Title
A Vocabulon Study of E.Coli Regulatory Sites with Feedback to Expression Array Analysis

Authors
Chiara Sabatti
Lars Rohlin
Kenneth Lange
et al.

Publication Date
2011-10-25

A Vocabulon study of E.Coli regulatory sites with feedback to expression array analysis

Chiara Sabatti*, Lars Rohlin†, Kenneth Lange*, and James Liao†

* Department of Human Genetics and Statistics, UCLA, Los Angeles CA 90095-7088

and † Department of Chemical Engineering, UCLA, Los Angeles CA 90095

UCLA Statistic Department Preprint

369

October 2003

Running head E. Coli regulatory protein binding sites

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

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

1 Introduction

The identification of binding sites for regulatory proteins in the up-stream region of genes is an important ingredient towards the understanding of transcription regulation. In recent years, novel experimental techniques, as gene expression arrays, and the availability of entire genome sequences have opened the possibility for more detailed investigations in this domain. Traditionally, the reconstruction of the profile of a binding site and the localization of all its occurrences in a sequence are 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 or pattern known. One successful approach to such reconstruction problem is based on a probabilistic model of the sequence, represented as concatenation of background and motif stochastic words. Maximum likelihood or maximum a-posteriori estimates are obtained with EM or Gibbs-sampler algorithms [13, 14].

The second problem is approached considering one or multiple sequences of variable length; the pattern characterizing the motif is assumed known. Possible locations are identified on the base of scoring functions that highlight the similarity of the motif with the sequence portions. Cut off values for such similarity scores are hard to determine: ad hoc solutions or estimations on a training set are often adopted [17, 18]. Typically these techniques are used to scan one sequence of interest against a data-base of known binding sites. While there are historical and practical reasons to consider these two problems as separate, the current post-genomic era, where we are confronted with large abundance of sequence, calls for a different approach. Consider the problem, tackled in [18], of identifying all the the binding sites of the known regulatory proteins in the genome of E. 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 genome-wide search. On the 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 one should use the information in the identified locations to update the current pattern description. On the other hand, given that the output is not

going to include a small number of sites, that can be further investigated, but a large collection of them, the assessment of significance cut-off should be based on proper probabilistic statements. To address these issues, one would need a probability model for the entire genome sequence, that can lead to evaluation of specific a-posteriori probabilities of appearance of a binding site in any given location, and whose parameters can be estimated on the base of data. At the same time, given the scale of the problem, the model should be suitable for rapid computation. In an attempt to address such need we introduce here the Vocabulon model.

Section 2 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. Given that genome-wide information on the location of binding sites is not available, we used results of gene expression array experiments to corroborate our results, arguing in favor of a novel perspective in array analysis.

2 Methods

The first suggestion of a dictionary-oriented probabilistic model for DNA sequence is due to Bussemaker et al. [4]. These authors propose conceiving the genome as a concatenation of words selected from a dictionary independently from each other, and with word-specific probabilities. In this framework, a word of length one represents meaningless background, or a "space filler", while longer words identify functional sites. There are three points worth noting on the characteristics of this model. Firstly, the hypothesis of independence across consecutive words is certainly inadequate to account for some specific regulatory proteins interactions described in the literature [10]. However, it provides a very significant speed-up from a computational stand point, that is difficult to do with out. A second important advantage of Bussemaker's dictionary [4] is that its probability model for the entire DNA sequence provides a framework where the decision on the presence of a binding site in a given location can be based on well defined conditional probabilities. Unfortu-

nately, [4] contains slightly imprecise i calculations, so that the original algorithm would not lead to exact evaluations of the probability of interest. A third crucial aspect of this dictionary model is that it relies on deterministic words—that is words that admit only one spelling. This simplification allowed the authors to attempt to reconstruct the DNA dictionary starting from no other data then a sequence, but represents a serious limitation if we want to use such models for binding site reconstruction.

The limitations/strengths of the original dictionary model [4] have prompted further investigations. In one direction, we [20] describe in detail a dictionary-style model that, overcoming some of the computational difficulties in [4], allows exact computation of conditional probabilities. In an other domain, while deterministic words are the only one considered by Bussemaker et al., their same notion of DNA as concatenation of independently selected semantic units can be extended to encompass motifs or fuzzy words with variable spellings. Such extension can be found originally in the theoretical papers by Sabatti and Lange [20], andGupta and Liu [9] (These two extensions were developed independently and their computational algorithm differ: [20] uses a deterministic one, while [9] propose a MCMC approach) We hence obtain the model at the core of *Vocabulon*, with the same macroscopic features as Bussemaker's dictionary, but, at a finer scale, similar to the one used in motif finding analysis [13]. Vocabulon is a generic name for a french society game based on word guessing and recognition, which we though fit well the nature of our problem. Following is a description of the main characteristics of the model.

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. 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

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

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

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

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

To be robust to the presence of 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 w. 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 Bussemarker et al. [4] can be found in [20]). When we observe a DNA sequence, we do not have information on where the boundaries between words lie. 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 sequence. We use the symbol π to indicate a such partition

$\pi = (\pi_1, \ldots, \pi_{|\pi|})$, with π_i the set of indices corresponding to one word. With $s[\pi_i])$ we indicate the portion of the sequence corresponding to the indexes in π_i. With the above notation and assumptions, the likelihood of a sequence is

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

In Sabatti and Lange (2002) we discuss in detail the definition of such likelihood. We also give algorithms for likelihood computation that resemble Baum's forward and backward algorithms from the theory of hidden Markov chains [2, 8].

For estimation purposes, we adopt a Bayesian framework, attractive because it allows the incorporation of prior information on experimentally identified binding sites. We use independent priors on q, r, and ℓ. It is convenient to choose Dirichlet distributions, as they are conjugate priors for multinomial densities. In the case of q, for example, this implies choosing a prior distribution of the following 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 $\alpha_1, \ldots, \alpha_{k_{\max}}$, 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. It should be clear, at this point, how information on binding sites contained in various databases can be used to define the prior counts of the appropriate Dirichelet distribution. Maximum a posteriori estimates are obtained with a M-M gradient algorithm [12] described in [20].

The current implementation of Vocabulon is in Fortran 95. It requires two input files: the first contains the sequences to be searched in FASTA format; the second lists the dictionary structure (list of words with given length) and prior information on the spelling. The algorithm can be run in two modes: the default one estimates the value of all the parameters q, r, ℓ. The no-spelling option

fixes the matrix of multinomial probabilities. The output of the program includes a list of all the locations in the analyzed sequences where a motif was detected (posterior probability higher than a given threshold); the expected counts of each motifs for each sequence; and the estimated values of the parameters.

3 Results

3.1 The Crp binding site

For uniformity with the rest of the literature, we started considering the classical benchmark case of the reconstruction of Crp binding site from a collection of 18 microbial sequences. The specific sequences used are the same as in [14] and were kindly provided by these authors. No prior information on the DNA pattern corresponding to the Crp binding site was given. We used 0.80 as a cut-off for the posterior probability of motif. We identified 19 locations that are also presented by [14]. We did not identify the 5 remaining reconstructed by those authors and we individuated an additional 23 putative sites. The reconstructed spelling matrix corresponds well with the one known to characterize Crp.

This first test case was the occasion to note that—like other related motif finding algorithm—Vocabulon is subject to the problem of local modes. In particular, we noticed that Vocabulon can be trapped in non optimal local patterns that are a shifted form of the optimal one, as described in [14]. We have been able to overcome this difficulty by using multiple runs in cases as Crp and thank to the prior information in the larger problems. In the future, we plan to augment our algorithm with “shift moves” like the ones described in [14].

3.2 The lexA binding site

Both to consider a problem that is closer to the reality faced by investigators and to appreciate the specific features of our algorithm, we turned to the study of the binding site for LexA. We considered a set of published microarray experiments on E. Coli: Courcelle et al. [7] investigated the dynamic effects of UV irradiation of E. coli using microarray technology for 4290 genes. They collected two different time-courses, one where they exposed wild type to UV and one where they exposed a lexA- strain to the same treatment. In each time course they collected 5 time points, 5, 10, 20, 40 and 60 min. It is well known that exposure to UV should activate the LexA regulon. We analyzed the gene expression values with a very conservative procedure to identify a set of genes that may be regulated by LexA. In a first pass, we isolated all genes that where either up regulated or down regulated 2 fold or more at all time points in the wild type and in the lexA- strain shown neither up or down regulation. 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 of E. Coli as in [3] we extracted 600 base pairs prior to and 100 after the start codon for these genes. We run our algorithm on these sequences, hypothesizing a dictionary with a background word 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 palindromic (even if no prior information was used to this effect). There are 6 sites almost perfectly conserved. This may be due to the selection procedure for the input sequences: our stringent criteria may have led us to isolate those sequences that have the most exclusive and efficient binding sites. Once this first description of the site was obtained, we used our algorithm to search for all other possible binding

sites for the same regulatory protein in E. Coli. For this purpose, we compiled a list of the 700 bp in the promoter regions of 3277 genes in E. Coli. These were selected, out of the 4290 total genes, to take into account operon structure. Indeed, a considerable number of genes in E. Coli are transcribed as an operon: that is, they are adjacent, in the same direction of transcription, and all regulated by the promoter region up-stream the first one. When searching for binding sites of regulatory proteins, then, one can eliminate the up-stream regions of genes that are in an operon, but not in the first position. Unfortunately, the entire operon structure in E. Coli is not known, but we used the predictions described in [21], with a cut-off of 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 run 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. The procedure a) is the most similar to scoring function searches, in that the pattern describing the motif remains untouched. Notice, however, that we are still estimating the parameters q and r, and obtaining 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. Clearly, being able to refine the motif description, as more locations were identified, represented a benefit for the algorithm. As we are searching for a weak signal through a large number of sequences, however, it would not make sense to discount much the prior information—that is why we used the strong prior.

So far, we have based our evaluation of the algorithm on the comparison with the 19 known binding sites, but this information in typically not available when searching for novel sites. One instrument that is, instead, quite generally available to refine the results of a motif search is the comparison with gene expression array data. We considered the second time-point of the UV

experiment described before. 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 between the genes that are estimated not to have a binding site for LexA and those that appear to have one, which offers a generic validation of our results. More specifically, one can look at the expression value for each single gene and compare it with the posterior estimated probability for a binding site in its upstream region. Such comparison is carried out in Figure 3 with a scatter plot of probabilities of a binding site versus Log10 of the p-value for a null hypothesis of no expression change in the second time point of the UV series of microarray experiments. Such plots can be used to determine adaptively a cut-off for significance of the binding site. For example, the results presented so far are based on a 0.5 cut off. However, in light of Figure 3 one might decide that 0.8 is a more appropriate cut off, (as starting from such value we record the first gene expression changes which are significantly different from zero). From figure 3, where points corresponding to experimentally verified binding sites are in red, we can also gather that, while some of the imputed sites appear spurious, there are at least 6 that have expression values similar to the known LexA regulated genes, so that their prediction receives some corroboration.

3.3 A dictionary of motifs

The experiments illustrated so far have shown how our algorithm can reconstruct the pattern of an unknown binding site and estimate all its occurrences in a genome—with a comparative advantage over other procedures, due to the fact that the pattern description can be refined as more occurrences are encountered.

Another substantial strength of our method is in its ability to deal with a large number of binding sites—we will explore now this feature. When dealing with a large number of sequences and a large dictionary of words, with variable spellings, one has to pay attention to identifiability issues. Unless some form of constraint is introduced, for example, it is very plausible that two

formally different words in the dictionary end up describing the same real-life motif. Similarly, to select the appropriate size of the dictionary one would need to define appropriate complexity penalties. We have not thoroughly investigated this problem. In the E. Coli case study presented in what follows, we opted for the use of a pre-defined dictionary, with a fix number of words of known length and with strong prior information on their spelling. This makes it possible to monitor the identity of the reconstructed binding sites, and still provides a scientifically challenging problem, as testified by the work of Robinson et al. [18].

To compile the dictionary and define prior information, we referred initially to the similar study by [18]. We then modified the original list of binding sites according to the following criteria: (1) we added some proteins that have been studied in greater detail since [18]; (2) we eliminated words that were by definition overlapping with others in the dictionary; (3) we eliminated proteins whose binding site has very low information content. According to criteria (1), we included in the dictionary fliA and creB [16, 1]. An example of (2) is the case of phoB and phoB3. The latter consisting in three overlapping instances of the first. Since our model does not admit overlapping words, such definition is clearly inconsistent. In such cases, we included in the dictionary only the smaller, modular, word. According to (3), we eliminated from our dictionary binding sites such as ihf, lrp and hns, which has been proven not to be effectively recognizable on the base of sequence pattern alone. A total of 17 words were deleted because their binding sites only occur less than 4 times in the prior. RpoD was also delete due to low information in the whole word. RpoS and rpoH were divided up into two different words each in order to better represent the binding site. We compiled then a dictionary of 41 binding sites and a background word reported in table 3.

Initially, we tested the performance of our algorithm on the set of 233 sequences, each 700 bp long, that contain the experimentally identified binding sites used in the definition of the priors on motifs. The performance of our algorithm is illustrated in table 3 and in figures 4, 5, 6, and 7. Using a cut-off of 0.5 for the posterior probability of a word, we reconstruct 80% of the known binding sites. A summary of how this overall percentage breaks down by sequence and motif is given in

Figure 4. The proportion of recovered sites raises above 90% if the cut-off is 0.2, as illustrated in Figure 6, which provides specificity and sensitivity as a function of the posterior probability cutoff. In the following, when analyzing the performance of the algorithm in detail, we will consider recovered any site that has posterior probability greater than 0.2—which is considerably greater than the average value of zero.

A substantial portion (1/3) of the missed motifs can be explained considering that the undetected motif is overlapping with a detected one—an example is given in Figure 5. As described in the method section, the Vocabulon model for DNA sequence consists in the concatenation of non-overlapping words. Since the output of our algorithm is not a single segmentation of the DNA sequence, but gives, for each position, the posterior probability that a given word appears there, it is possible that overlapping motifs may be detected; indeed, this happens in roughly 1/2 of the instances. However, given the fact that two words offer a plausible explanation for the same portion of sequence, both their posterior probabilities are reduced with respect to the value that they might had in absence of overlap. This translates sometimes in one of the motifs going undetected. Which motif is detected depends on the length of the motifs and their degree of conservation.

The influence of the information content of a motif on its chance of being recovered is illustrated in Figure 7. For each position in the motif, following [22] we define as information the quantity $2 - \sum_i p_i \log p_i$, with $i = A, C, G, T$. As an index of degree of conservation of the entire motif we consider the total information across position (known as Rsequence). From Figure 7 it seams to appear that the cumulative information offers a reasonable predictor of the chance for a motif to be detected.

After having explored the performance of our algorithm on the test set above, we analyzed 3277 up-streams regions for E. Coli genes (see description in the previous section) with the goal of identifying all the binding sites for the regulatory proteins in our dictionary. This is a real scientific challenge, rather than a test problem, and we cannot compare our results to the "correct" answer. In order to assess, however, how reasonable our predictions might be, we (1) reassessed

the percentage of recovered motifs in the test set; (2) compared our counts with the ones obtained for the same problem in [18]; (3) checked our results against gene expression data.

The fact that the signal to noise ratio decreases substantially in this problem (only 7.5% of the sequences have known binding sites), did not affect dramatically our performance on the test set: the percentage of recovered sites is 70% with a 0.5 posterior probability cut-off (see table 2). Figures 8 and 9 illustrate the correspondence between the estimated number of binding sites obtained with our algorithm and in Robinson et al. (1989). Those authors actually propose two criteria to identify possible binding sites, one more conservative than the other. The counts in Figure 8 are according the most stringent criteria. Our estimated counts appear to be somewhat in the middle of the two obtained by Robinson et al. (1989), while following the same general trend. Perhaps the most interesting validation for our method relies in its comparison with the results from micro-array experiments, described in the following section.

4 Interpreting gene expression results in the light of motif imputation

Since the advent of gene expression arrays, there has been a substantial interest in analyzing their results in connection with the presence of regulatory-protein binding motifs in the up-stream sequence of the studied genes (see [19] for a review). The main goal of researchers has been the identification of novel regulatory motifs. Two approaches have been successful so far. The first starts from the analysis of expression data, identifies few genes that exhibit a very similar expression pattern and searches for shared motifs in their up-stream region with algorithm like the one described in [13]. The second approach starts by creating a long collection of putative motifs by searching for small deterministic words that appear with sizeable frequency in the promoter regions of the genes studied. Linear regression of the results of one array experiment against the collection

of putative motifs is used to weed out the spurious ones (see [5, 11, 6]). As outlined, both 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 our final goal, and suggest using the available motif information as supporting evidence. To clarify our viewpoint, consider what can be achieved in E. Coli, using literature information on binding site positions. Relying on the information reviewed in [18], for example, one can determine, for each gene in the array experiment, if it has a binding site for each regulatory protein under consideration. Using these binding sites presence/absence scores as regressors, one can easily obtain information on which regulatory proteins are activated in the experiment analyzed with gene expression arrays. Indeed, if we apply this strategy to the analysis of the second time-point of the UV experiment described before, we obtain that LexA is the most significant regressor (p-value in the order of 10^{-16}, with the next most significant p-value of the order 10^{-3}). Had we not known that UV activates the LexA regulon, we would have learn it. A more sophisticated and powerful method of analysis for array data that exploits the same principle as above is given in [15]. However, using only information available through the literature and based on experimentally verified binding sites, has a clear limitation: the number of genes whose expression can be explained is fairly modest. The Vocabulon model offers an important contribution at this level: by providing expected locations of binding sites across the genome, it significantly increases the amount of information that can generally be extracted 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 the expected number of binding sites for each regulatory protein in the dictionary and each gene. The results are in table 3, and in Figure 10. Again, lexA appears as the most significant explanatory variable. The plot of change in expression vs expected number of binding sites for LexA, Figure 10, illustrates the additional information that Vocabulon gathered. Points corresponding to experimentally

verified binding sites for LexA are in red. There are quite a number of genes that are not known in the literature to have a binding site for LexA that both have an affinity for LexA in their promoter region sequence, and an expression value comparable to the one of known LexA-regulated genes. Considering at the same time these two pieces of information allows us to place them more confidently in the LexA regulon. Notice that an analysis based on expression values only may not lead to the identification of such genes, whose expression changes are not very significant.

4.1 Conclusion

Bussemaker et al. [4] proposed the use of language parsing algorithm to study DNA sequence. Extensions of such model and description of algorithmic procedures for estimating it can be found in [20, 9], where the discussion is, however, limited to simple tests or benchmark examples. This paper describes the challenges encountered and the results of the first genome-wide investigation of regulatory protein binding sites conducted with such models. The results are encouraging. The feature of our algorithm that allows refining of the prior information on binding sites, while rapidly scanning a genome for its presence, proved useful. The total expected counts of binding sites per regulatory proteins correspond to scientific expectations. It also appears how the outcome of the Vocabulon model can be effectively used in a novel analysis of gene expression arrays.

Acknowledgments

C. Sabatti thankfully aknowleges support from NSF (grant DMS0239427) and from NASA/Ames (grant NCC2-1364).

References

[1] Avison MB, Horton RE, Walsh TR, Bennett PM. (2001) Escherichia coli CreBC is a global regulator of gene expression that responds to growth in minimal media. J Biol Chem. 2001 Jul 20;276(29):26955-61.

[2] L. E. Baum, "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes," *Inequalities*, vol. 3, pp. 1–8, 1972.

[3] Blattner FR, Plunkett G 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y. The complete genome sequence of Escherichia coli K-12. Science. 1997 Sep 5;277(5331):1453-74.

[4] H. J. Bussemaker, H. Li, and E. D. Siggia, "Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis," *PNAS*, vol. 97, pp. 10096–10100, 2000.

[5] Bussemaker, Li, Siggia (2001) Regulatory element detection using correlation with expression, Nature Genetics 27:167-171.

[6] E Conlon, X. Liu, J Lieb, and J Liu Integrating regulatory motif discovery and genome-wide expression analysis PNAS 2003 100: 3339-3344.

[7] 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.

[8] P. A. Devijver, "Baum's forward-backward algorithm revisited," *Pattern Recognition Letters*, vol. 3, pp. 369–373, 1985.

[9] M. Gupta and J. Liu, "Discovery of conserved sequence patterns using a stochastic dictionary model," *Journal of the American Statistical Association*, vol. 98, pp. 55–66, 2003.

[10] Jennings, M., I.R. Beacham Co-dependent positive regulation of the ansB promoter of Escherichia coli by CRP and the FNR protein: a molecular analysis.Mol Microbiol. 1993 Jul;9(1):155-64.

[11] Keles, van der Laan, and Eisen (2002) Identification of regulatory elements using a feature selection method, Bionformatics 18:1167-1175.

[12] K. Lange, D. R. Hunter, and I. Yang, "Optimization transfer using surrogate objective functions (with discussion)," *Journal of Computational and Graphical Statistics*, vol. 9, pp. 1–59, 2000.

[13] C. E. Lawrence, and A. A. Reilly, "An expectation maximization (em) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences," *Proteins*, vol. 7, pp. 41–51, 1990.

[14] C. E. Lawrence, S. F. Altschul, M. S. Bogouski, J. S. Liu, A. F. Neuwald, and J. C. Wooten, "Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment," *Science*, vol. 262, pp. 208–214, 1993.

[15] Liao, J., R. Boscolo, Y. Yang, L. Tran, C. Sabatti, and V. Roychowdhury (2003) "Network component analysis: Reconstruction of regulatory signals in biological systems" to appear in *PNAS*

[16] Kiejung Park, Sookyoung Choi, Minsu Ko, Chankyu Park (2001) Novel (F-dependent genes of Escherichia coli found using a specified promoter consensus. FEMS Microbiology Letters 202, 243-250.

[17] Quandt, K., K. Frech, H. Karas, E. Wingender, T. Werner, “MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data,” *Nucleic Acids Res.* vol 23, pp. 4878–4884, 1995.

[18] K. Robison, A. M. McGuire, and G. M. Church, “A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K12 genome,” *Journal of Molecular Biology*, vol. 284, pp. 241–254, 1998.

[19] Sabatti, C. (2003) “Sequence and gene expression array information: a feedback loop?” invited by and to appear in *Current Genomics*.

[20] C. Sabatti and K. Lange, “Genomewise motif identification using a dictionary model,” *IEEE Proceedings*, vol. 90, pp. 1803-1810, 2002.

[21] C. Sabatti, L. Rohlin, M. Oh, J. Liao, “Co-expression pattern from DNA microarray experiments as a tool for operon prediction,” *Nucleic Acid Research*, vol. 30, pp. 2886–2893, 2002.

[22] T. D. Schneider and R. M. Stephens, Sequence Logos: A New Way to Display Consensus Sequences Nucl. Acids Res. 18: 6097-6100, 1990.

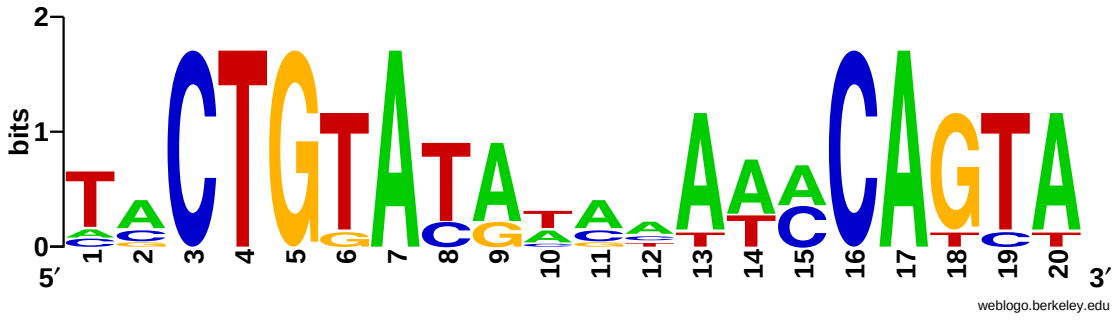

Figure 1: Profile of the binding site for LexA as reconstructed by Vocabulon starting from 6 E. Coli sequences. Sequence logo was depicted using the server `http://weblogo.berkeley.edu/`

Word	recovered sites	missed sites	imputed sites
araC	6	0	6
arcA	8	5	28
argR	15	2	24
cpxR	11	1	29
creB	8	0	9
crp	36	13	131
cspA	4	0	4
cytR	2	3	7
dnaA	7	1	41
fadR	7	0	8
fis	8	7	36
fliA	12	0	14
fnr	12	0	14
fruR	12	0	18
fur	8	1	18
galR	7	0	10
gcvA	4	0	4
glpR	7	6	20
hipB	2	2	2
lexA	19	0	24
malT	4	6	6
metJ	6	3	8
metR	5	3	10
nagC	6	0	9
narL	7	3	9
narP	8	0	4
ntrC	4	1	4
ompR	5	4	28
oxyR	4	0	4
phoB	10	2	12
purR	21	1	25
rpoH2	6	1	6
rpoH3	8	0	8
rpoN	6	1	11
rpoS17	5	10	9
rpoS18	4	3	8
soxS	11	6	22
torR	3	1	5
trpR	4	0	4
tus	5	0	5
tyrR	13	4	19
	340	90	663

Table 1:

Word	recovered sites	missed sites	imputed sites
araC	6	0	9
arcA	6	7	60
argR	15	2	108
cpxR	7	5	99
creB	8	0	19
crp	34	15	610
cspA	3	1	12
cytR	1	4	55
dnaA	6	2	96
fadR	6	1	21
fis	8	7	200
fliA	12	0	25
fnr	11	1	43
fruR	11	1	43
fur	8	1	69
galR	5	2	10
gcvA	4	0	6
glpR	6	7	71
hipB	0	4	2
lexA	19	0	46
malT	0	10	0
metJ	5	4	13
metR	4	4	44
nagC	6	0	22
narL	4	6	18
narP	8	0	7
ntrC	4	1	6
ompR	4	5	238
oxyR	4	0	4
phoB	9	3	35
purR	17	5	47
rpoH2	6	1	9
rpoH3	8	0	13
rpoN	6	1	22
rpoS17	1	14	4
rpoS18	3	4	5
soxS	9	8	61
torR	3	1	14
trpR	4	0	6
tus	5	0	5
tyrR	10	7	54
	296	134	2231

Table 2:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-2.021e-02	7.223e-03	-2.798	0.00518	**
malT	2.201e-01	3.320e-01	0.663	0.50737	
torR	3.703e-04	1.692e-01	0.002	0.99825	
dnaA	4.662e-02	5.521e-02	0.844	0.39851	
metR	-5.957e-02	9.084e-02	-0.656	0.51204	
arcA	-6.376e-02	7.604e-02	-0.839	0.40181	
cpxR	-4.336e-02	6.356e-02	-0.682	0.49521	
creB	1.419e-01	1.274e-01	1.113	0.26560	
metJ	1.074e-01	1.501e-01	0.715	0.47451	
rpoN	4.535e-02	1.372e-01	0.330	0.74109	
fruR	-7.020e-03	6.206e-02	-0.113	0.90995	
narL	-4.918e-02	1.565e-01	-0.314	0.75334	
narP	-9.861e-02	1.726e-01	-0.571	0.56788	
galR	-5.055e-02	1.060e-01	-0.477	0.63348	
fadR	-1.570e-01	1.075e-01	-1.461	0.14417	
ntrC	4.059e-02	1.691e-01	0.240	0.81035	
fur	-8.345e-02	6.062e-02	-1.377	0.16875	
argR	4.024e-02	3.695e-02	1.089	0.27625	
cytR	-2.131e-02	8.390e-02	-0.254	0.79955	
soxS	1.803e-01	7.741e-02	2.329	0.01994	*
gcvA	-3.036e-01	1.533e-01	-1.981	0.04771	*
ompR	5.058e-02	3.729e-02	1.357	0.17504	
glpR	-1.230e-02	5.249e-02	-0.234	0.81477	
cspA	5.521e-02	1.176e-01	0.469	0.63889	
lexA	3.841e-01	4.572e-02	8.401	< 2e-16	***
tyrR	8.268e-03	4.717e-02	0.175	0.86089	
fnr	3.284e-02	7.832e-02	0.419	0.67501	
crp	-6.978e-03	1.521e-02	-0.459	0.64648	
phoB	1.522e-01	8.682e-02	1.753	0.07964	.
nagC	1.815e-01	9.739e-02	1.864	0.06244	.
tus	7.899e-02	2.519e-01	0.314	0.75384	
trpR	-3.488e-01	2.234e-01	-1.561	0.11867	
purR	1.278e-01	5.306e-02	2.409	0.01604	*
fliA	-8.639e-02	1.031e-01	-0.838	0.40230	
rpoS17	3.678e-01	1.796e-01	2.049	0.04060	*
rpoS18	-3.606e-02	2.856e-01	-0.126	0.89955	
hipB	4.159e-01	2.494e-01	1.668	0.09552	.
fis	1.428e-01	3.312e-02	4.313	1.67e-05	***
oxyR	4.259e-02	2.878e-01	0.148	0.88236	
araC	4.065e-05	1.293e-01	0.000314	0.99975	
rpoH3	4.380e-01	1.444e-01	3.034	0.00244	**
rpoH2	4.005e-01	1.772e-01	2.260	0.02387	*

Table 3:

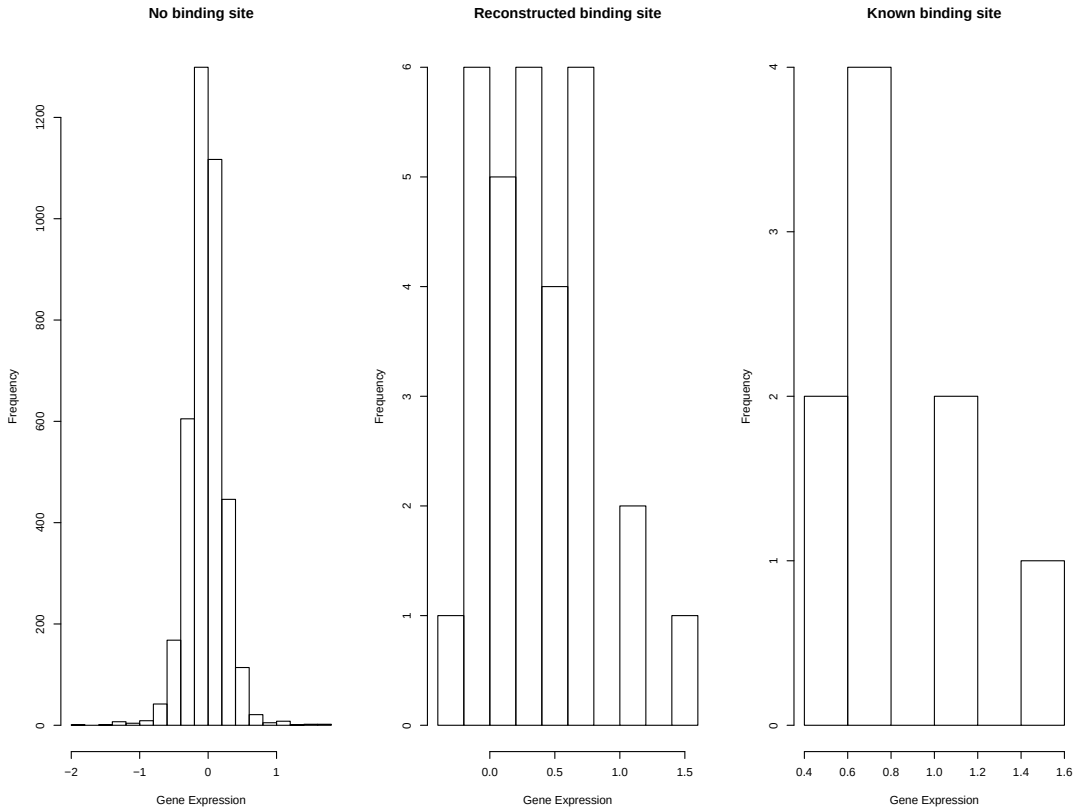

Figure 2: Histograms of the gene expression values in the second timepoint of the UV experiment. From left to right: genes that do not have a binding site for lexA according to the Vocabulon reconstruction; genes that do have a binding site for LexA according to Vocabulon; and genes that are known to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

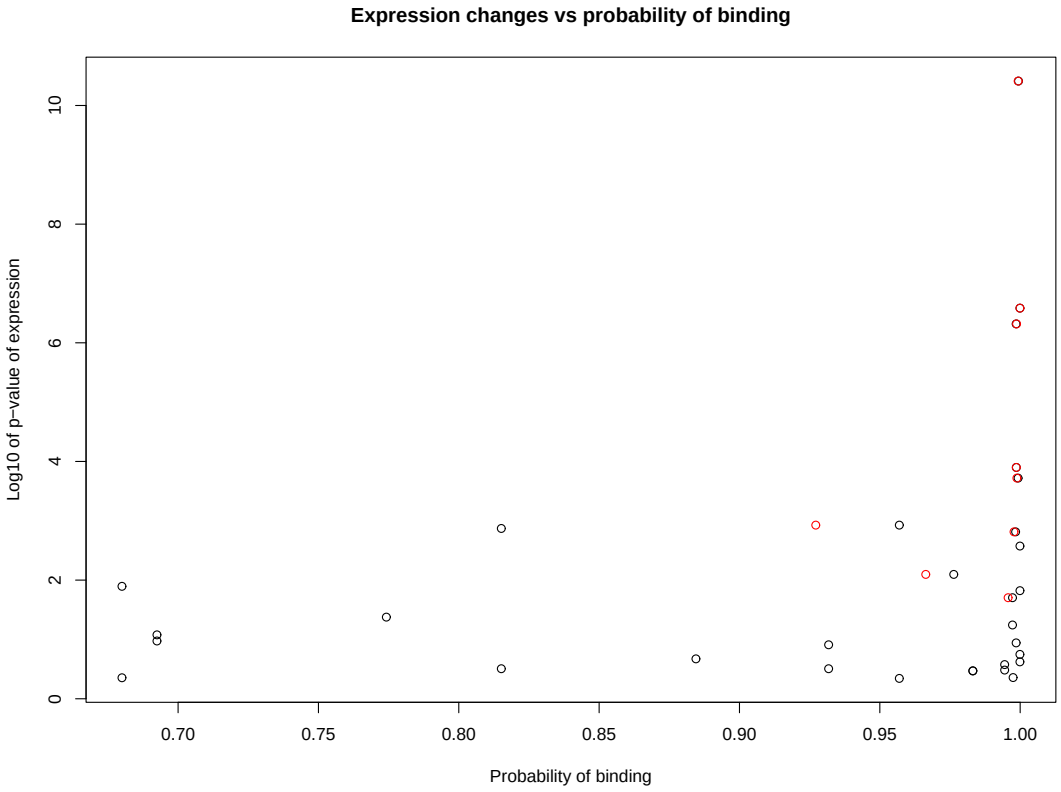

Figure 3: Scatter plot of significance of expression changes vs probability of binding site for LexA. Each circle represents a gene. Red circles represent genes that are known in the literature to have a binding site for LexA. On the x axis, the posterior probability of a LexA site as reconstructed by Vocabulon. On the y axis, the logarithm base 10 of the p-value of a test for the null hypothesis of no expression change in time point 2 of the UV experiment.

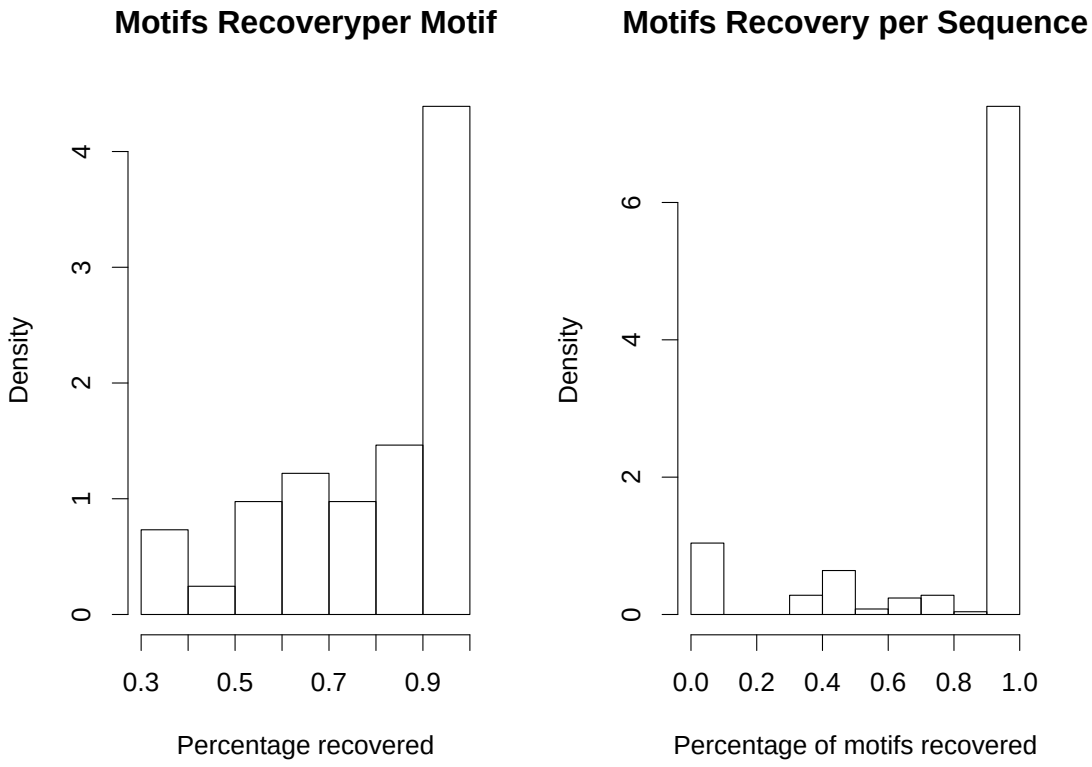

Figure 4: On the left, histogram of the percentage of recovered sites, for each of the 42 motifs. On the right, histogram of the percentage of recovered sites for each of the 233 sequences.

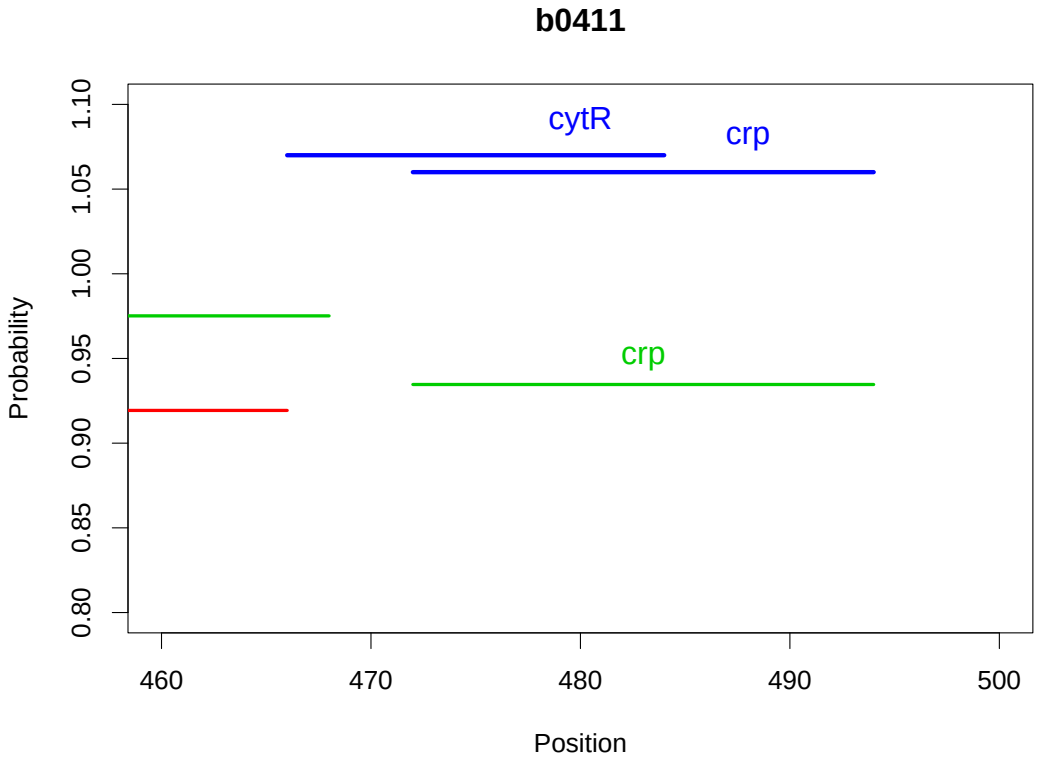

Figure 5: Graphical illustration of part of the Vocabulon output for the up-stream sequence of gene b0411. On the x axis position in base pairs, on the y axes value of the posterior probability reconstructed for binding sites (values above one are used to displayed known binding sites). This is an example of overlapping motifs, where only one of the motifs is reconstructed (crp).

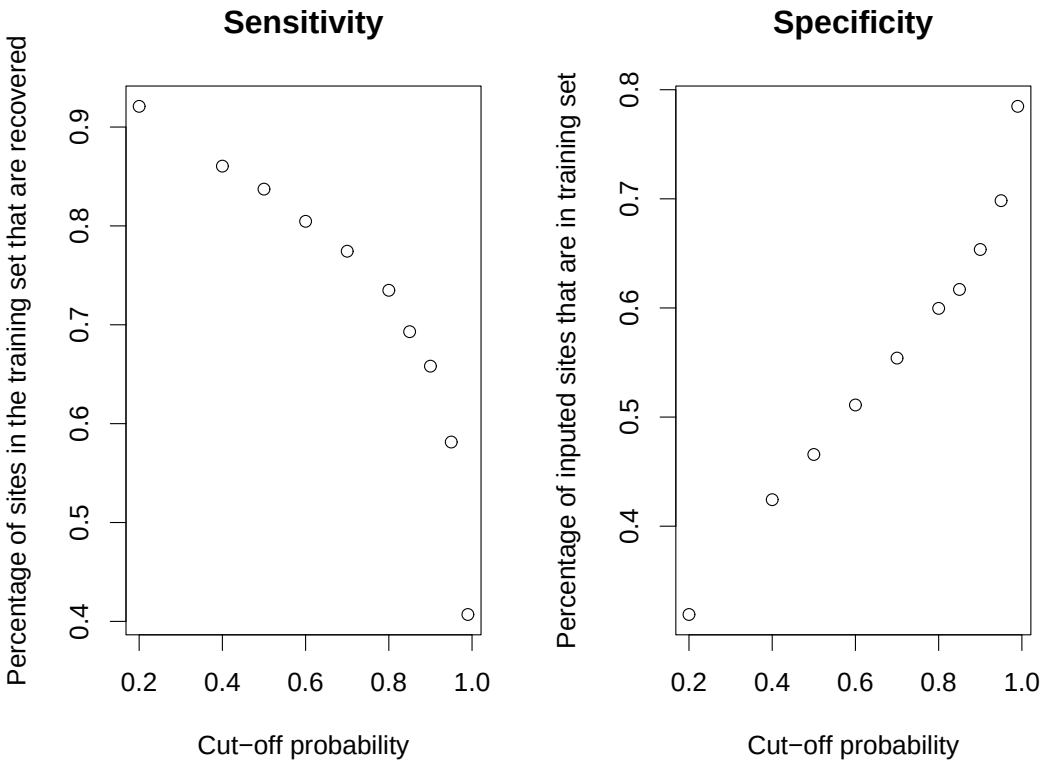

Figure 6: Sensitivity (left) and specificity (right) of the motif reconstruction on the 233 sequence set as a function of the cut-off value for the posterior probability.

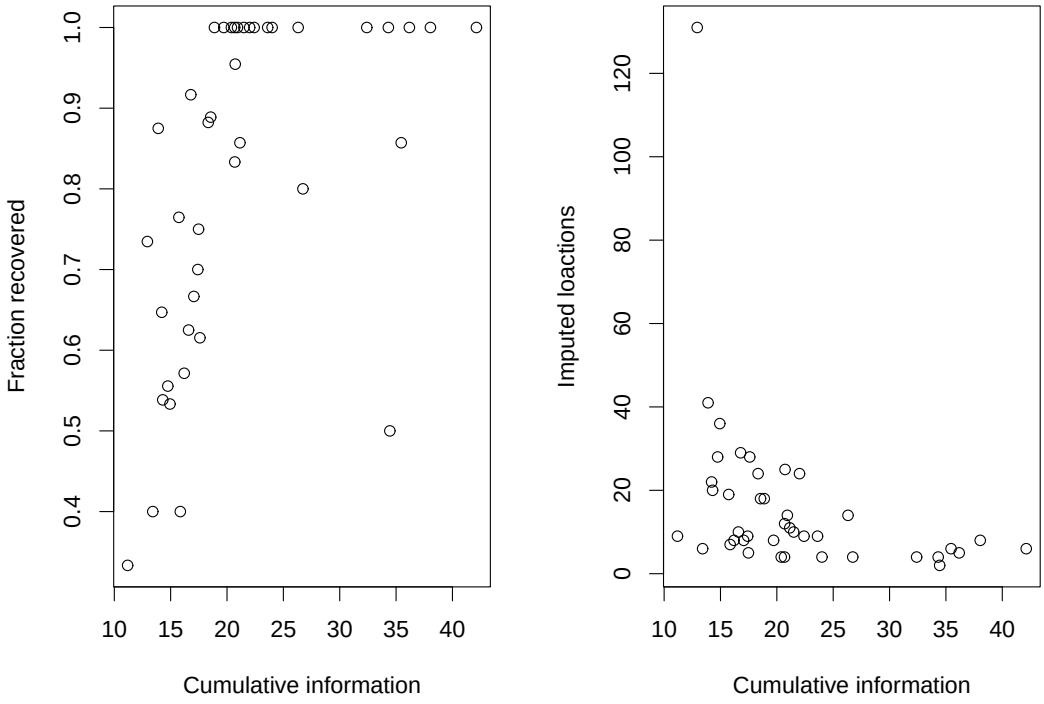

Figure 7: Fraction of recovered motif and number of imputed motif as a function of cumulative information.

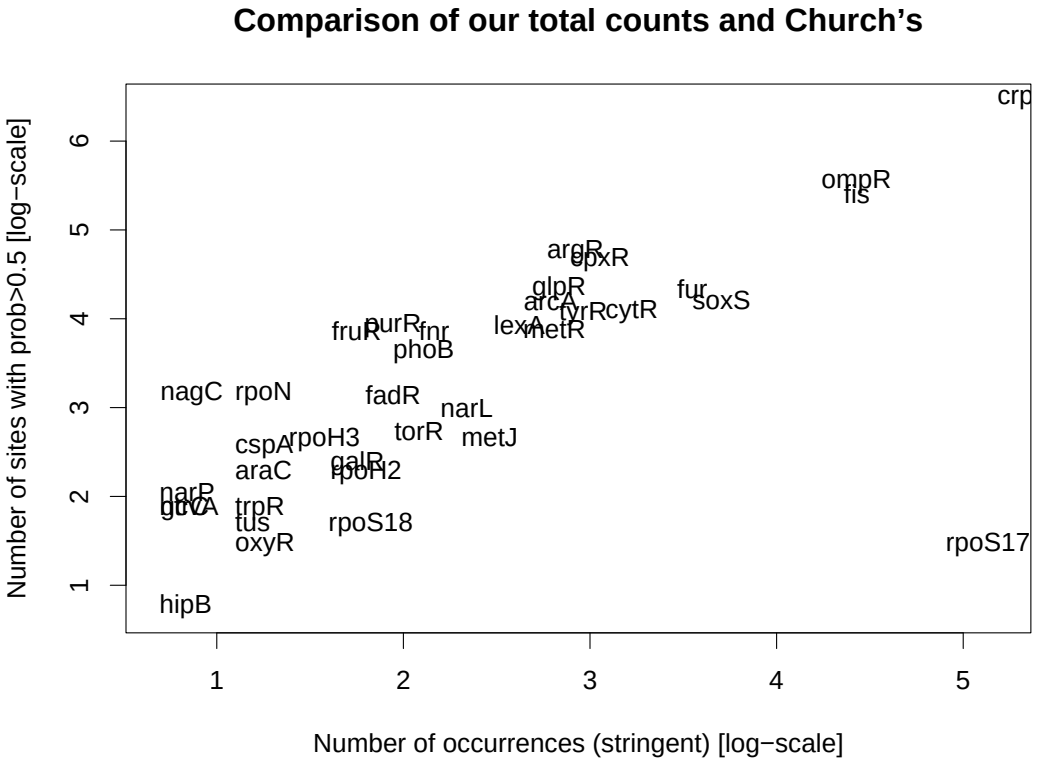

Figure 8: Scatter plot of the total number of predicted sites for the 42 motifs in our dictionary with Vocabulon (y axis) and with the strict criteria in [18] (x axis). Note that numbers are in the log-scale.

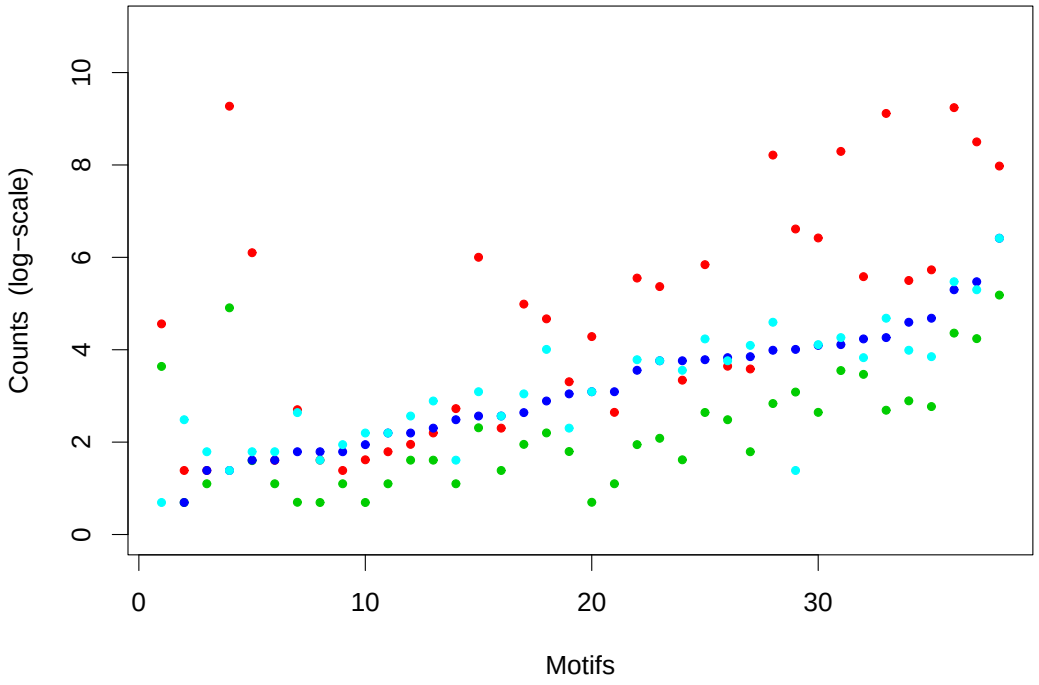

Figure 9: Comparison of predicted counts per motif. In red predictions of [18], in green stringent predictions of [18]. In blue expected counts with Vocabulon, and in cyan, number of sites with posterior probability greater than .05 with Vocabulon.

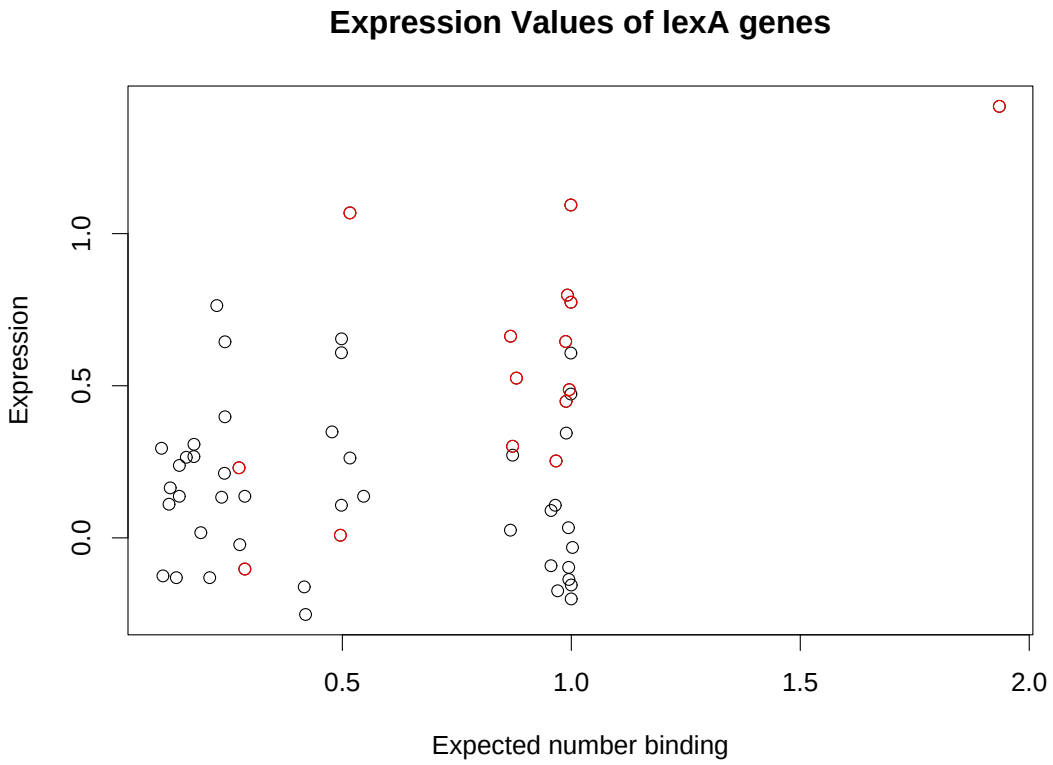

Figure 10: Scatter plot of the expression values in time point 2 of the UV experiment vs the predicted probability of binding site for LexA. Genes with known binding site for LexA are in red.