

Chapter 1

Methods for Discovery and Characterization of DNA Sequence Motifs

Philipp Bucher

1. Introduction

Motif discovery is considered to be an important problem in bioinformatics, as documented by a large number of papers. It is also believed to be a hard and still partly unsolved problem, despite considerable efforts by many distinguished researchers. Finally, it is an old problem with a long tradition in bioinformatics. An early example is the discovery of the Pribnow box in *E. coli* promoters.¹ Although this motif was found by visual inspection of DNA sequences, it was probably instrumental in defining the paradigm that subsequently led to the formalization of the motif discovery problem in its modern form and to the development of algorithms to solve it.

A DNA motif, such as the Pribnow box shown in Fig. 1, is defined by a set of short subsequences from longer sequences with high similarity. The subsequences share some common features, which typically are described by a consensus sequence or weight matrix. A motif must be overrepresented in a biologically defined collection of genome sequences, i.e. it must occur more frequently than



Fig. 1. The Pribnow box, an early discovered DNA motif. The Pribnow box is a promoter element of *E. coli* promoters, originally discovered by visual inspection of six experimentally characterized promoter sequences.¹ (a) Input sequence set. (b) Consensus sequence proposed in the original paper (R is the IUPAC (International Union of Pure and Applied Chemistry) code for A or G). (c) Input sequence set with highlighted motif instances. The set of motif instances is also referred to as “motif annotation” in this chapter. (d) Base count frequency matrix. (e) Base probability matrix estimated by adding one pseudocount to each element of the base count frequency matrix (probabilities are given as percentages). (f) Weight matrix. The position-specific weights of corresponding bases are summed up to compute a score for a DNA sequence of the same length as the motif. The weights of the matrix were computed as a natural log-likelihood ratio from the base probabilities, multiplied by 10, and rounded to the nearest integer (see Chapter 2).

one would expect by chance. The DNA motif has become a central concept of molecular biology, a research field which has its roots in biology as well as in physics. In order to understand why motifs are of interest, a brief look at the leading paradigms of both disciplines will be useful.

1.1. Motif Discovery from a Biological Perspective

The basis of modern biology is the theory of evolution by natural selection introduced by Darwin and Wallace. One of the tenets of this theory is that any genetically encoded biological structure is subject to the randomizing forces of mutation and eventually will disappear if not conserved by natural selection. According to Williams,² constancy and complexity are biological proof of function, even in the absence of a conceivable mechanism by which a conserved structure might contribute to the organism's fitness. The lateral organ of fishes is cited as an example. The high complexity of this organ and its high degree of conservation across species prompted biologists to carry out experiments, which eventually led to the identification of its function as a sensory organ. This is exactly the biological motivation behind motif discovery. Sequence conservation is evidence of natural selection and thus justifies an investment of experimental work to elucidate the function of a motif. In fact, this approach has been very successful in the study of protein function. There also, the discovery of a new conserved domain has often preceded the characterization of its molecular function. Even though this chapter is focused on DNA motifs, many of the concepts and methods introduced extend readily to RNA and protein sequence motifs.

A minimal degree of complexity is an essential property of a motif, as motifs of low complexity may frequently occur by chance and thus cannot meet the condition of overrepresentation. While the complexity of a morphological structure can be judged by human visual intuition, the complexity of DNA sequence motifs is typically evaluated by a conditional entropy-based index borrowed from information theory.³ Unlike the search for new protein motifs, DNA motif discovery is often targeted to a particular function, which may, however, be broadly defined. For instance, by searching eukaryotic promoter sequences for conserved DNA motifs, one typically expects to find the target binding sites for a variety of *a priori* unknown transcription factors.

Sequence motifs can also be viewed as taxonomic entities. Mastering a bewildering diversity of phenomena through classification is a typical

biological approach that can be traced back to Carl von Linné's *Systema Naturae*. Since classification has contributed so much to our understanding of the living world, the discovery of a new species or the definition of a new medical syndrome is rightly considered as a scientific achievement in its own right. A relatively recent article in *Nature* on the discovery of a new mammalian species⁴ documents this view.

1.2. DNA Motifs from a Physical Perspective

To a physicist, the definition of a taxonomic entity hardly represents the endpoint of a research project. The physical approach aims at causal relationships between observable events, and at quantitative models that can predict the outcome of experiments. Surprisingly, DNA motif discovery has found important applications in such a research setting too. A classical example is the characterization of transcription factor binding sites, where the DNA motif becomes a quantitative model to predict the binding energy of a protein–DNA complex (Fig. 2). In fact, the Pribnow box mentioned before is also a part of a DNA–protein binding site, the one recognized by bacterial RNA polymerase. The Berg and von Hippel⁵ statistical mechanical theory of protein–DNA interactions provides a connection between motif complexity (conservation) and binding energy. Interestingly, the standard descriptor used for representing a protein binding site, the energy matrix, is mathematically equivalent to the weight matrix used in *de novo* discovery of evolutionarily conserved motifs. However, the logic of the scientific inference process that leads to the definition of the matrix is reversed; here, the starting point is a known molecular function and the endpoint is an initially unknown motif which can be considered as the “genetic code” for the function.

From a computational chemistry viewpoint, an energy matrix for a transcription factor binding site is a special case of quantitative structure–activity relationship (QSAR) model.⁶ There is a wealth of literature about QSAR models that is only sparsely cited in DNA motif discovery papers. Machine learning methods exploiting quantitative activity data have been widely used in the QSAR field. Interestingly, the first weight matrix-like structure used for representation of a nucleic acid

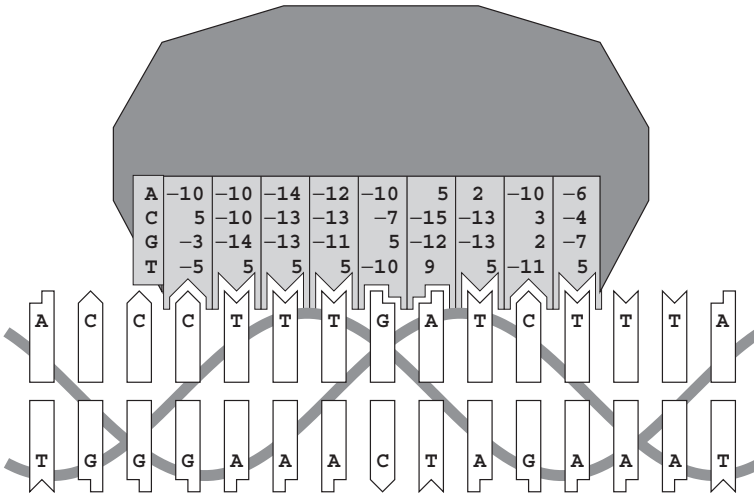


Fig. 2. Energy matrix for a transcription factor binding site. An energy matrix represents one possible physical interpretation of a weight matrix. Each element of the matrix quantitatively defines the binding energy (in arbitrary units) between a DNA base pair and a corresponding compartment of the DNA-binding surface. Energy matrices can be used to compute binding constants for DNA protein complexes, and therefore represent a special case of QSAR (quantitative structure–activity relationship) models. Note that the sign of the energy units is reversed; a high weight matrix score signifies low energy value, and thus high binding strength.

sequence motif (ribosome binding sites) was inspired by a machine learning method called “perceptron”.⁷

In summary, DNA motif discovery is not an isolated, specialized topic for a closed circle of bioinformaticians. DNA motifs have many facets and have different meanings to different researchers. Mathematically equivalent descriptors have been used in many more fields, even outside life sciences. Hidden Markov models,⁸ for instance, were developed in the speech recognition field.

In the following sections, a personal view of motif discovery will be presented inspired partly by the author’s own work. The focus will be on essential concepts and open questions. Methods will be presented in their most basic version; a comprehensive review of current state-of-the-art motif discovery algorithms is beyond the scope of this chapter. Further references on methods can be found in recent reviews.^{9,10}

2. Motif Discovery in a Nutshell

Knowing the inputs and outputs is central to the understanding of a computational problem. The data structures involved in motif discovery are shown in Fig. 1. The input consists of a set of sequences, not necessarily of fixed length. The output consists of a list of motif instances and/or a motif description: a consensus sequence, a probability matrix, or a weight matrix. The motif instances are subsequences of the input sequence, and can be defined by a sequence name and a starting position. For the type of motifs considered here, they are of fixed length. The motif description and motif annotation are intertwined entities in that the motif description defines the motif annotation of the input data set, and the set of motif instances can be used to derive the motif description.

A consensus sequence is a short sequence (k -letter word) from the DNA alphabet or from an extended alphabet containing IUPAC (International Union of Pure and Applied Chemistry) codes for incompletely specified bases in nucleotide sequences.¹¹ A threshold number of mismatches may be permitted. The consensus sequence, together with the maximal number of allowed mismatches, defines the motif in a deterministic and qualitative manner. Specifically, it defines the subset of all k -letter words which qualify as motif instances.

The position-specific scoring matrix, introduced in its standard form by Staden,¹² is a more flexible representation of a sequence motif. Its use is motivated by the assumption that not all mismatches to consensus sequences are equally detrimental. Therefore, the relative fit of a particular base to a given motif position is expressed by a number. Matrix descriptions for sequence motifs come in two forms: base probability matrices and additive scoring matrices, henceforth called weight matrices. The former reflects the expected frequencies of each base at each position. The latter serves to compute a motif score for a particular k -letter sequence by adding up the matrix elements corresponding to all bases at each position in the sequence. The parameters of a weight matrix can have positive or negative values.

The base probabilities are often estimated by adding one pseudo-count to the observed base count of base b at position i :

$$p(i, b) = \frac{c(i, b) + 1}{4 + \sum_{b'=A}^T c(i, b')} \quad (1)$$

The elements of a weight matrix may be computed from a probability matrix as a log-likelihood ratio:

$$w(i, b) = \ln \frac{p(i, b)}{p_0(b)} \quad (2)$$

Here, $w(i, b)$ and $p(i, b)$ are the weight and probability of base b at position i of the motif, respectively, and $p_0(b)$ is the background probability of base b . Note, however, that log-likelihood ratios are not universally used in the field. One of the best known motif search programs, MATINSPECTOR, uses a different way of scoring transcription factor binding sites with a base probability matrix.¹³

Like a consensus sequence, a weight matrix in conjunction with a cut-off value defines a subset of k -letter words which qualify as motif instances. However, the power of the matrix representation lies in the quantitative evaluation of candidate k -letter words, which can be exploited, for instance, for transcription factor binding site affinity prediction. On the other hand, a base probability matrix defines a motif in a probabilistic manner, a property which is exploited by probabilistic motif optimization methods such as expectation maximization.

The goal of the motif search problem is to find the best motif for a given set of input sequences. The complete statement of the problem requires the specification of a quality criterion (objective function) related to overrepresentation. A specific motif discovery method is thus characterized by three components: (a) the motif descriptor, (b) the objective function, and (c) the algorithm to scan the search space of possible motifs.

The large diversity of published motif search algorithms is readily classified along these lines. The dual nature of the motif discovery output, motif annotation and motif description, has implications on the search space that needs to be scanned. If the motif annotation is considered to be the primary result, then the search space consists of all possible motif annotations and the optimization problem becomes a gap-free local multiple alignment problem. If the motif description is considered the primary result, then the search space consists of all possible consensus sequences or weight matrices.

Some variations of the standard scheme outlined above deserve to be mentioned. Figure 1 suggests that each input sequence contains exactly one sequence motif. This is not a general requirement. In fact, motifs may occur only in a subset of the input sequences or more than once in a particular sequence. The popular motif discovery program MEME¹⁴ has established the following nomenclature for the three different motif search modes: “oops” for one occurrence per sequences, “zoops” for zero or one occurrence per sequence, and “anr” for any number of repetitions. Furthermore, a given input sequence set may contain more than one overrepresented motif type. Many algorithms and computer programs offer, in fact, the possibility to search for multiple motifs in one run. Finally, since genomic DNA is usually double-stranded, the motif search may be extended to the reverse-complementary strands of the input sequences.

3. Overview of the Methods

3.1. Objective Functions

As explained before, candidate motif descriptions need to be ranked by an index that reflects the degree of overrepresentation in a statistical sense. Two very different types of probabilities are used for this purpose:

- (a) the probability that a given motif occurs at least a certain number of times in the input sequence set; and
- (b) the probability of the input sequences, given the motif.

Probability (a) is minimized and, from a statistical viewpoint, reflects the classical frequentist approach exemplified, for instance, by the objective function introduced by van Helden *et al.*¹⁵ Probability (b) is maximized and is inspired by Bayesian statistics; in this case, the motif is part of a probabilistic generative model such as a hidden Markov model (HMM)⁸ or a mixture model.¹⁶ The two types of probabilities will be illustrated by examples below.

Let us first turn to the question of how to compute the probability that a motif occurs at least n times in an input sequence set. Here, and in all following examples, we will assume that motif frequencies were determined in the “anr” search mode. An exact solution to this problem is hard to obtain because of the statistical nonindependence of overlapping words.¹⁷ In fact, the probability distribution of a k -letter word to occur zero to N times in a sequence of length $N + k - 1$ depends on the internal repeat structure of the word. To bypass this difficulty, motif discovery algorithms often rely on approximations, which are debatable from a mathematical viewpoint. According to the frequently used Poisson approximation, which assumes independence between motif occurrences, the probability of finding a motif exactly n times is given by

$$\text{Prob}(n, E_i) = \frac{1}{n!} E_i^n \exp(-E_i). \quad (3)$$

Here, E_i is the expected number of occurrences of a given motif i , which is the product of the search space N and the probability p_i that a random sequence of length k constitutes an instance of motif i . The search space is the number of all possible starting positions for a motif of length k in the input DNA sequence set.

If the motif description consists of a consensus sequence based on the four-letter DNA alphabet, with a maximal number of m mismatches allowed, then the probability p_i may be computed as follows:

$$p_i = \sum_{j=0}^m \binom{k-j}{j} 0.25^{(k-j)} 0.75^j. \quad (4)$$

The assumption underlying this formula is that all bases occur with an equal probability of 0.25 in random sequences. This is the simplest background (null) model that can be used in this context. Markov chains, which assume unequal probabilities for different bases and dependencies between consecutive bases, are more realistic background models for genomic DNA sequences. Algorithms have been presented for computing p_i for such a model, as well as for consensus sequences including ambiguous positions represented by IUPAC codes¹⁸ and also for weight matrices.¹⁹

The Bayesian approach will be illustrated with the mixture model used by the program MEME. Again, we assume the “arn” search mode. To circumvent the mathematical difficulties of overlapping words statistics, the input sequence set is usually evaluated as if it were to consist of N nonoverlapping k -letter subsequences (N is the search space defined before). In the simplest case, the mixture model consists of two components, a motif model given by a probability matrix and a background model given by a base probability distribution. The probability of the sequences given the model is then computed as

$$\text{Prob}(x, M, M_0, q) = \prod_j \left(qP(x^j | M) + (1 - q)P(x^j | M_0) \right). \quad (5)$$

In this notation, x denotes the total set of overlapping k -letter subsequences contained in the input sequences, and x^j is an individual member of it. $P(x^j | M)$ and $P(x^j | M_0)$ are the probabilities of subsequence x^j given the motif model and the background model, respectively. q is the mixture coefficient indicating the sequence-independent probability that a given subsequence constitutes a motif. The models M and M_0 both define probability distributions over all k -letter words. The probabilities of sequence x^j under the motif and background models, respectively, are defined as follows:

$$P(x^j | M) = \prod_{i=1}^k p(i, x_i^j) \quad (6)$$

$$P(x^j | M_0) = \prod_{i=1}^k p_0(x_i^j). \quad (7)$$

Note that the mixture coefficient q is part of the model and, thus, the target of optimization by the motif discovery algorithm. It plays a similar role as the threshold value of allowed mismatches to a consensus sequence in the frequentist motif evaluation framework. This raises the interesting question of whether the two approaches are equivalent with regard to defining the optimal threshold value for a consensus sequence or weight matrix. The answer is, to my knowledge, not known.

3.2. Scanning the Search Space

How do we find the best motif among all possible motifs? Three strategies can be distinguished depending on the structure of the search space, which may consist of

- (a) all k -letter words of a given alphabet;
- (b) all probability matrices of length k ; or
- (c) all motif annotations (all subsets of k -letter subsequences of input sequences).

3.2.1. Finding the best consensus sequence

For consensus sequences based on the four-letter DNA alphabet, the size of the search space remains computationally manageable up to a word length of about 15. The optimal motif can thus be found by enumeration, i.e. by evaluating a frequentist-type objective function for each k -letter word. Algorithms to this end are reasonably fast, as the input sequences need only to be scanned once. The word index is first initialized with zeros. Then, one word frequency is incremented each time a subsequence is processed (if mismatches are allowed, multiple motif frequencies are updated for one subsequence). For longer words, heuristic algorithms have to be used instead of exact methods. An old trick, introduced in the

early 1980s,²⁰ is to restrict the search space to those k -letter words which actually occur in the input sequence set. A popular implementation of the word search strategy is provided by the program Weeder.²¹

For consensus motifs based on the complete 15-letter IUPAC code, the search space becomes too large for enumerative approaches. Progress has recently been reported in developing efficient heuristics under certain constraints.²² It is debatable, however, whether such algorithms are really needed. Consensus sequences with ambiguous symbols are still less flexible than position-specific scoring matrices, and efficient and effective algorithms are readily available for optimizing such matrices.

3.2.2. Optimizing a base probability matrix

The search space for probability matrices is continuous and, thus, potentially infinite. Rigorous algorithms guaranteed to find the optimal motif are not available or even conceivable for now. The standard heuristic approach to optimize a probability matrix is to start from an initial motif description referred to as a seed, and to use an iterative refinement algorithm to reach a local maximum for the objective function. Expectation maximization (EM) is the classical approach used in this context, introduced to computational biology by Lawrence and Reilly.²³ The iterative part is straightforward. Based on a current model M^k , one uses Bayes' formula to compute for each subsequence x^j a weight w_j^k , which is the posterior probability that the subsequence constitutes a motif instance. For the mixture model introduced above, one obtains

$$w_j^k = P(M^k | x^j) = \frac{q^k P(x^j | M^k)}{q^k P(x^j | M^k) + (1 - q^k) P(x^j | M^0)}. \quad (8)$$

These probabilities are then used as weights to compute a new probability matrix by adding up weighted base contributions from all subsequences of the input sequence set:

$$p^{k+1}(i, b) = \frac{\sum_{j=1}^N w_j^k \delta(x_i^j = b)}{\sum_{j=1}^N w_j^k}, \quad (9)$$

where $\delta(x_i^j = b)$ is 1 if $x_i^j = b$ and 0 otherwise. Note that, in practice, only a few subsequences ... the likely motif instances ... will make significant contributions to the new model. The new mixture coefficient q is obtained as the sum of posterior probabilities for all k -letter subsequences divided by the search space:

$$q^{k+1} = \left(\frac{1}{N} \right) \sum_{j=1}^N w_j^k. \quad (10)$$

Expectation maximization is a deterministic algorithm that can get trapped in a local optimum. To have a chance of reaching the globally optimal motif, it has to start from a good seed, which is already relatively close to the target. Two strategies are used for this purpose. One is to trigger the algorithm from a large number of random seeds; this is time-consuming and thus relatively inefficient. A better way is to use a consensus sequence obtained from a fast exact or heuristic word search algorithm, as described above. The combination of a word search algorithm for the seeding step and EM for refinement is probably the most effective motif discovery strategy used nowadays, and is implemented in various forms in many software tools. MEME, for instance, uses seeds obtained from exhaustive pairwise comparison of k -letter subsequences, an approach which leads to good seeds but has the unfavorable time complexity $O(N^2)$ for the seeding step.

As explained before, the mixture-model-based EM approach does not account for the nonindependent occurrence of overlapping subsequences. The HMM framework offers an explicit way of computing posterior motif probabilities, under the constraint that motif instances must not overlap, via the choice of an appropriate model architecture. In all other respects, the “training” of a HMM, i.e. the iterative optimization of the model with respect to the input sequence set, is identical to the optimization of a mixture model.

3.2.3. *Optimizing the motif annotation*

Keeping the same probabilistic mixture model framework, the motif discovery problem can also be formulated in such a way that the motif positions

are the target of optimization. The search space thus becomes discrete and finite, but still remains astronomically large. The classical algorithm in this category is the Gibbs sampler,²⁴ a stochastic variant of EM. This approach focuses on actual motif instances rather than a generalized description of the motif. In other words, one would primarily like to know the locations of the functional elements in the given sequences rather than be able to predict motif instances in new sequences. From a theoretical and computational viewpoint, the differences are relatively minor, since both EM and Gibbs sampling update a probability matrix. However, whereas EM assigns k -letter subsequences probabilistically to the two mixtures of the model, Gibbs sampling attributes them explicitly to one or the other class. The decision whether a given subsequence x^j is included in the motif set for the next iteration is made randomly based on its current posterior probability w_j^k , as defined previously. The new probability matrix is typically estimated by a maximum *a posteriori* (MAP) likelihood method, for instance, by adding one pseudocount to each base frequency:

$$p^{k+1}(i, b) = \frac{1 + \sum_{j \in \mathbf{M}^k} \delta(x_i^j = b)}{|\mathbf{M}^k| + 4}, \quad (11)$$

where $\delta(x_i^j = b)$ is 1 if $x_i^j = b$ and 0 otherwise. \mathbf{M}^k denotes the current set of motif instances, and $|\mathbf{M}^k|$ the number of elements therein.

The fact that Gibbs sampling has an in-built stochastic element helps to overcome the risk of getting trapped in a local optimum far away from the global optimum. Another advantage of a nondeterministic algorithm is that it potentially returns different results when run several times from the same seed, which also increases its chances of finding the globally optimal solution. Optimization in motif annotation space can also be done in a deterministic fashion. For instance, one could define the new motif instances by accepting all k -letter subsequences with posterior probabilities higher than 0.5. This leads to an iterative, multiple alignment algorithm, similar to the PATOP algorithm described in Sec. 5.

3.2.4. Finding multiple motifs

In exploratory applications, such as mining promoter sequences for new transcription regulatory motifs, one often expects to find more than one motif. For instance, a landmark paper on *Drosophila* promoters²⁵ reported 10 *ab initio* discovered motifs returned in one program run by MEME. Fortunately, there is a simple and efficient way to extend the basic algorithms presented above to multiple motif discovery. The principle is to proceed iteratively by searching for one motif at a time, and by progressively excluding motif instances found from subsequent iterations. More formally, this means that, after each cycle, the k -letter subsequences attributed to the newly discovered motif are removed from the search space — a process that is commonly referred to as “masking” in the sequence analysis literature. A theoretically more proper approach would use multi-component mixture models for synchronous optimization of several motifs at a time by EM, Gibbs, or a progressive local multiple alignment algorithm.

3.2.5. Estimating the significance of a newly discovered motif

The different types of probability values used as objective functions for motif optimization do not provide an answer to the question of whether the best motif found is significant or not, as they apply to single motifs and thus are not corrected for multiple tests. With consensus sequence motifs, a Bonferroni correction is sometimes applied; see, for instance, Xie *et al.*²⁶ However, this approach is likely to yield overly conservative P-value estimates, as consensus word frequencies are highly dependent on each other, especially if mismatches are tolerated. The program MEME provides significance estimates for matrix-based motif models based on a maximum likelihood ratio test (LRT), which takes into account the number of free parameters of the model.¹⁶ This approach is quite sensitive to the properties of the null model, and in practice tends to assign low E-values to questionable motifs. A good way to corroborate the significance of a newly found motif is to rerun the motif discovery program with randomized or shuffled sequences as a control, so as to get an idea of what P-values or E-values could be

expected for fortuitous motifs. Shuffling methods which preserve higher-order Markov chain properties²⁷ of the real sequences are recommended for this purpose.

4. Bottlenecks and Limitations

DNA motif discovery is considered to be a tough problem. This is surprising in view of the apparently simple structure of DNA motifs, as compared to protein sequence motifs. The perception that the problem is difficult is partly based on the poor track record in terms of important discoveries made by this approach, which were later confirmed by experimental follow-up studies. This contrasts with the great success of similar methods in discovering new protein sequence motifs.²⁸ Recent evaluation studies based on representative and realistic benchmark sequence sets indeed confirmed that current state-of-the-art motif discovery programs are highly ineffective in rediscovering experimentally characterized motif instances (transcription factor binding sites) hidden in gene regulatory sequences.^{29,30} However, these results have to be interpreted with caution. Let us first have a look at the benchmarking procedure.

4.1. Benchmarking Procedures for Motif Discovery

Claims about poor performance of motif discovery algorithms require that the community agrees on how performance is measured. In this sense, the recent benchmarking papers have made an invaluable contribution to structuring the field by better defining the problem. It is thus of paramount importance to the newcomer to understand how these tests were set up. The procedure is schematized in Fig. 3. The benchmark sets consist of DNA sequences of a few hundred base pairs in length containing annotated transcription factor binding sites, which constitute the motifs to be discovered. The experimental motif annotations of the eukaryotic benchmarking set were taken from TRANSFAC,³¹ while the prokaryotic test set is based on RegulonDB.³² Both resources are manually curated databases relying on experimental results published in journal articles.

One test per motif is carried out. An input sequence set consists of all sequences containing a particular motif. The experimental

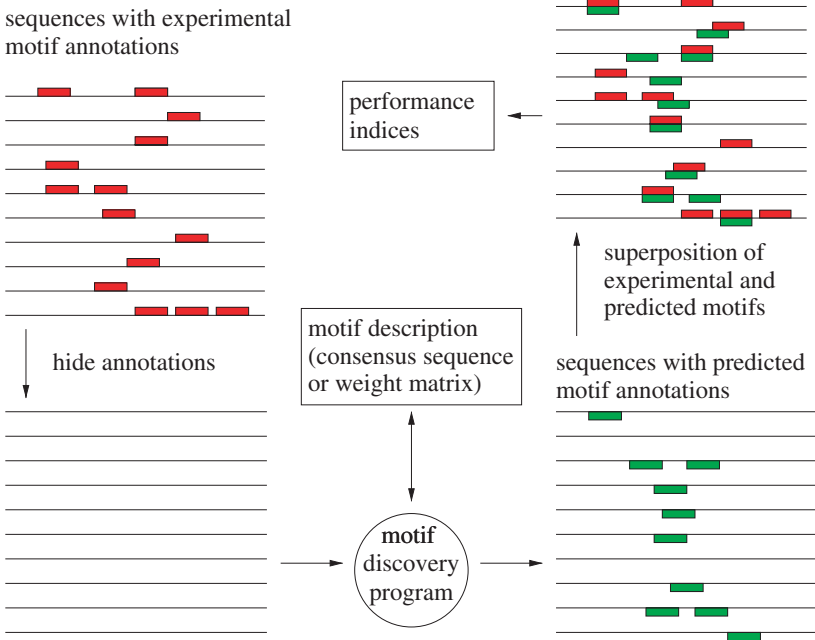


Fig. 3. Benchmarking protocol for evaluation of motif discovery algorithms. A test consists of DNA sequences which experimentally mapped binding sites to a particular transcription factor (experimental motif annotations are shown in red). The naked sequences (without motif annotations) are given as input to the motif discovery algorithm, which returns a set of predicted motif instances plus, optionally, a motif description. The experimental and predicted motif annotations are then superimposed for computation of a variety of performance indices, such as sensitivity and specificity. Partial overlaps between known and predicted motifs are usually counted as success. This protocol has been applied in the recent benchmarking studies described in Hu *et al.*²⁹ and Tompa *et al.*³⁰

annotations, of course, are hidden to the program. The task of the motif discovery program is to rediscover the hidden motif and to return the coordinates of the corresponding motif instances. The performance is evaluated on the basis of the overlap between experimental and predicted motif annotations, and is expressed by standard measures such as sensitivity (percentage of true motif instances overlapped by predicted motif instances) and specificity (percentage of predicted motif instances overlapped by true motif instances).

As mentioned before, the results of these studies were disappointing. With the eukaryotic benchmarking set, sensitivity and specificity varied between 10% and 30%,³⁰ and only marginally better results were obtained with the prokaryotic test set.²⁹ Similar results were obtained with synthetic sequences, where the experimentally defined motif instances were hidden in computer-generated random sequences corresponding to a Markov chain model.

The primary reason for the poor performance is probably related to the characteristics of the input sets, typically consisting of only a few (in the order of 10) rather long sequences. The total number of motif instances hidden in the test sequences was also relatively low, below 20 in most cases. These are unfavorable conditions for motif discovery. A large number of short sequences, highly enriched in a given motif, would have made the task of the motif discovery program much easier.

4.2. Why is Protein Domain Discovery Easier?

At first glance, regulatory proteins resemble gene regulatory regions in that they have a modular architecture. The modules are motifs; in proteins, they are also called conserved domains. The key difference, however, lies in the complexity of the modules.

Regulatory DNA elements are short and based on a smaller alphabet. Complexity is expressed by the information content IC , which is computed from a base probability matrix as follows³:

$$IC = \sum_{i=1}^k \sum_{b=A}^T p(i, b) \log_2 \frac{p(i, b)}{0.25}. \quad (12)$$

The definition of information content has the form of a conditional entropy, where the null model consists of uniform base probabilities of 0.25 for each base. The information content, which is expressed in bits, indicates the random occurrence probability of a motif, which is 2^{-IC} . Typical transcription factor binding site matrices in TRANSFAC have an IC value of about 10 bits, which means that they are expected to occur about once in every 1000 bp. Consequently, about one

match to a motif is expected to occur by chance in the input sequence sets that were used for benchmarking. Under these conditions, it is principally impossible to infer the true motif instances with high reliability. Since the probability matrices returned by motif discovery algorithms are derived from hypothetical motif instances in the input set, their quality is compromised by the contamination with false matches.

Conversely, the motifs corresponding to protein domains have a much higher complexity, often in the range of 30 bits or more. The higher information content is due to the increased length (up to 100 amino acids) and the larger size of the protein alphabet (20 instead of 4). Motifs with this degree of complexity are unlikely to occur by chance in a protein sequence of average length, and thus can be located with near certainty. The higher complexity also explains why protein domain discovery has often been initiated by a single statistically significant pairwise sequence match retrieved by database search.

4.3. Reasons for the Limited Success of DNA Motif Discovery

Based on the above considerations, I have doubts whether motif discovery is rightly considered to be a tough problem. The poor benchmarking results reported in recent papers are perhaps mostly due to the inadequacy of the input data sets. Shorter sequences would be needed to localize motifs with high confidence, and a larger number of motif instances would be required to obtain reliable base frequency estimates. Failure by the heuristic algorithms to find the optimal motif is unlikely to be a major reason for the poor benchmarking results. This could be tested by comparing the true versus predicted motif annotations in terms of the objective function used by the motif discovery program. My conjecture is that the true motif annotation will look less good in such a test. The fact that the consensus sequence-based motif search program Weeder²¹ showed the best performance in the above-described tests supports this hypothesis.

In an overfitting situation due to sparse data, methods based on a simpler model with fewer degrees of freedom tend to perform better.

The true problem with DNA motif discovery is that biologists have published consensus sequences and weight matrices based on very few sequences for too many years, whereas computational biologists were mostly concerned with algorithmic improvements aimed at finding the globally optimal motif with higher probability and in shorter time. In the future, more efforts should be spent on analyzing the limitations of motif discovery in light of a statistical inference problem.

5. Locally Overrepresented Sequence Motifs

This last section summarizes a variant of the classical motif search problem, introduced by the author about 25 years ago for the study of promoter sequences. This method, named signal search analysis (SSA),³³ takes into account the fact that certain classes of DNA sequences such as promoters are experimentally defined by “positions” rather than by “borders”. First, let us elaborate on these two concepts. A set of regulatory genomic sequence regions defined by deletion mutations, or a set of oligonucleotides shown to bind a particular transcription factor *in vitro*, constitutes a DNA sequence set defined by borders. The sequences are of defined length, and the biologist has good reason to believe that a particular sequence motif is hidden anywhere within the sequences. This is exactly the experimental scenario to which the standard formulation of the motif discovery problem applies. On the other hand, promoters exemplify a sequence type defined by position; they are defined by the location of the transcription start site (TSS), which can be mapped experimentally. Promoter motifs are supposed to occur in the vicinity of a TSS, but there is no experimental protocol that would allow delineating the sequence range within which they must occur. Computational biologists have to cope with this missing information problem in some way.

5.1. Modification of the Problem Statements

A common way to proceed in promoter analysis is to define promoters operationally as sequences extending from arbitrarily chosen distances

upstream to downstream from the TSS. For instance, Ohler *et al.*²⁵ used sequences between relative positions -60 and $+40$ for the identification of core promoter elements in *Drosophila*. A principle limitation of this approach is that it ignores a motif's specific positional distribution around the TSS, which varies widely between motifs. For instance, the eukaryotic TATA box occurs at a rather fixed distance of about $30 \text{ bp} \pm 5 \text{ bp}$ upstream from the TSS; conversely, the CCAAT box occurs within a large region of about 150 bp with a maximum at -80 (Fig. 4).³⁴ Realistic objective functions for promoter motif discovery have to account for such differences. At equal frequency, a motif predominantly occurring within a narrow distance range should be considered more significant than one that is evenly distributed over the entire promoter region considered.

Signal search analysis (SSA) is an early method that takes positional distributions of motifs into account. It is based on the concept of a locally overrepresented sequence motif, which leads to a reformulation of the motif discovery problem, as will be explained below. The input to SSA is a set of genome sequences together with a list of so-called "functional positions", e.g. a list of TSSs. The result is a locally overrepresented motif

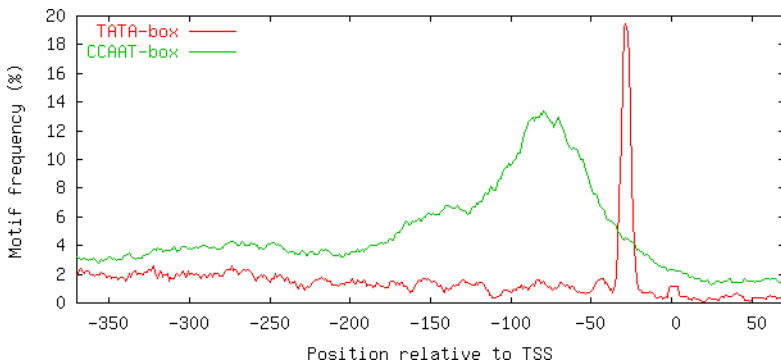


Fig. 4. Positional distributions of promoter sequence motifs around a human transcription start site (TSS). Shown are the distributions of the TATA and CCAAT boxes relative to 1867 precisely mapped TSSs from the Eukaryotic Promoter Database (EPD), release 93.³⁶ The plot is based on the weight matrices published in Bucher.³⁴ The motif frequencies were determined in overlapping windows of 20 bp for the TATA box, and 50 bp for the CCAAT box.

consisting of a motif description (consensus sequence or weight matrix + cut-off value) plus a region of preferential occurrence defined by 5' and 3' borders relative to the functional site. The key difference to the classical motif search problem statement is that the location of the motif relative to the reference position is transferred from input to output. As a consequence, the borders of the preferred region become targets for optimization and arguments of the objective function.

SSA uses a nonprobabilistic measure of local overrepresentation as an objective function for assessing motif quality. The computation of this measure is illustrated in Fig. 5. Briefly, the frequency of a given motif is determined in a series of adjacent, nonoverlapping windows of identical size, including the preferred region of occurrence as an individual window. The total length of the analyzed sequence region is chosen ad hoc.

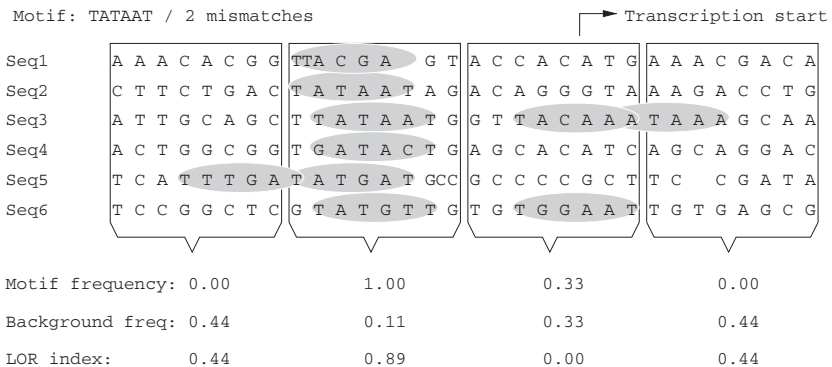


Fig. 5. Local overrepresentation — the objective function used by signal search analysis (SSA). The example sequence set consists of the six *E. coli* promoter sequences which led to the discovery of the Pribnow box motif.¹ This motif typically occurs about 10 bp upstream of the TSS. The frequency of this motif (here, defined as TATAAT with two mismatches allowed) is analyzed in a series of adjacent, nonoverlapping windows of 8 bp. The motif frequency is defined as the fraction of sequences per window that contain at least one motif instance (motifs spanning window boundaries are not counted). The background frequency for a particular window is defined as the mean of the motif frequencies in all other windows. The index of local overrepresentation (LOR) is simply the difference between the local motif frequency and the corresponding background frequency. In this example, the analyzed motif is highly overrepresented in the second window of the series, extending from relative positions -13 to -6.

The motif frequency is defined as the fraction of sequences in a window that contain at least one motif occurrence. Motifs overlapping window borders are ignored for this purpose. Local overrepresentation (*LOR*) is defined as the motif frequency within the window of preferential occurrence minus the average motif frequency determined in all other windows:

$$LOR_j = f_j - \frac{\sum_{i \neq j} f_i}{N - 1}. \quad (13)$$

Here, f_j is the motif frequency in window j , and N is the total number of windows. Note that the series of windows used to compute the background frequency needs to be adjusted to the specific region for which *LOR* is computed. The motif frequency outside the preferred regions is called background frequency, and serves the same function as the null model in the classical motif discovery framework. In fact, a major strength of SSA is its usage of a realistic null model based on natural sequences from the same genomic environment. This may explain why the weight matrices for major eukaryotic promoter elements, which were derived by this method almost 20 years ago, are still in use.

5.2. Search Algorithms for Locally Overrepresented Sequence Motifs

Two algorithms have been developed for the discovery of locally overrepresented sequence motifs, one for consensus sequence motifs³⁵ and one for weight matrices.³⁴ The former enumerates k -letter words, possibly containing free positions represented by a wildcard character and allowing a specified number of mismatches. The search space of the preferred region is defined by a preselected fixed window, with the 5' and 3' borders of the complete sequence range being taken into consideration. Using the computing power available at the time this method was conceived, the enumerative approach was possible up to a word length of about 6. To provide a heuristic search strategy for longer

motifs, the original algorithm offered the possibility of restricting motif evaluation to a random subset of k -letter words.

The algorithm for iterative refinement of a locally overrepresented sequence motif was published under the name PATOP.³⁴ The three different components of the motif description — the borders of the preferred region, the weight matrix, and the cut-off value — are updated

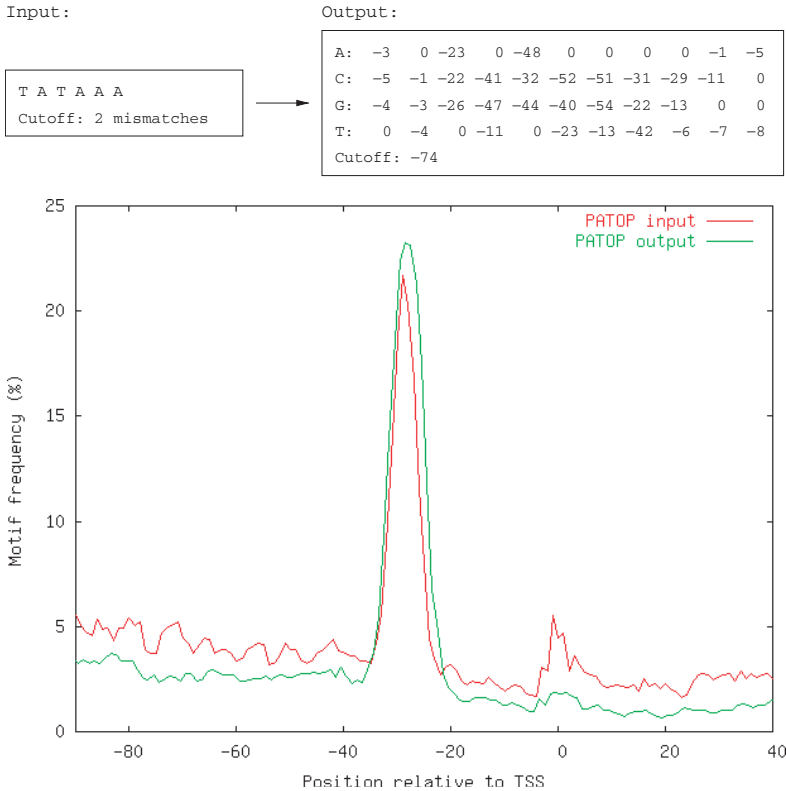


Fig. 6. Optimization of the TATA box motif by the PATOP algorithm. A new weight matrix description for the TATA box motif was optimized using local overrepresentation (LOR), as explained in Fig. 5, as an objective function. A consensus sequence-based motif description was used as the seed, and the 1867 promoter sequences from EPD (see legend for Fig. 4) served as the training set for optimization. The positional distributions of the input and output motifs are shown at the bottom. Note that the optimized weight matrix has both a higher peak frequency near position -30 and a lower background frequency elsewhere.

one at a time, in an alternating fashion. The center position and width of the preferred region, as well as the cut-off value, are optimized in an enumerative fashion based on the previously introduced objective function. User-specified upper and lower bounds and increment values define the search space of all combinations of these three parameters. The weight matrix is updated by counting the base frequencies in the current set of putative motif instances, which is unambiguously defined by the current ensemble of motif parameters. The base frequencies are converted into log-likelihood weights, as detailed in Sec. 2. A powerful additional feature of PATOP is that it can shrink or extend the length of the weight matrix on each iteration. This is achieved by including a number of additional, adjacent positions in the base frequency matrix compiled from the current motif instances. The new limits of the matrix are then defined on the basis of the observed skew of base composition in a matrix column, evaluated by a χ -squared test.

Figure 6 illustrates the effect of the refinement by PATOP with the eukaryotic TATA box as an example. The initial motif consists of the consensus sequence TATAAA with two mismatches allowed. The length of the final matrix is 11 base pairs. The DNA sequences and TSS positions used for refinement correspond to the human promoter subset of the Eukaryotic Promoter Database (EPD), release 93,³⁶ 1867 sequences in total. The plot shows the motif frequencies, evaluated in overlapping windows of width 8 and 20, respectively. In this example, the gain in local overrepresentation results from an increase in the peak signal frequency and from a decrease in the background frequency. In other words, the resulting optimized weight matrix has both higher sensitivity and higher specificity than the input consensus sequence.

6. Conclusions and Perspectives

The success of motif discovery depends, to a large extent, on the suitability of the input data. Ideally, the data set should consist of a large number of short sequences highly enriched in one particular motif, but otherwise random. In certain application areas, such data sets are available. For instance, the SAGE/SELEX technique³⁷ can produce thousands of short sequences that bind to a particular transcription factor *in vitro*.

Today, one also has access to large promoter sets defined by high-throughput methods such as expressed sequence tag (EST) sequencing of oligo-capped cDNAs³⁸ and CAGE.³⁹ However, in promoter analysis, one still faces the problem that many motifs may only be weakly over-represented and located at a highly variable distance from the TSS. An enrichment of motifs in genomic sequences can be achieved in various ways. For instance, the motif search could be targeted at specific regulatory motifs by restricting the input sequences to promoters of genes that are regulated in a particular way; see, for instance, Roth *et al.*⁴⁰ Likewise, genome-wide chromatin immunoprecipitation profiles for more than 200 transcription factors have been used to select yeast promoters that are occupied *in vivo* by a given factor in order to define its cognate binding site motifs with six different motif discovery methods.⁴¹ A presumably very accurate binding site weight matrix has been derived from over 13 000 *in vivo* mapped sites for the insulator protein CTCF.⁴² Another currently very trendy approach, exemplified in Xie *et al.*²⁶ is to restrict the motif search to sequence regions that are conserved across genomes. Specialized motif discovery algorithms, such as PhyloGibbs,⁴³ can exploit information about phylogenetic conservation contained in a multiple sequence alignment given as input to the method.

The classical weight matrix model, which assumes that motifs are of fixed length and that the contribution of individual bases at different positions are additive and independent of each other, is likely to be an oversimplification in most cases (see Benos *et al.*⁴⁴ for a critical discussion of this issue). The weight array method takes nearest-neighbor dependencies into account by scoring overlapping dinucleotides rather than individual bases.⁴⁵ Target motifs recognized by multimeric DNA-binding proteins often consist of two or more short motifs separated by spacers of slightly variable length. In fact, the Pribnow box shown in Fig. 1 is one of two conserved motifs characteristic of the major class of *E. coli* promoters. The classical EM algorithm for motif optimization is readily extensible to such bipartite, variable-length motif structures.⁴⁶ *Ab initio* discovery of composite motifs (also referred to as transcription regulatory modules), including combinations of motifs that may occur in different order and orientation, is another currently very active research direction; see Kel *et al.*⁴⁷ for an example.

A last point worth mentioning is that research on transcription factor binding sites is increasingly based on physical models of protein–DNA interactions and quantitative affinity data. Protein–binding microarrays (PBMs)⁴⁸ and related technologies⁴⁹ allow measurement of thousands of interaction energies in parallel. The TRAP model,⁵⁰ which defines the quantitative relationship between DNA sequence and PBM signal based on a weight matrix, provides a theoretical basis for inference of a binding energy matrix from quantitative affinity data. Likewise, thermodynamic modeling of the SELEX process led to the proposal of a new mathematical formula to convert base probabilities from motif discovery into base pair–protein interaction energies.⁵¹

There is definitely more work to be done in the field of DNA motif discovery.

References

1. Pribnow D. (1975) Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. *Proc Natl Acad Sci USA* **72**(3): 784–8.
2. Williams C. (1966) *Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought*. Princeton, NJ: Princeton University Press, pp. 10–11.
3. Schneider TD, Stormo GD, Gold L, Ehrenfeucht A. (1986) Information content of binding sites on nucleotide sequences. *J Mol Biol* **188**(3): 415–31.
4. Dung VV, Giao PM, Chinh NN *et al.* (1993) A new species of living bovid from Vietnam. *Nature* **363**: 443–5.
5. Berg OG, von Hippel PH. (1987) Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J Mol Biol* **193**(4): 723–50.
6. Selassie CD, Mekapati SB, Verma RP. (2002) QSAR: then and now. *Curr Top Med Chem* **2**(12): 1357–79.
7. Stormo GD, Schneider TD, Gold L, Ehrenfeucht A. (1982) Use of the ‘Perceptron’ algorithm to distinguish translational initiation sites in *E. coli*. *Nucleic Acids Res* **10**(9): 2997–3011.
8. Durbin R, Eddy S, Krogh A, Mitchison G. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge, UK: Cambridge University Press, pp. 46–79.
9. Sandve GK, Drablos F. (2006) A survey of motif discovery methods in an integrated framework. *Biol Direct* **1**: 11.
10. Das MK, Dai HK. (2007) A survey of DNA motif finding algorithms. *BMC Bioinformatics* **8**(Suppl 7): S21.

11. Cornish-Bowden A. (1985) Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Res* **13**(9): 3021–30.
12. Staden R. (1984) Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Res* **12**(1 Pt 2): 505–19.
13. Quandt K, Frech K, Karas H. (1995) MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res* **23**(23): 4878–84.
14. Bailey TL, Williams N, Misleh C, Li WW. (2006) MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res* **34**(Web Server issue): W369–73.
15. van Helden J, Andre B, Collado-Vides J. (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J Mol Biol* **281**(5): 827–42.
16. Bailey TL, Elkan C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* **2**: 28–36.
17. Pevzner PA, Borodovsky M, Mironov AA. (1989) Linguistics of nucleotide sequences. I: The significance of deviations from mean statistical characteristics and prediction of the frequencies of occurrence of words. *J Biomol Struct Dyn* **6**(5): 1013–26.
18. Attesen K. (1998) Calculating the exact probability of language-like patterns in biomolecular sequences. *Proc Int Conf Intell Syst Mol Biol* **6**: 17–24.
19. Staden R. (1989) Methods for calculating the probabilities of finding patterns in sequences. *Comput Appl Biosci* **5**(2): 89–96.
20. Queen C, Wegman MN, Korn LJ. (1982) Improvements to a program for DNA analysis: a procedure to find homologies among many sequences. *Nucleic Acids Res* **10**(1): 449–56.
21. Pavesi G, Mereghetti P, Mauri G, Pesole G. (2004) Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res* **32**(Web Server issue): W199–203.
22. Carlson JM, Chakravarty R, Khetani RS, Gross RH. (2006) Bounded search for de novo identification of degenerate cis-regulatory elements. *BMC Bioinformatics* **7**: 254.
23. Lawrence CE, Reilly AA. (1990) An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins* **7**(1): 41–51.
24. Lawrence CE, Altschul SF, Boguski MS *et al.* (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262**(5131): 208–14.
25. Ohler U, Liao GC, Niemann H, Rubin GM. (2002) Computational analysis of core promoters in the *Drosophila* genome. *Genome Biol* **3**(12): RESEARCH0087.

26. Xie X, Mikkelsen TS, Gnirke A *et al.* (2007) Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proc Natl Acad Sci USA* **104**(17): 7145–50.
27. Coward E. (1999) Shufflet: shuffling sequences while conserving the *k*-let counts. *Bioinformatics* **15**(12): 1058–9.
28. Copley RR, Ponting CP, Schultz J, Bork P. (2002) Sequence analysis of multidomain proteins: past perspectives and future directions. *Adv Protein Chem* **61**: 75–98.
29. Hu J, Li B, Kihara D. (2005) Limitations and potentials of current motif discovery algorithms. *Nucleic Acids Res* **33**(15): 4899–913.
30. Tompa M, Li N, Bailey TL *et al.* (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol* **23**(1): 137–44.
31. Wingender E. (2008) The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Brief Bioinform* **9**(4): 326–32.
32. Gama-Castro S, Jiménez-Jacinto V, Peralta-Gil M *et al.* (2008) RegulonDB (version 6.0): gene regulation model of *Escherichia coli* K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res* **36**(Database issue): D120–4.
33. Bucher P, Bryan B. (1984) Signal search analysis: a new method to localize and characterize functionally important DNA sequences. *Nucleic Acids Res* **12**(1 Pt 1): 287–305.
34. Bucher P. (1990) Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J Mol Biol* **212**(4): 563–78.
35. Bucher P, Trifonov EN. (1986) Compilation and analysis of eukaryotic POL II promoter sequences. *Nucleic Acids Res* **14**(24): 10009–26.
36. Schmid CD, Perier R, Praz V, Bucher P. (2006) EPD in its twentieth year: towards complete promoter coverage of selected model organisms. *Nucleic Acids Res* **34**(Database issue): D82–5.
37. Roulet E, Busso S, Camargo AA *et al.* (2002) High-throughput SELEX–SAGE method for quantitative modeling of transcription-factor binding sites. *Nat Biotechnol* **20**(8): 831–5.
38. Wakaguri H, Yamashita R, Suzuki Y *et al.* (2008) DBTSS: database of transcription start sites, progress report 2008. *Nucleic Acids Res* **36**(Database issue): D97–101.
39. Kawaji H, Kasukawa T, Fukuda S *et al.* (2006) CAGE Basic/Analysis Databases: the CAGE resource for comprehensive promoter analysis. *Nucleic Acids Res* **34**(Database issue): D632–6.
40. Roth FP, Hughes JD, Estep PW, Church GM. (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat Biotechnol* **16**(10): 939–45.

41. Harbison CT, Gordon DB, Lee TI *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**(7004): 99–104.
42. Kim TH, Abdullaev ZK, Smith AD *et al.* (2007) Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* **128**(6): 1231–45.
43. Siddharthan R, Siggia ED, van Nimwegen E. (2005) PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput Biol* **1**(7): e67.
44. Benos PV, Bulyk ML, Stormo GD. (2002) Additivity in protein–DNA interactions: how good an approximation is it? *Nucleic Acids Res* **30**(20): 4442–51.
45. Zhang MQ, Marr TG. (1993) A weight array method for splicing signal analysis. *Comput Appl Biosci* **9**(5): 499–509.
46. Cardon LR, Stormo GD. (1992) Expectation maximization algorithm for identifying protein-binding sites with variable lengths from unaligned DNA fragments. *J Mol Biol* **223**(1): 159–70.
47. Kel A, Konovalova T, Waleev T *et al.* (2006) Composite Module Analyst: a fitness-based tool for identification of transcription factor binding site combinations. *Bioinformatics* **22**(10): 1190–7.
48. Berger MF, Bulyk ML. (2006) Protein binding microarrays (PBMs) for rapid, high-throughput characterization of the sequence specificities of DNA binding proteins. *Methods Mol Biol* **338**: 245–60.
49. Maerkl SJ, Quake SR. (2007) A systems approach to measuring the binding energy landscapes of transcription factors. *Science* **315**(5809): 233–7.
50. Roeder HG, Kanhere A, Manke T, Vingron M. (2007) Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics* **23**(2): 134–41.
51. Djordjevic M. (2007) SELEX experiments: new prospects, applications and data analysis in inferring regulatory pathways. *Biomol Eng* **24**(2): 179–89.