = k), \end{aligned}$$

where z denotes the family membership indicator. A useful variant of this mixture model is to replace the mixture with the maximal-likelihood component:

$$p(\theta|\mathbf{A}, \{\alpha, \pi, B\}_{k=1}^{K}) \equiv p(\theta|\mathbf{A}, \{\alpha, \pi, B\}_{k^*}), \quad \text{where} \quad k^* = \arg\max_k \; p(\mathbf{A}|\{\alpha, \pi, B\}_k).$$

To evaluate the quality of the estimated PWM in terms of its discriminability of true motifs over background noises, we compare the likelihood of a true motif substring with the likelihoods of background substrings, all scored under the estimated PWM of the motif under test. To get an objective evaluation for this comparison, we performed the following experiment: (*i*) for a set of aligned instances of a motif, compute the Bayesian estimation of the PWM from 66% of the instances, and then use it to score (*i.e.*, compute the likelihood of) the remaining 34% of the instances in terms of their joint log likelihood; (*ii*) use the same PWM to score M sets of background strings, each having the same length and number of instances as the motif instances being scored in step *i*; (*iii*) compute the mean log likelihood odds between the motif and the background substrings (over M sets of randomly sampled background substrings). For each motif, we repeat this procedure three times so that each motif substring will be scored exactly once. The performance on each motif is summarized by the average log likelihood odds per motif instance. (Larger odds means that the background substrings are less likely to be mistakenly accepted as motif instances and, thence, a smaller false-positive rate).

Since the original aligned motif sequences corresponding to the count matrices used for MotifPrototyper training are not provided in TRANSFAC and are hard to retrieve from the original literature, we compiled an independent collection of aligned motif instances for 161 TFs in TRANSFAC, each of which has at least six binding sites whose sequence information is available (table 2). We simulated background substrings from a uniform and random model. This corresponds to examining the log likelihood odds under a motif model with respect to a uniform and random null hypothesis. Sampling of backgrounds substring from genuine genomic sequence as the null hypothesis was also done at a small scale (for some motifs) and yields largely the same results. But since the motifs we studied are from diverse genomic sources, a comprehensive evaluation in this manner is tedious and hence omitted.

The results of our evaluation are highlighted in Fig. 8. We compared four PWM estimation schemes: maximum likelihood (ML) estimation (*i.e.*, plain relative frequencies); Bayesian smoothing using a single symmetric Dirichlet prior; Bayesian estimation using a mixture of profile models, and Bayesian estimation using the ML profile model from the mixture. Depicted as the bars in Fig. 8 for reference is the Bayesian estimation using a single profile model corresponding to the original family label of each motif, an unrealistic scenario in *de novo* motif detection.

As evident from Fig. 8, in most cases, a mixture of profile models leads to significantly improved log likelihood odds compared with the standard ML estimation. In particular, in cases where only a small number of instances are available for estimation, a mixture of profile models still leads to a good estimation that generalizes well to new instances and results in high log likelihood odds, whereas the ML estimation does not generalize as well.

De novo **motif discovery.** We tested on 28 well represented yeast motifs from the SCPD database (http://cgsigma.cshl.org/jian/). Each motif has 5 to 32 recorded instances, all of which have been identified/verified by biological experiments and hence considered as "authentic." For each motif, we create a test dataset by planting each of the "authentic" instances of that motif at a random position in a 500-bp simulated background sequence (i.e., one motif per sequence). To further increase the difficulty of the motif detection task, we also insert a "decoy" signal, which is an artificial pattern resulting from randomly permuting the positions in the motif [2]. Since each sequence has only one true motif occurrence, prediction was made by finding the position with the maximal log likelihood ratio (for the substring that begins with that position) under the estimated motif PWM (obtained at the convergence point of a procedure that iterates between computing posterior distribution of motif locations based on current estimation of PWM, and computing the Bayesian estimation of the PWM based on current posterior distribution of motif locations), and under the background nucleotide-distribution (assumed to be the nucleotidet-frequencies estimated from the entire sequence). This scenario frees us from modeling the global distribution of motif occurrences, as needed for more complex sequences [compare the LOGOS model, ref. (1)], and therefore demonstrates the influence of different models for motif patterns on *de novo* detection.

We evaluate the performance based on *hit rate*, the ratio of correctly identified motif instances (within $\pm$ 3-bp offset with respect to the locations of the authentic instances) versus the total number of instances to be identified. To obtain robust estimation, for each motif we performed 40 experiments, each with a differently created test dataset (i.e., with different background sequences, motif and decoy locations, and decoy patterns), and we report the median of the hit rates over all experiments for each motif.

[1] Xing, E. P, Wu, W, Jordan, M. I, & Karp, R. M. (2004) *Journal of Bioinformatics and Computational Biology* **2(1)**, 127–154.

[2] By permutation we mean that the same permuted order is applied to all the instances of a motif so that the multinomial distribution of each position is not changed but their order is changed.