
**Vocabulon: a dictionary model approach for reconstruction and localization
of transcription factor binding sites.**

Chiara Sabatti¹, Lars Rohlin², Kenneth Lange^{1,3}, and James C. Liao²

¹ Departments of Human Genetics and Statistics, UCLA, Los Angeles CA 90095-7088,

² Department of Chemical Engineering, UCLA, Los Angeles CA 90095,

³ Department of Biomathematics, UCLA, Los Angeles, CA 90095

Revised July 23rd, 2004

Running head A dictionary model for transcription factors

Keywords Motif search; transcription regulation; dictionary model; gene regulation; gene expression arrays.

Corresponding author Chiara Sabatti

Department of Human Genetics

UCLA School of Medicine

695 Charles E. Young Drive South

Los Angeles, California 90095-7088 (USA)

FAX: (310) 794-5446

Phone: (310) 794-9567

e-mail: csabatti@mednet.ucla.edu

Abstract

We describe the first implementation of dictionary-style models to the study of transcription factors binding sites in an entire genome. Vocabulon's unique feature is that it can both reconstruct binding sites characterized by unknown motifs and impute locations of known binding sites in long sequences by simultaneous search. On one hand, the dictionary model specifies a probability for the entire sequence taking simultaneously into account all the possible binding sites. This greatly reduces the number of false positives. On the other hand, the possibility of refining motif description, as an increasing number of binding sites are identified, augments the sensitivity of the method. We illustrate these properties with examples in *Escherichia Coli*. The results of gene expression arrays are used both to guide the search and corroborate it.

1 Introduction

The identification of binding sites for regulatory proteins in the up-stream region of genes is an important step towards understanding transcription regulation. In recent years, novel experimental techniques, such as gene expression arrays, and the availability of entire genome sequences have permitted more detailed investigations in this area. Traditionally, the reconstruction of the profile of a binding site and the localization of all its occurrences in a sequence have been treated as separate problems. The first is tackled using a small group of sequences, known or suspected to contain the binding site, but with neither position nor pattern known. One successful approach to such reconstruction problems is based on a probabilistic model that represents a sequence as a concatenation of background and motif stochastic words. Maximum likelihood (or maximum a-posteriori) estimates for the parameters of this model are obtained with EM (or Gibbs-sampler) algorithms (Lawrence and Reilly, 1990; Lawrence *et al.*, 1993).

The second problem involves localizing motifs with known patterns, in one or multiple sequences of variable length. Possible locations are identified on the base of scoring functions that

highlight the similarity of the motif with sequence segments. Cutoff values for such similarity scores are hard to determine, and ad hoc solutions or estimates from a training set are often adopted (Quandt *et al.*, 1995; Robison *et al.*, 1998). Typically these techniques are used to scan one sequence of interest against a data-base of known binding sites. When they are used to investigate the presence of binding sites in the entire genome, they lead to a large number of possible locations.

While there are historical and practical reasons to tackle separately the two described tasks, the current post-genomic era calls for a different approach. Consider the problem, addressed in Robison *et al.* (1998), of identifying all of the binding sites of the known regulatory proteins in the genome of *Escherichia Coli*. While formally similar to blasting a small sequence of interest against a data-base of known regulatory proteins, there are substantial differences in these genomewide searches. On one hand, as one scans through the genome for binding sites of LexA—to take one example—and finds a substantial number of them, it seems appropriate to use the information in the identified locations to update the current pattern description (Lawrence and Reilly, 1990; Lawrence *et al.*, 1993). On the other hand, given that the output may involve a large collection of potential sites, the determination of a similarity score cut-off should be done systematically, relying on probabilistic evaluations. Furthermore, the investigator does not have the option of validating the results of this bioinformatics search with experiments, as it happens in the case of one short sequence leading to few putative sites. To increase the specificity of the search, it is important to scan the genome sequence for binding sites of a collection of proteins simultaneously, rather than one binding site at the time.

To address these issues, one needs a precise probability model for the entire genome sequence that can be used to evaluate specific a-posteriori probabilities for the presence of a binding site at any given location. The parameters of this model should be estimated from data. Moreover, given the scale of the problem, it is also important to guarantee a rapid computation. In an attempt to address this need, we introduce here the Vocabulon model. The next section gives a description of the probability model we employ, its differences from others in the literature, and its current

implementation. We then present the results of multiple investigations on *E. Coli* sequence.

2 Methods

The first suggestion of a dictionary-based probabilistic model for DNA sequence is due to Bussemaker *et al.* (2000). These authors propose modeling the genome as a concatenation of words selected independently from a dictionary with word-specific probabilities. In this framework, a word of length one represents meaningless background, or a “space filler”, while longer words identify functional motifs. There are three points worth noting about this model. First, the hypothesis of independence across consecutive words cannot account for some interactions between regulatory protein described in the literature (Jennings and Beacham, 1993). Because the independence assumption provides a very significant computational speed-up, it is difficult to relax. Secondly, an important advantage of the dictionary model is that it provides a conditional probability framework for deciding whether a binding site occurs at a given sequence location. A third point about this dictionary model is that it relies on deterministic words—that is words that admit only one spelling. Although this simplification allows the authors to reconstruct a DNA dictionary starting from sequence alone, it represents a serious limitation if we want to use the model for binding site reconstruction.

Both the limitations and the strengths of the model proposed by Bussemaker *et al.* (2000) have prompted further investigations. For example, Sabatti and Lange (2002) describe in detail an algorithm that allows exact, rather than approximate, computation of sequence probabilities. More importantly, Gupta and Liu (2003) and Sabatti and Lange (2002) extend the model to encompass motifs or fuzzy words with variable spellings. (These two extensions were developed independently and entail different computational algorithms: we propose a deterministic one, Gupta and Liu an MCMC one.) The extended model retains the macroscopic features of the original dictionary model, but also adds, at a finer scale, microscopic features used in motif finding (see, for

example Lawrence and Reilly, 1990). We propose calling our extended model and algorithm Vocabulon, after a generic name for a French society game based on word guessing and recognition. A description of the main characteristics of the model follows.

The building blocks of a sequence are words, intended as irreducible semantic units, or, in the genetic context, motifs. Each word may admit more than one spelling. Thus, in English, “theater” and “theatre” represent the same word. Two different words may share a spelling. In our model, a word w always has the same number of letters $|w|$. Hence, alternative spellings such as “night” and “nite,” with different number of letters, are disallowed. The letters of a word are independently sampled from different multinomial distributions. This is known as product multinomial sampling. It is convenient to group words according to their lengths and to impose a maximum word length k_{\max} on our dictionary. Note that the total number of words, and their lengths, are taken as known for the purpose of this implementation. We will discuss at the end of the methods section how to relax this assumption. In summary, the Vocabulon model requires a static dictionary with a list of alternative spellings and probability distributions determining which words and spellings are selected. The parameters of the model can, then, be grouped as follows:

1. The probability of choosing a word of length k is q_k . Here k ranges from 1 to k_{\max} , and $\sum_{k=1}^{k_{\max}} q_k = 1$. If there are no words of length k , then $q_k = 0$.
2. Conditional on choosing a word of length k , a particular word w with $|w| = k$ is selected with probability r_w . Hence, $\sum_{|w|=k} r_w = 1$.
3. The letters of a word w follow a product multinomial distribution with success probabilities

$$\ell_{wi} = (\ell_{wiA}, \ell_{wiC}, \ell_{wiG}, \ell_{wiT})$$

for the letters A, C, T, and G at position i of w .

A randomly chosen word of length k , then, exhibits the spelling $s = (s_1, \dots, s_k)$ with probability

$$p(s) = \sum_{|w|=k} r_w \prod_{i=1}^k \ell_{wis_i}. \quad (1)$$

To accommodate missing data, we represent missing letters by question marks and introduce the additional letter probability $\ell_{wi?} = 1$ for each word w and position i within it. A random sequence S is constructed from left to right by concatenating random words, with each word and each spelling selected independently. In the Vocabulon model, we assume that the stretch of DNA observed is a fragment of text from an infinitely long sequence. (A detailed description of the implications of this assumption and its difference from the dictionary model proposed by Bussemaker *et al.* (2000) can be found in Sabatti and Lange (2002).) When we observe a DNA sequence, we do not have information on word boundaries. We will call the portion of a sequence between two consecutive word boundaries a “segment” and the set of word boundaries dividing a sequence an “ordered partition” of the indices of the sequence. If we use the symbol π to indicate a such partition $\pi = (\pi_1, \dots, \pi_{|\pi|})$, then π_i is the set of indices corresponding to the i th segment $s[\pi_i]$ of s . With the above notation and assumptions, the likelihood of a sequence is

$$\mathcal{L}(s) = \Pr(S = s) = \frac{1}{\sum_{i=1}^{k_{\max}} i q_i} \sum_{\pi} \prod_{i=1}^{|\pi|} q_{|\pi_i|} p(s[\pi_i]),$$

where $p(s[\pi_i])$ is determined by equation (1). Sabatti and Lange (2002) derive in detail this likelihood. They also give algorithms for likelihood computation that resemble Baum’s forward and backward algorithms from the theory of hidden Markov chains (Baum, 1972; Devijver, 1985).

For estimation purposes, a Bayesian framework is attractive because it allows the incorporation of prior information on experimentally identified binding sites. It is convenient to impose independent Dirichlet priors on q , r , and ℓ , since they are conjugate priors for multinomial densities. In the case of q , for example, this implies choosing a prior distribution of the form

$$\frac{\Gamma(\sum_{k=1}^{k_{\max}} \alpha_k)}{\prod_{k=1}^{k_{\max}} \Gamma(\alpha_k)} \prod_{k=1}^k q_k^{\alpha_k - 1}.$$

In selecting the prior parameters $\alpha_1, \dots, \alpha_{k_{\max}}$, it is helpful to imagine a prior experiment and interpret $\alpha_k - 1$ as the number of successes of type k in that experiment. The sum $\sum_{k=1}^{k_{\max}} \alpha_k - k_{\max}$ gives the number of trials in the prior experiment, and hence determines the strength of the prior. Note that the special case where all $\alpha_k = 1$ yields a posterior density that coincides with the likelihood. Information on binding sites contained in various databases can be used to define the prior counts of the appropriate Dirichlet distribution. For example, the values of α_k can be set to represent the relative frequency of a motif in the database, and corresponding parameters for the spelling probabilities can be chosen to reflect the frequency with which a particular base pair is observed at a position in an experimentally identified motif. Maximum a posteriori estimates are obtained with a MM gradient algorithm (Lange *et al.*, 2000), described in Sabatti and Lange (2002).

Once the maximization procedure is completed and we have obtained parameter estimates for the model, it is possible to evaluate the probability that a specific word occupies any location in the sequence under analysis. The probability that a particular word w fills the sequence segment extending from index i to index j is

$$\rho_{ij}(w) = \frac{f_{i-1} q_{j-i+1} r_w \prod_{k=1}^{j-i+1} \ell_{w k s_{i+k-1}} b_{j+1}}{\mathcal{L}(s)}, \quad (2)$$

where f_{i-1} is the joint probability of the letters of the sequence up to position $i - 1$ and the event that a new words starts at position i and b_{j+1} is the probability of the sequence from position $j + 1$ onward, conditional on a word ending at position j . Both the forward probabilities f_{i-1} and the backward probabilities b_{j+1} are calculated as part of the likelihood evaluation. Note that $\rho_{ij}(w)$ depends not only on the similarity of the sequence $s[i : j]$ to the motif under study, but also on the composition of all the rest of the sequence. In particular, because the numerator depends on the partitioning of the entire sequence, the probability of occurrence of a longer word is not necessarily always smaller than the probability of occurrence of a word that constitutes a subset of it. This is a desirable property in that increases the chances of longer words being identified.

Computing $\rho_{ij}(w)$ for every position i and word w allows us to infer the location of binding sites in the genome. In particular, we will impute a binding site for word w at every location i whenever $\rho_{ij}(w) > T$, for some threshold T , specified by the user.

The choice of T is not arbitrary. For example, T is compared with the probability that a word occupies a specific portion of the sequence: thus, there is an immediate interpretation available for any selected threshold. From a decision theoretic perspective, the cost incurred when one misses a binding site (C_{FN}) and the cost incurred in falsely identifying (C_{FP}) one determine an explicit value for the cut off. When such costs are equal, the value of 0.5 is the optimal cut off; in general, $T = C_{FP}/(C_{FN} + C_{FP})$. This formula implies that as the ratio C_{FP}/C_{FN} increases, the optimal threshold value increases. Generally speaking, in our problem, false positives are less harmful than false negatives. Identified binding sites are considered as tentative and often their validation is pursued with other methods, so that false positives can be weeded out. On the other hand, there is no “safety net” against false negatives: once a binding site is lost, it falls off the experimental radar screen. Because of these considerations, it may be wise to choose a T value smaller than 0.5. Indeed, our experience suggests that when analyzing an entire genome, for multiple motifs, the values of $\rho_{ij}(w)$ across positions i and words w are generally very close to zero, so that a $\rho_{ij}(w)$ even as low as 0.2 is uncommon enough to provide the researcher with a strong signal that a binding site may be present. In the opposite situation, where a few relatively short sequences enriched for a single binding site are analyzed, the posterior probabilities of occurrence will tend to be higher, so that one may want to select a value of $T > 0.5$.

Finally, we would like to briefly discuss the selection of word lengths and the size of the dictionary. For clarity, consider first a dictionary consisting only of background and one unknown word of unknown length. To run Vocabulon in such a case, we arbitrarily assume a given word length (20 bp, for example). Once the word is identified, it is easy to extend or trim it on both ends by comparing the distribution of the four bases in each of the positions in question with their distribution in the background sequence. This is standard procedure for algorithms based on EM

and Gibbs sampling. Although one could also re-run either algorithm assuming different lengths and compare the results, the method just described above is effective and computationally efficient.

For dictionaries containing multiple words, two situations should be considered: a limited extension of the case above, where relatively short sequences are enriched with two or three motifs rather than only one, and the case of a genomewide investigation, aiming at the analysis of tens of motifs. In the first case, it is possible to reconstruct the entire dictionary without prior information. One starts by identifying a first word, adds a second word in a second round of estimation, and so forth. This strategy does not represent a substantial departure from what is achievable with a single motif (Lawrence and Reilly, 1990; Lawrence *et al.*, 1993). Stopping rules for small dictionaries are outlined in Gupta and Liu (2003). We are more interested in the radically different situation where we analyze an entire genome and expect sparse occurrences of tens of motifs. Reconstructing a dictionary with many words and no prior information is likely to be an ill-posed problem when the words have variable spelling. For this reason, we consider here a static dictionary defined on the basis of available experimental information. This less ambitious approach is consistent with our practical interest in the analysis of *E. Coli*. To date, two attempts to build dynamic dictionaries have been documented in the literature: Bussemaker *et al.* (2000) and Gupta and Liu (2003). Bussemaker *et al.* consider only words that are perfectly conserved. Although Gupta and Liu describes a theoretical dictionary building methodology, they only apply it to an easy example with a single word.

Our Fortran95 implementation of Vocabulon requires two input files; one containing the sequences to be searched in FASTA format and another listing the words of the dictionary and prior information on their alternative spellings. The Vocabulon program can be run in two modes. The default option estimates the value of all the parameters q, r, ℓ . The no-spelling option fixes the matrix of multinomial spelling probabilities and estimates only the q and r parameters. The output of the program includes a list of all the locations in the analyzed sequences where a motif was detected ($\rho_{ij}(w) > T$), the probability of the motif at that location, the expected count of each motif

for each sequence, and the estimated values of the parameters. We have currently implemented *Vocabulon* on the Linux cluster of 2.5 GHz computers in the UCLA Human Genetics Department; the code has not been parallelized. To obtain an executable of *Vocabulon*, please contact Chiara Sabatti.

3 Results

The purpose of this section is to illustrate the performance of *Vocabulon* and its comparative advantages. *Vocabulon* performs two tasks, which we want to test: the reconstruction of binding sites profiles and the localization of their occurrences in the sequences under analysis. In setting up a series of test-examples, the following points should be kept in mind: (a) the problems need to be identifiable, in the sense that there must be sufficient information in the data to discriminate among alternative solutions; (b) the available experimental information to be used as a benchmark has collection biases; and (c) comparison with competing algorithms and models depends on the specific task at hand. With regard to point (a), it is reasonable to attempt the reconstruction of a single binding site, in the absence of prior information, when analyzing a group of relatively short sequences enriched for that motif. However, the same task becomes impossible when a large collection of motifs, sparsely represented in long sequences, has to be identified. Point (b) refers to asymmetry between false negative and false positives. The false negative rate of a computational model can be evaluated by comparison of predicted sites with the experimentally verified sites. In contrast, sites identified as false positives may turn out to be true positives after experimental techniques are perfected. Because the problem of identifying all the binding sites in the genome, for any regulatory protein, is an open one, it is impossible to evaluate exactly the error rates of a procedure. Compared to computing methods, *Vocabulon* is unique in its ability to scan a genome for the location of binding sites and, at the same time, refine their description. This does not rule out the possibility that a different strategy could better reconstruct the binding sites in a genome.

Statistical methods are always tied to the data actually used, and to the experimental sophistication with which it is gathered.

On the base of the considerations described above, we have considered examples that illustrate the ability of *Vocabulon* to: a) reconstruct a binding site in absence of prior information from enriched sequences; b) identify all the locations of such a binding site in the genome; and c) identify occurrences in the genome of a collection of binding sites on which prior information is available.

3.1 Reconstructing an unknown binding site from enriched sequences: Crp

As a standard test, we considered first the classical benchmark problem of reconstructing Crp binding sites from a collection of 18 microbial sequences. The specific sequences were kindly provided by the authors of the original paper (Lawrence *et al.*, 1993). No prior information on the DNA pattern corresponding to the Crp binding site was given. We assumed the length of the motif to be 22 bp. The purpose of this example is to illustrate that, indeed, our algorithm for binding site reconstruction works appropriately. Since our model is extremely close to the model of Lawrence *et al.* (1993) for a dictionary with a single binding site, we do not expect any substantial difference in performance—modulo the fact that the code implementing the results in Lawrence *et al.* (1993) is based on MCMC and has been optimized over a decade, while our code uses MM and is recent. Indeed, if anything, we would expect that popular algorithms inspired by this model would perform better than *Vocabulon* in this particular test. The advantages of our approach will be apparent in sections 3.2 and mainly 3.3, where we deal with problems that algorithms such as that of Lawrence *et al.* (1993) do not handle.

In view of our earlier discussion, we used 0.80 as a cutoff for the posterior probability of a motif. We identified 19 of the 24 previously noted locations. We also found an additional 23 putative sites. Our reconstructed spelling matrix corresponds well with the one known to characterize

Crp. Results did not change substantially when the cutoff value of 0.5 was used.

This first test case demonstrated that Vocabulon—like other related motif finding algorithms—can converge to a local mode. In particular, Vocabulon can be trapped in a non optimal local pattern that is a shifted form of the optimal one, as described in Lawrence *et al.* (1993). We have been able to overcome this difficulty by using multiple runs in cases such as Crp and imposing prior information in larger problems. In the future, we plan to augment our algorithm with “shift moves” like the ones described in previous literature,. The analysis required 10 minutes, using 50 different random starting values.

3.2 Reconstructing an unknown binding site and finding all its location in a genome: *lexA*

To consider a problem that is closer to experimental reality, and to illustrate the specific features of Vocabulon, we turned to the study of the binding site for LexA. Again, we assumed no prior information on the site, and we first reconstructed it, using a set of enriched sequences, and, subsequently, we identified all of its locations in the genome. To guide our selection of enriched sequences, we used a set of published micro-array experiments on *E. Coli*. Courcelle *et al.* (2001) investigated the dynamic effects of UV irradiation of *E. Coli* using microarray technology for 4290 genes. Two different time-courses were collected, one where they exposed wild type *E. Coli* to UV, and one where they exposed a strain where LexA had been knocked out (*lexA*-) to the same treatment. During each time course, they collected data at 5, 10, 20, 40 and 60 min. It is well known that exposure to UV activates the LexA regulon. We analyzed gene expression values, with a very conservative procedure, to identify those genes regulated by LexA. In a first pass, we isolated all genes that were either up regulated or down regulated at least 2 fold at all time points in the wild type and showed no changes in the *lexA*- strain. A total of 87 genes fitted this criterion. The selected genes were clustered with an agglomerative hierarchical method, based on correlation and

complete linkage. The genes from the highest conserved sub-cluster were selected: *sulA* (b0958), *dinI* (b1061), *umuD* (b1183), *ruvA* (b1861), *recN* (b2616) and *recA* (b2699). These are indeed genes that are known to be regulated by LexA and whose LexA binding sites have been experimentally determined. Using the genome sequence of *E. Coli* published in Blattner *et al.* (1997), we extracted 600 base pairs prior to and 100 after the start codon for these genes. We ran our algorithm on these sequences, hypothesizing a dictionary with a background word of length one and a word of length 20. We found a total of 8 sites for LexA, corresponding perfectly to the ones identified experimentally in these sequences. The reconstructed LexA pattern can be seen in Figure 1. The identified motif is clearly palindromic, even though no prior information was used to this effect. Six sites are almost perfectly conserved. This may be a consequence of our stringent criteria isolating those sequences that have the most specific and efficient binding sites.

Once this first description of the site was obtained, we used our algorithm to search for other possible binding sites for LexA in *E. Coli*. For this purpose, we focused on 700 bp segments in the promoter regions of 3277 genes in *E. Coli*. These genes were selected, out of the 4290 total genes, to take into account operon structure. Indeed, many genes clusters in *E. Coli* are transcribed as an operon, in the sense that the genes are adjacent, read in the same direction, and regulated by a common promoter region upstream of the first gene. Although the entire operon structure in *E. Coli* is unknown, we were able to use the predictions described in Sabatti *et al.* (2002), with a cut-off posterior probability of being in an operon equal to 0.9. To check the effectiveness of our algorithm, we had at our disposal 19 known binding sites for LexA, eleven more than the 8 ones reconstructed in the previous experiment. We ran our algorithm using as prior information the reconstructed pattern and a) the no-spelling option, and b) the standard procedure with a strong contribution from the prior. Procedure a) is similar to a scoring function search, in that the pattern describing the motif remains untouched. Note, however, that we still estimated the parameters q and r , to gain a specific probabilistic description of the sequence. Procedure a) led to the identification of 12 of the known binding sites, and a total of 25 imputed ones. Procedure

b), where the pattern describing the motif was allowed to adapt, led to the identification of 15 of the known binding sites, and a total of 35 imputed ones. Being able to refine the motif description resulted in increased sensitivity with more true locations identified. The accompanying increase in the total number of detected sites somewhat tempers our enthusiasm for this result. It is interesting that while updating spelling probabilities introduces more degrees of freedom, it is not obvious that these should result in identification of a higher number of true binding sites, since no information on the identity of these is given to the algorithm.

So far, we have based our evaluation of the *Vocabulon* algorithm on comparisons with 19 known binding sites, but such information is typically not available when searching for novel sites. It is also possible to refine the results of a motif search by comparison with gene expression array data—often available. We considered the second time-point of the UV experiment described before (the first time point reflects too little changes and all time-points after 2 behave in similar manner). The histograms of the expression values for the genes, divided in groups with respect to their LexA status, is given in Figure 2. Clearly, there is a shift in expression values for the genes that are not imputed to have a binding site for LexA versus those that are imputed to have one. Furthermore, these shifts are in the direction of the expression values of the genes that are known to be regulated by LexA. This provides evidence, that while there probably are some false positives among the imputed sites, the majority of the imputed sites are true positives.

3.3 Localization of a dictionary of motifs

The experiments discussed so far have shown that the *Vocabulon* method can reconstruct the pattern of an unknown binding site and predict its occurrences in a genome. It even has an edge over competing procedures (such as MatInspector) that predict the localization of a defined binding site in the genome, because the pattern description can be refined as more occurrences are encountered. However, the real goal of *Vocabulon* is its ability to deal with a large number of binding sites and

identify all of their locations in a genome.

The identification of all binding sites in a genome for a group of regulatory proteins is an open scientific problem. Its resolution would represent a considerable advance in the understanding of regulatory networks. The work of Robison *et al.* (1998), based on similarity scores, was the first attempt to solve this problem in *E. Coli*. The results were an important step forward in the field. While we still do not have experimental validation or disproof of the imputed sites, it is clear that their total number is too high and many false positive are to be expected. To give a sense of the extent of the problem, we briefly summarize the findings of Robison *et al.* (1998). Two possible estimates of binding sites are provided—one corresponding to a very stringent cut-off (leading on average to missing 50% of the real sites) and one corresponding to the cut-off suggested by the authors. Often, the order of magnitude of the counts obtained with the stringent cut-off (in the hundreds) are more reasonable than the ones obtained with the suggested cut-off (in the thousands).

While it is generally true that sensitivity is more important than specificity in searching for binding sites, increasing specificity is critical in any attempt to reconstruct a genomewide regulatory network. This was indeed one of our goals in developing *Vocabulon*. Recently Djordjevic *et al.* (2003), have attacked the specificity problem by modeling the sequence-specific binding energy of transcription factors. We turn now to the results of our study, which is more empirical. To evaluate our performance, we will study how well we reconstruct the set of known binding sites that provide current prior information; compare our results with the ones in Robison *et al.* (1998); and assess their consistency with an array experiment.

To compile our dictionary and define prior information, we modified the original list of binding sites in Robison *et al.* (1998) by (a) adding some proteins subsequently studied in greater detail, (b) eliminating words that were by definition overlapping with others in the dictionary, and (c) eliminating proteins whose binding site has very low information content. According to criteria (a), we included in the dictionary *fliA* and *creB* (Avison *et al.*, 2001; and Park *et al.*, 2001). An example of the application of criterion (b) is the case of *phoB* and *phoB3*. The latter binding site

consists in three overlapping instances of the first. Since our model does not admit overlapping words, such a definition is clearly inconsistent. In such cases, we included in the dictionary only the smaller, more modular word. According to criterion (c), we eliminated from our dictionary binding sites such as *ihf*, *lrp* and *hns*, which appear to be unrecognizable on the basis of sequence pattern alone. A total of 17 words were deleted because their binding sites occur four or fewer times in the database. *RpoD* was also deleted due to low information content. *RpoS* and *rpoH* were divided up into two different words each in order to better represent their binding sites. Having made these decisions, we compiled the dictionary with 41 binding sites plus a background letter as reported in table 1.

Initially, we tested the performance of Vocabulon on the set of 233 sequences, each 700 bp long, containing the experimentally identified binding sites used in the definition of the motif priors. The performance of our algorithm is illustrated in Table 1 and in Figures 3, 5, and 4. Using a cutoff of 0.5 for the posterior probability of a word, we reconstruct 80% of the known binding sites. A summary of how these percentages by sequence and motif is given in Figure 3. The proportion of recovered sites rises above 90% if the cutoff is 0.2. Figure 4 illustrates the tradeoff between sensitivity and “specificity” of Vocabulon’s predictions as the posterior probability cutoff varies. In the remainder of this paper, we will consider a site recovered if its posterior probability equals or exceeds 0.2, unless otherwise specified. As we have explained, false negatives are worse than false positives.

A substantial portion (1/3) of the missed motifs can be explained on the basis of undetected motifs overlapping detected motifs. A typical example is given in Figure 5. As described previously, the Vocabulon constructs a DNA sequence by concatenating non-overlapping words. Since Vocabulon outputs posterior probabilities position by position, it is possible that overlapping motifs may be detected. Indeed, roughly 1/2 of all overlaps are detected. However, when two words both offer a plausible explanation for the same portion of sequence, their posterior probabilities are reduced below what we might expect in the absence of overlap. This sometimes translates into

one of the motifs going undetected. Which motif is detected depends on the length of the motifs, and their degree of conservation.

Having explored the performance of Vocabulon on the above test set, we analyzed 3277 upstream regions for the *E. Coli* genes described in the previous section. This required roughly one hour of computing time. Again, we evaluated how many of the binding sites used to define the prior information were correctly recovered. Notice that now, the 233 sequences that contain these are a small fraction of the total sequences analyzed and hence the signal to noise ratio is substantially lower. Although only 7.5% of the sequences have known binding sites, the percentage of recovered sites is 70% with a 0.5 posterior probability cutoff (see Table 1), suggesting that our false negative rate is still acceptable.

To evaluate our false positive rate, we compared our results to the results in Robison *et al.*, (1998). The counts in Figure 6 reflect the more stringent criterion of Robison *et al.*, while both counts are represented in Figure 7. Our estimated counts appear closer to the stringent ones by Robison *et al.*, with, however, a much lower false negative rate (see above).

Perhaps the most interesting validation of our method involves comparison with micro-array experiments. Recently, there has been substantial interest in interpreting gene expression results in terms of the presence of regulatory-protein binding motifs in the up-stream sequence of the studied genes. The main goal of researchers has been the identification of novel regulatory motifs. Two approaches have been successful so far. The first analyzes changes in expression levels, identifies those few genes that exhibit similar expression responses, and searches for shared motifs in their up-stream regions with standard algorithms like the one described in Lawrence and Reilly (1990). The second approach creates a long collection of putative motifs by searching for small deterministic words that appear with sizable frequency in the promoter regions of the genes studied. Linear regression of the results of an array experiment against the collection of putative motifs is used to weed out spurious motifs (see Bussemaker *et al.*, 2001; Colon *et al.*, 2003; and Keles *et al.*, 2002).

As outlined, both of these strategies see the identification of motifs as the ultimate goal and the analysis of gene expression array data as a tool. We here propose a different perspective. We consider the interpretation of the results of a micro-array experiment as the ultimate goal and use the available motif information as supporting evidence. To clarify our viewpoint, consider what can be achieved in *E. Coli* based on prior information on binding site positions. Relying on the motif library reviewed in Robison *et al.* (1998), for example, one can determine, for at least some of the genes in an array experiment, the presence of binding sites for the regulatory proteins under consideration. Using binding sites presence/absence scores as regressors, one can gain insight into which regulatory proteins are activated in a gene expression experiment. Indeed, if we apply this strategy to the analysis of the second time-point of the UV experiment described earlier, we can analyze 233 genes and LexA is clearly the most significant regressor. Its p-value, on the order of 10^{-16} , is far lower than the next most significant p-value, on the order 10^{-3} . Had we not known that UV activates the LexA regulon, we would have learn this fact. Liao *et al.* (2003) pursue this strategy to its logical conclusion. (Again, note that we use the second experiment because the first is done at time zero, so that there is no substantial effect to be detected. The remaining experiments lead to similar results.)

The described analysis has the merit of leading to the identification of the appropriate regulatory protein in the specific situation under analysis. However, it is limited in scope, as it is based on a significantly reduced fraction of the genes. It does not allow to make inference about the expression values of the remaining genes and it would be blind to the effects of regulatory proteins that operate mainly on gene outside this limited group of 233. The Vocabulon model allows us to overcome this impasse: exploiting the expected locations of putative binding sites across the genome, we can significantly increases the amount of information derived from array experiments. We can more effectively learn which regulatory proteins are involved, and, at the same time, which genes are affected by such changes. To illustrate these possibilities, we analyzed again the expression values of the second time point of the UV experiment series, using as regressors for

each gene the expected number of binding sites for each regulatory protein in the dictionary. This time the expression values of 3277 genes were involved in the analysis. Again, *lexA* is the most significant explanatory variable, with a p-value of 10^{-16} , followed by *fis* with a p-value of 10^{-5} . This illustrates two points: (a) our predictions of binding sites are accurate enough to enable us to correctly identify the regulatory protein mainly involved in the biological process; (b) a tool such as Vocabulon that provides genomewide predictions of binding sites can be potentially very useful in this approach to the analysis of microarray experiments.

3.4 Conclusion

Bussemaker *et al.* (2000) proposed the use of a language parsing algorithm to study DNA sequence. Extensions of their dictionary model and a description of methods for estimating its parameters can be found in Gupta and Liu (2003) and Sabatti and Lange (2002). The present paper goes beyond the benchmark examples previously treated and describes the results of the first genomewide investigation of regulatory protein binding sites conducted with a dictionary model. The results are encouraging. Vocabulon simultaneously refines prior information on binding sites, while rapidly scanning a genome. The expected number of binding sites per regulatory protein corresponds to scientific expectations. The output of Vocabulon model can also be effectively used to understand gene expression array experiments.

While we believe that this represents a considerable advance, there are still many unmet challenges in motif recognition. It is our view that a “solution” to the problem is not likely to come from a single approach, but from the convergence of a variety of methods and data. Vocabulon offers the possibility of integrating the genomewide dictionary view and the flexible motif description found in Lawrence *et al.*, (1993). However, the challenge presented by the analysis of genome sequences of higher organisms, such as humans, is likely to require additional developments. For example, it is clear that cross-species comparison will strengthen motif signals (McCue *et al.*,

2001). Moreover, a careful physical model of the bio-physical interactions that define binding may ultimately be necessary to capture binding sites with very low information content (Djordjevic *et al.*, 2003). Some of the innovations coming from these different approaches can be incorporated in Vocabulon. The dictionary model itself can be improved by including interactions between binding sites. Even in absence of these refinements, the results reported in this manuscript underscore the important role that dictionary models will have in the identification of transcription factor binding sites.

Acknowledgments

C. Sabatti acknowledges support from NSF (grant DMS0239427) and from NASA/Ames (grant NCC2-1364). K. Lange was supported in part by USPHS grants GM53275 and MH59490.

Avison, M.B., Horton, R.E., Walsh, T.R. and Bennett, P.M. (2001) Escherichia coli CreBC is a global regulator of gene expression that responds to growth in minimal media, *J. Biol. Chem.*, **29**, 26955–61.

Baum, L.E. (1972) “An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes,” *Inequalities*, **3**, 1–8.

Blattner, F.R., Plunkett, G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B., Shao, Y (1997) The complete genome sequence of Escherichia coli K-12, *Science*, **277**, 1453–74.

Bussemaker, H.J., Li, and Siggia, E.D. (2000) Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis, *PNAS*, **97**, 10096–10100.

- Bussemaker,H.J., Li,H. and Siggia,E.D. (2001) Regulatory element detection using correlation with expression, *Nature Genetics*, **27**, 167–171.
- Colon,E., Liu,X., Lieb,J., and Liu,J.S. (2003) Integrating regulatory motif discovery and genome-wide expression analysis, *PNAS*, **100**, 3339–3344.
- Courcelle,J., Khodursky,A., Peter,B., Brown,P.O. and Hanawalt,P.C. (2001) Comparative gene expression profiles following UV exposure in wild-type and SOS-deficient Escherichia coli, *Genetics*, **158**, 41–64.
- Devijver,P.A. (1985) Baum’s forward-backward algorithm revisited, *Pattern Recognition Letters*, **3**, 369–373.
- Djordjevic,M., Sengupta,A.M. and Shraiman,B.I. (2003) A biophysical Approach to transcription factor binding site discovery, *Genome Research*, **13**, 2381–2390.
- Gupta,M. and Liu,J.S. (2003) Discovery of conserved sequence patterns using a stochastic dictionary model,” *Journal of the American Statistical Association*, **98**, 55–66.
- Jennings,M., Beacham,I.R. (1993) Co-dependent positive regulation of the ansB promoter of Escherichia coli by CRP and the FNR protein: a molecular analysis, *Mol. Microbiol.*, **9**, 155–64.
- Keles,M., van der Laan,M. and Eisen,M. (2002) Identification of regulatory elements using a feature selection method, *Bioinformatics*, **18**, 1167–1175.
- Lange,K., Hunter,D.R. and Yang,I. (2000) Optimization transfer using surrogate objective functions (with discussion), *Journal of Computational and Graphical Statistics*, **9**, 1–59.
- Lawrence,C.E. and Reilly,A.A. (1990) An expectation maximization (em) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences, *Proteins*, **7**, 41–51, 1990.

- Lawrence,C.E, Altschul,S.F., Bogouski,M.S., Liu,J.S., Neuwald,A.F. and Wooten,J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment, *Science*, **262**, 208–214.
- Liao,J., Boscolo,R., Yang,Y., Tran,L., Sabatti,C. and Roychowdhury,V. (2003) Network component analysis: Reconstruction of regulatory signals in biological systems, *PNAS*, **100**, 15522–15527.
- McCue,L.A., Thompson,W., Carmack,C.S., Ryan,M.P., Liu,J.S., Derbyshire,V. and Lawrence,C.E. (2001) Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes, *Nucleic Acids Research*, **29**, 774–782.
- Park,K., Choi,S., Ko,M. and Park,C. (2001) Novel (F-dependent genes of Escherichia coli found using a specified promoter consensus, *FEMS Microbiology Letters*, **202**, 243–250.
- Quandt,K., Frech,K., Karas,H., Wingender,E. and Werner,T. (1995) MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data, *Nucleic Acids Res.* **23**, 4878–4884.
- Robison,K., McGuire,A.M. and Church,G.M. (1998) A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K12 genome, *Journal of Molecular Biology*, **284**, 241–254.
- Sabatti,C. and Lange,K. (2002) Genomewide motif identification using a dictionary model, *IEEE Proceedings*, **90**, 1803–1810.
- Sabatti,C., Rohlin,L., Oh,M. and Liao,J. (2002) Co-expression pattern from DNA microarray experiments as a tool for operon prediction, *Nucleic Acids Research*, **30**, 2886–2893.
- Schneider,T.D. and Stephens,R.M. (1990) Sequence Logos: A New Way to Display Consensus Sequences, *Nucleic Acids Research*, **18**, 6097–6100.

Word	233 sequences			3277 sequences		
	recovered	missed	imputed	recovered	missed	imputed
araC	6	0	6	6	0	9
arcA	8	5	28	6	7	60
argR	15	2	24	15	2	108
cpxR	11	1	29	7	5	99
creB	8	0	9	8	0	19
crp	36	13	131	34	15	610
cspA	4	0	4	3	1	12
cytR	2	3	7	1	4	55
dnaA	7	1	41	6	2	96
fadR	7	0	8	6	1	21
fis	8	7	36	8	7	200
fliA	12	0	14	12	0	25
fnr	12	0	14	11	1	43
fruR	12	0	18	11	1	43
fur	8	1	18	8	1	69
galR	7	0	10	5	2	10
gcvA	4	0	4	4	0	6
glpR	7	6	20	6	7	71
hipB	2	2	2	0	4	2
lexA	19	0	24	19	0	46
malT	4	6	6	0	10	0
metJ	6	3	8	5	4	13
metR	5	3	10	4	4	44
nagC	6	0	9	6	0	22
narL	7	3	9	4	6	18
narP	8	0	4	8	0	7
ntrC	4	1	4	4	1	6
ompR	5	4	28	4	1	6
oxyR	4	0	4	4	0	4
phoB	10	2	12	9	3	35
purR	21	1	25	17	5	47
rpoH2	6	1	6	6	1	9
rpoH3	8	0	8	8	0	13
rpoN	6	1	11	6	1	22
rpoS17	5	10	9	1	14	4
rpoS18	4	3	8	3	4	5
soxS	11	6	22	9	8	61
torR	3	1	5	3	1	14
trpR	4	0	4	4	0	6
tus	5	0	5	5	0	5
tyrR	13	4	19	10	7	54
Total	340	90	663	296	134	2231

Table 1: Recovered, missed, and imputed binding sites for the 41 regulatory proteins in the dictionary, when Vocabulon is run on a set of 233 sequences, each containing at least one experimentally identified binding site; and when Vocabulon is run on 3277 sequences, representing all the regulatory regions in *E. Coli*. A binding site is identified in positions where its posterior probability is larger than 0.5.

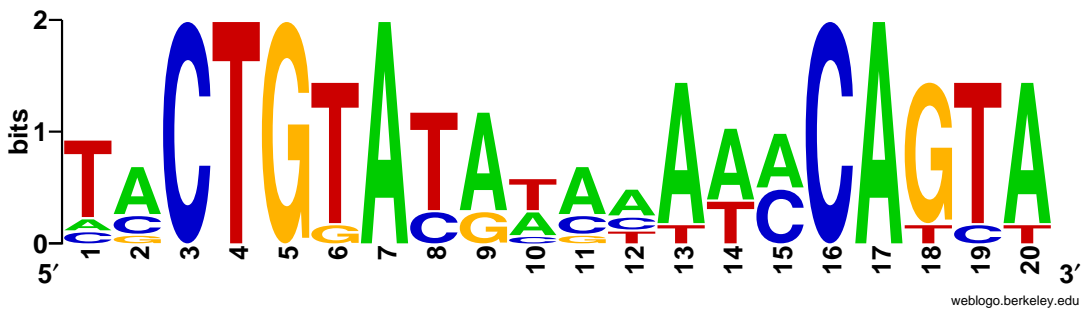


Figure 1: Profile of the binding site for LexA as reconstructed by Vocabulon starting from 6 *E. Coli* sequences. The graphics software used is available on the server <http://weblogo.berkeley.edu/> and is based on the idea of a sequence logo described in Schneider and Stephens (1990).

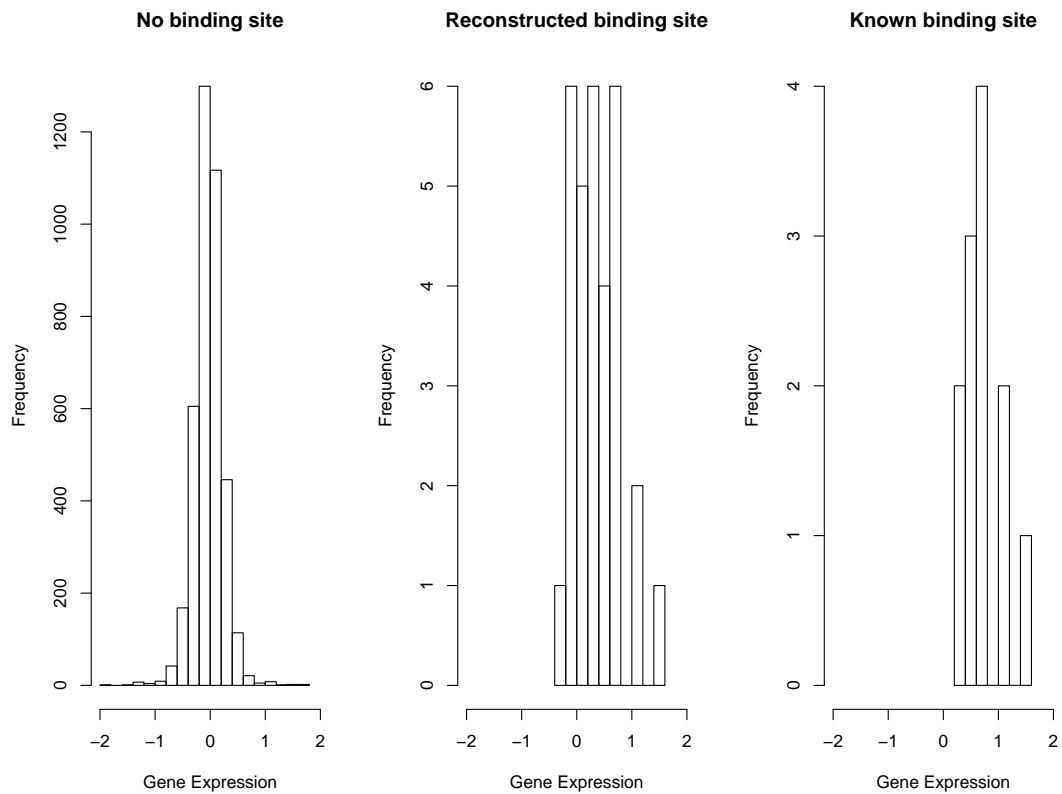


Figure 2: Histograms of the gene expression values at the second timepoint of the UV experiment. From left to right: genes that do not have a binding site for *lexA* according to the Vocabulon reconstruction; genes that do have a binding site for *LexA* according to Vocabulon; and known genes with a binding site for *LexA* from the literature.

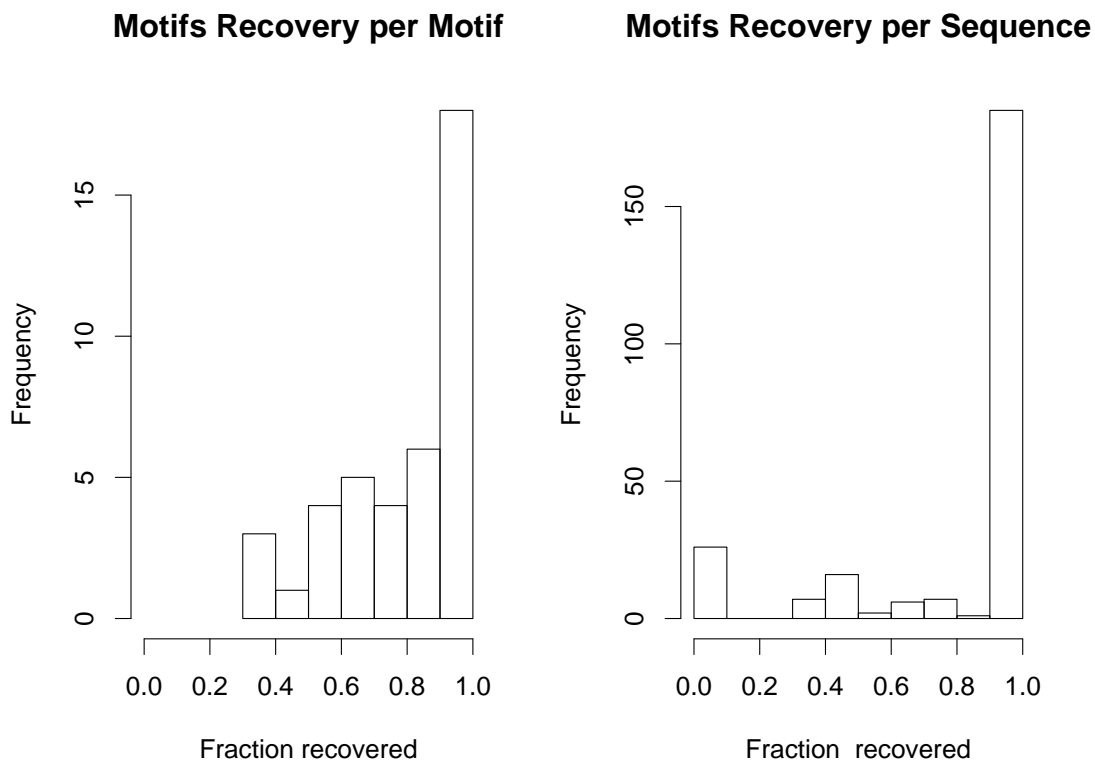


Figure 3: On the left, histogram of the percentage of recovered sites, for each of the 42 motifs. On the right, histogram of the percentage of recovered sites for each of the 233 sequences. A binding site is identified in positions where its posterior probability is larger than 0.5.

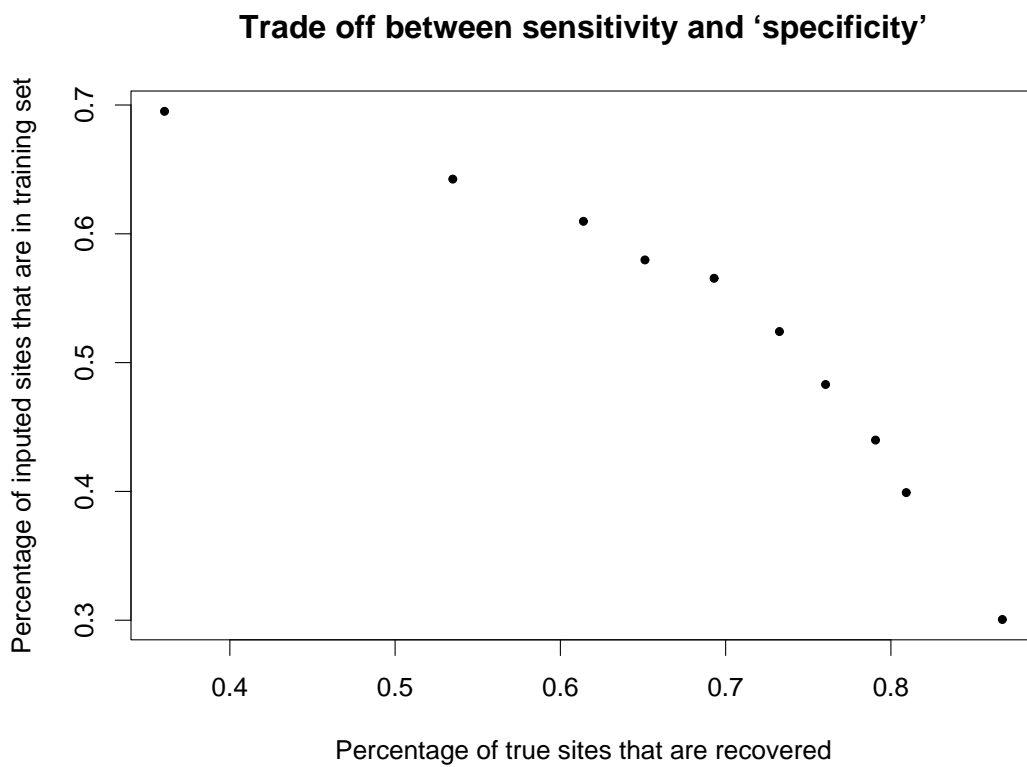


Figure 4: Sensitivity (x axis) and “specificity” (y axis) of the motif reconstruction on the 233 sequence set as a function of the cut-off value for the posterior probability.

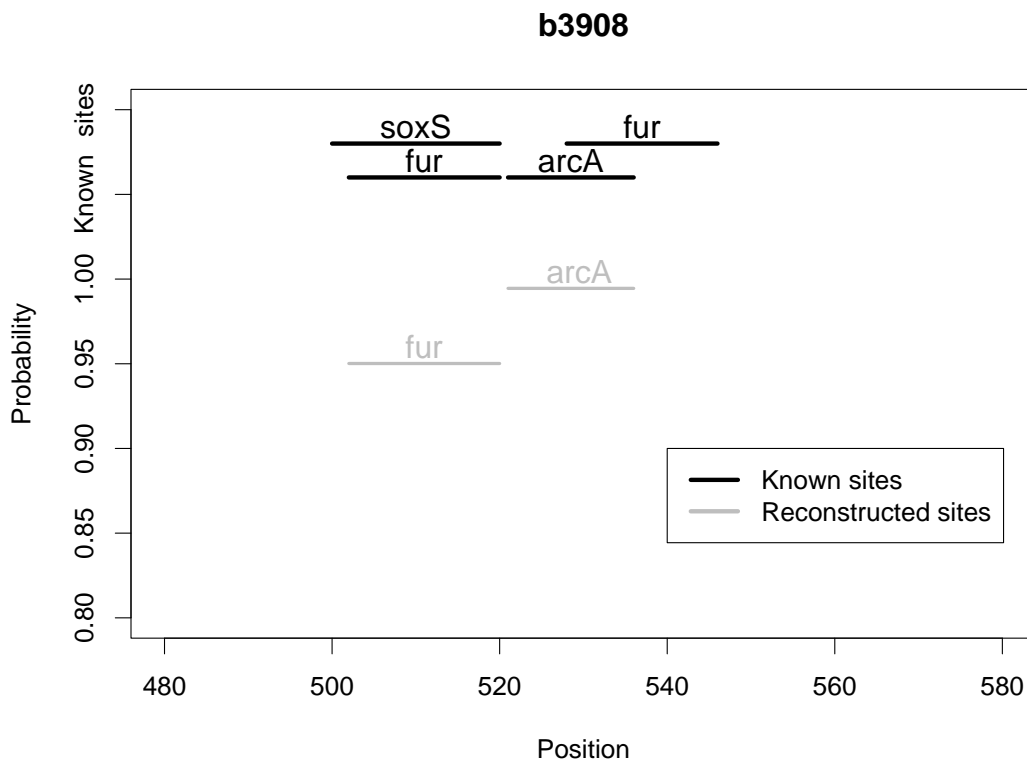


Figure 5: Graphical illustration of part of the Vocabulon output for the up-stream sequence of gene b3908. On the x axis is the position in base pairs, on the y axis is the posterior probability for binding sites. The values above one are used to displayed known binding sites. This is an example of overlapping motifs, where only one of the motifs is reconstructed—*fur* in one case, and *arcA* in the other.

Vocabulon's and Robison et al.'s stringent counts

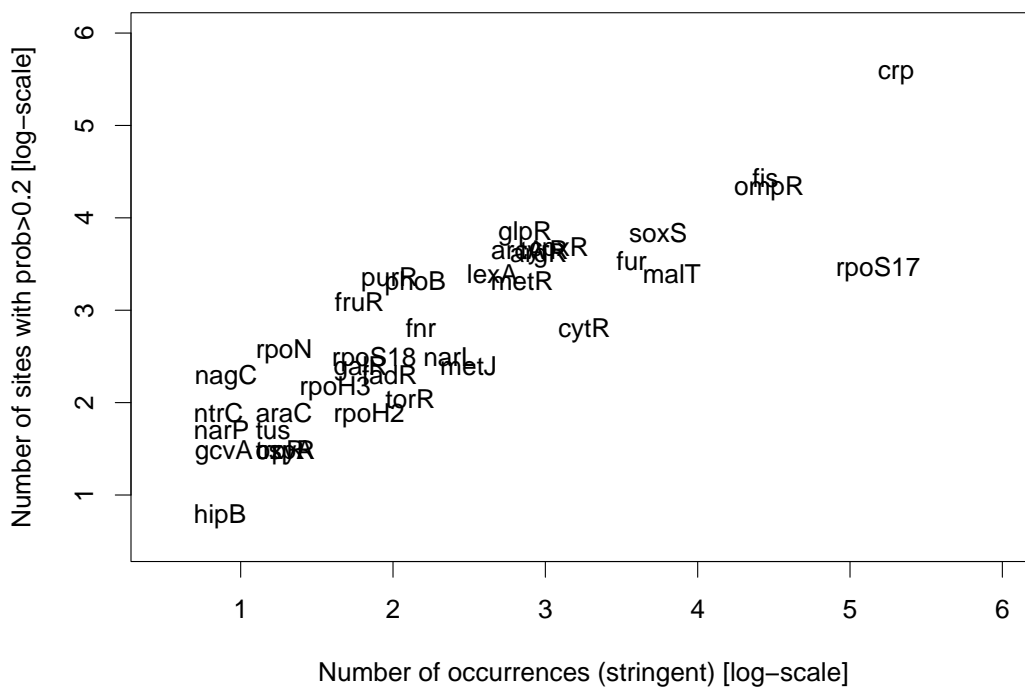


Figure 6: Scatter plot of the total number of predicted sites for the 41 motifs in our dictionary via Vocabulon (y axis) and via the Robison *et al.* (1998) strict criteria (x axis). Note that numbers are expressed on a \log_{10} scale.

Vocabulon and Robison et al.'s Counts

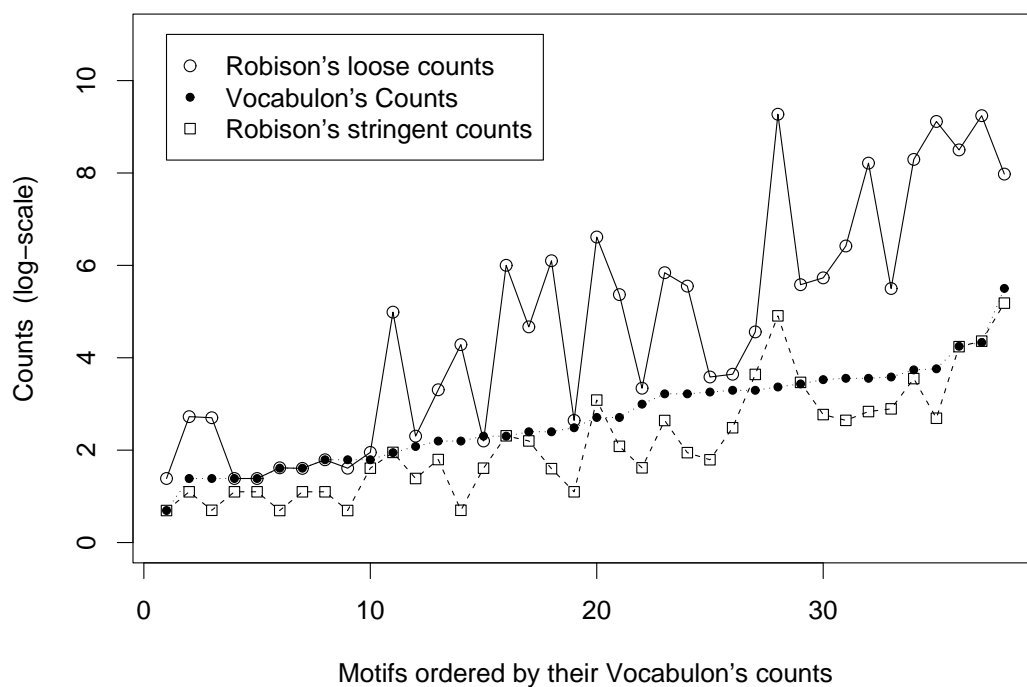


Figure 7: Comparison of predicted counts per motif. Motifs are ordered with respect to their expected counts computed by Vocabulon. We report the total expected counts according to Vocabulon, and the loose and stringent predictions of Robison *et al.* (1998), with symbols as in the legend.