In Silico Biology 9 (2009) S17–S39
DOI 10.3233/ISB-2009-0381
IOS Press

Extracting Signature Motifs from Promoter Sets of Differentially Expressed Genes

Ieva Mitašiūnaitė[a], Christophe Rigotti[a], Stéphane Schicklin[b], Laurène Meyniel[a], Jean-François Boulicaut[a] and Olivier Gandrillon[b,*]

[a]*Université de Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, F-69621, Villeurbanne, France*
[b]*Université Lyon 1, Centre de Génétique Moléculaire et Cellulaire, UMR5534, F-69622, Villeurbanne, France*

Edited by E. Wingender; received 22 July 2008; revised 23 October 2008; accepted 25 October 2008; published 13 December 2008

ABSTRACT: There is a critical need for new and efficient computational methods aimed at discovering putative transcription factor binding sites (TFBSs) in promoter sequences. Among the existing methods, two families can be distinguished: statistical or stochastic approaches, and combinatorial approaches. Here we focus on a complete approach incorporating a combinatorial exhaustive motif extraction, together with a statistical Twilight Zone Indicator (TZI), in two datasets: a positive set and a negative one, which represents the result of a classical differential expression experiment. Our approach relies on the existence of prior biological information in the form of two sets of promoters of differentially expressed genes. We describe the complete procedure used for extracting either exact or degenerated motifs, ranking these motifs, and finding their known related TFBSs. We exemplify this approach using two different sets of promoters. The first set consists in promoters of genes either repressed or not by the transforming form of the v-*erbA* oncogene. The second set consists in genes the expression of which varies between self-renewing and differentiating progenitors. The biological meaning of the found TFBSs is discussed and, for one TF, its biological involvement is demonstrated. This study therefore illustrates the power of using relevant biological information, in the form of a set of differentially expressed genes that is a classical outcome in most of transcriptomics studies. This allows to severely reduce the search space and to design an adapted statistical indicator. Taken together, this allows the biologist to concentrate on a small number of putatively interesting TFs.

KEYWORDS: Promoter, differential expression, complete pattern extraction, transcription factor, transcription factor binding site, twilight zone, extraction parameter tuning, exact matching pattern, soft matching pattern

INTRODUCTION

To understand the regulation of gene expression remains one of the major challenges in molecular biology. One of the elements through which the regulation works is the initiation of the transcription by the interaction between gene promoter elements at the level of DNA sequence and multiple activator and repressor proteins called transcription factors (TFs). This interaction occurs when a TF binds on its binding site on a gene promoter. Numerous efforts have given rise to a variety of computational methods to discover putative transcription factor binding sites (TFBSs) in sets of promoters of co-regulated genes. Among them two families can be distinguished: statistical or stochastic approaches, and combinatorial approaches [1].

*Corresponding author. E-mail: gandrillon@cgmc.univ-lyon1.fr.

Electronic publication can be found in *In Silico Biol.* **8**, 0043 <http://www.bioinfo.de/isb/2008/08/0043/>, 13 December 2008.

Concerning the family of statistical and stochastic approaches, a recent review of the most widely used algorithms exhibits rather limited results [2], and concludes to the necessity to go on exploring alternative methods. There are several reasons for their limited success, but it seems that the difficulty to separate the patterns from the *random background* is among the principal ones. Statistical methods make hypothesis about the distribution models and assumptions for computational as well as statistical reasons, but no one knows the correct stochastic process that nature uses, and what is the *biological randomness*. Moreover, this stochastic process seems to be different from species to species: many tools perform much better on the yeast datasets than on other species [2,3]. In addition to this, considering the employed measures of interest, statistical significance is very dependent on the choice of the length of the promoter sequences: considering longer promoters would allow to identify regulatory elements located further upstream, but conversely then random motifs become statistically as significant as the regulatory elements [4].

In this paper we focus on the family of combinatorial approaches that aims at an exhaustive motif extraction without *a priori* hypothesis on the underlying stochastic process. Exhaustive algorithms enumerate all objects they were built to find. According to [5], probably the best tools for finding consensus-based motifs in DNA sequences are the pattern-driven algorithms that test all the 4^L different patterns of L letters, score each pattern by the number of approximate occurrences and find the high-scoring patterns. The exhaustive search through all these 4^L patterns becomes impractical for large L, but the length of binding sites in promoter sequences is estimated to be between 5 and 15 base-pairs (bp) [6] and the mean of these lengths in TRANSFAC® [7] is 14.3 bp with standard deviation 4.7 bp [8]. These rather reasonable values of L turn the search to be tractable in practice. However, the exhaustive methods are often not selective enough to discriminate true sites from false positives, and thus, because of the large number of patterns obtained, the user has to rank them by different statistical measures of interest computed under different hypothesis. An effort on developing exhaustive and optimal approaches (i.e., with guarantee to find all the patterns having the highest or demanded fitness values) for the discovery of patterns in biosequences has resulted in a number of algorithms to search for putative TFBS, e.g., [9–15] (a systematized survey of main algorithmic ideas can be found in [16]). In practice, they all required some *fitness measure* used as a ranking and/or a selection criterion to help an user to differentiate the true positive from the false ones. Many different measures have been proposed, e.g., statistical significance [10,15], information content [11,17], ratio of the score of a pattern in a positive dataset divided by the score of the same pattern in a random dataset [14]. The approach of [3] takes one step further and, after having ranked the extracted patterns according to a measure of fitness, use the most significant ones as the seeds to build the motifs modelling the TFBSs (in the concrete, the position specific scoring matrices (PSSMs)).

Having in mind the difficulties to model statistically the biological randomness, we propose to postpone the phase of significant pattern selection, based on a statistical measure, and to use beforehand the supplementary biological information to constrain the search and reduce the number of extracted patterns. This additional information comes in the form of a second dataset representing an *opposite biological situation*. To collect this information, the method starts with a classical operation used in molecular biology: the search for differentially expressed genes.[1] This allows to obtain two groups of genes from which one can derive two sets of promoters. To look for putative TFBSs regulating the overexpressed genes, we choose the first set (the promoters of the over-expressed genes) to be used as a positive set,

[1] It consists in comparing two biological situations, A and B, in order to obtain two groups of genes: one that is up-regulated, and the other one that is down-regulated, when going from A to B.

and the second set as a negative one.[2] Then our method consists in finding the patterns occurring on at least $\alpha_{\min}$ promoters from the positive set and on at most $\alpha_{\max}$ promoters from the negative set, where the parameter $\alpha_{\min}$ (resp. $\alpha_{\max}$) is supposed to be a large (resp. small) threshold value. The originality of the proposed method w.r.t the other combinatorial algorithms, which allow to extract patterns from several datasets (e.g., SPEXS [14] or DRIM [18]), is that the maximal support threshold is set explicitly. This is particularly interesting, when there is a clear semantic cut between positive and negative datasets, and the negative dataset has an opposite biological sense (presence/absence of a mutation; addition or not of a given drug, etc.), and does not just represent random background. Two kinds of patterns are handled by our method: patterns having exact matches in the sequences, called Exact Matching Patterns (EMPs) and patterns having approximate matches within a maximum Hamming distance, called Soft Matching Patterns (SMPs). Interestingly, in both cases, the enrichment of the pattern discovery context, using a negative dataset, reduces the size of the solution set by several orders of magnitude. Even then, the set of the extracted patterns remains large, and thus we develop a set of complementary solutions to help to tune the parameters in order to focus on a manageable and potentially interesting set of patterns. In particular, we use a notion of rising patterns, and we rank/select the exceptional patterns according to a measure called Twilight Zone Indicator (TZI), based on *subtlety* [4]: a pattern M is considered to be subtle if we expect that some random patterns could occur at least as often as M in the positive dataset and at the same time no more often than M in the negative dataset. In the case of SMPs, we also cluster the patterns (hierarchical clustering) and compute the consensus pattern of each cluster of SMPs using a multiple alignment tool. Then, for both EMPs and consensus SMPs, we verify which patterns are known TFBSs in the TRANSFAC® database. Identification of the TFs that can bind on the patterns specific to the positive dataset can help to discover new regulators of the concerned biological process. Patterns that do not correspond to known TFBSs are equally interesting since they can be unknown elements of regulation.

METHODS

The pattern discovery tool Marguerite

To extract the patterns from gene promoter sequences, we use a tool called *Marguerite*. This tool performs a differential extraction of patterns between two sets of promoter sequences: set D^+, a dataset representing a positive situation, and D^-, a dataset representing a negative one. To run an extraction, the user has to set the four following constraints: $L_{\min}$ the minimal length of the patterns, $L_{\max}$ their maximal length, $\alpha_{\min}$ their minimal support in D^+, and $\alpha_{\max}$ their maximal support in D^-, where the support of a pattern in a dataset D is simply the number of sequences in D containing at least one occurrence of the pattern. *Marguerite* is complete in the sense that it finds all possible patterns satisfying the constraints according to the user setting.

Through minimum support constraint on D^+ and maximum support constraint on D^-, *Marguerite* enables differential extractions. Another approach to perform differential extractions is to use only a minimum support constraint on D^+ and to filter the patterns according to a score, which is a ratio of sequences, containing a pattern in D^+, divided by a ratio of sequences, containing a pattern in a D^- [14].

[2] Notice that to characterize the promoters of the repressed genes, one simply has to choose the repressed genes as the positive set.

These two approaches are not equivalent. The latter one is convenient, when D^- is a random dataset, since it picks out the patterns that are overrepresented in D^+ w.r.t. D^-. However, when D^- represents an opposite biological situation, and we want to extract patterns that are not implied in a biological process of that situation, we need to explicitly push an upper bound for pattern support in D^-.

It should be noticed that other existing tools to perform complete differential extractions, such as SPEXS [14], could be used in the process presented in this paper. In this case, the whole process remains unchanged, except that if different constraints are used, then we may have to change the way the TZI measure (see section "Twilight Zone Indicator") is computed. For instance, in the case of SPEXS, we have a slightly different parameter space, and the TZI needs to be adapted accordingly.

Marguerite can be used to compute both Exact Matching Patterns (EMPs) and Soft Matching Patterns (SMPs). Moreover, it does not use a predefined alphabet, and can, for instance, be used on sequences containing extra symbols, like the symbol N to indicate undefined bases. For EMPs, the support of a pattern M is the number of sequences containing at least one exact occurrence of M, while in the case of SMPs this is the number of sequences containing at least one *soft occurrence* of M (i.e., at least one approximated occurrence of M). A substring S is termed a soft occurrence of a pattern M if their Hamming distance (i.e., the number of substitutions necessary to obtain M from S) is at most α_{dist}, where α_{dist} is a user-specified threshold. When α_{min}, α_{max} and α_{dist} are given, *Marguerite* finds all SMP patterns M such that: (1) M has at least α_{min} soft occurrences in D^+, (2) M has at most α_{max} soft occurrences in D^-, and (3) M has at least one exact occurrence in D^+ (i.e., M occurs at least one time without modification in the positive dataset).

Marguerite is based on the generic algorithm FAVST [19], designed for the efficient extraction of strings under combination of constraints, taking advantage of the so called Version Space Tree (VST) [20] data structure. *Marguerite* [21,22] extends FAVST to degenerated patterns discovery through similarity and soft-support constraints. It is implemented in C/C++ and compiled for GNU/Linux, Mac OS and Windows operating systems. It is available upon request to the authors. On a MAC OS platform (Intel 2 Ghz processor, 1Gb of RAM), for the extractions reported in this study, the extraction times ranged from a few seconds (in the case of EMPs) to a few tens of minutes (for SMPs).

Procedure to find rising patterns

Let T_{min} (resp. T_{max}) be a set of possible values for α_{min} (resp. α_{max}) ordered by increasing values. Finding the rising patterns in the parameter space $T_{\mathrm{min}} \times T_{\mathrm{max}}$, is performed as follows:

1. Let α_{min} (resp. α_{max}) be the first element in T_{min} (resp. T_{max}).
2. Let S_p be the set of patterns obtained when running an extraction under the conjunction of constraints α_{min} and α_{max}.
3. If S_p is empty and α_{max} is not the last element in T_{max} then set α_{max} to the next value in T_{max}. Goto step (2).
4. Output S_p as a set of rising patterns.
5. If α_{min} is not the last element in T_{min} then set α_{min} to the next value in T_{min}. Goto step (2).

It should be noticed that if the set S is empty for a conjunction of constraints α_{min} and α_{max}, then S is also empty for any conjunction α'_{min} and α_{max} where $\alpha'_{\mathrm{min}} \geqslant \alpha_{\mathrm{min}}$. The procedure avoids the test of such useless conjunctions.

A more formal view of finding rising patterns would be to consider it as a multi-objective optimization problem [23]: maximizing α_{min}, minimizing α_{max}, under the constraint $N > 0$, where N is the number of patterns satisfying the α_{min} and α_{max} thresholds. However, in practice, a pure maximization of α_{min}

is too restrictive, because a value of $\alpha_{\min}$, that is slightly lesser than an optimal one, can lead to a few more patterns, and can also be interesting. Thus, we consider as rising patterns (in their definition and in the procedure to find them) the points in the parameter space that are solutions of this optimization problem (the Pareto optimal set), and also points that are suboptimal solutions (for each $\alpha_{\min}$ value we find the minimal $\alpha_{\max}$ such that $N > 0$) and extract the patterns for all these points in the parameter space.

Twilight Zone Indicator

The notion of twilight zone (TZ) [4] has been originally proposed to characterize the *subtle motifs*, i.e., motifs that can not be distinguished (in the statistical sense) from random patterns (patterns due to the random background). In this context the TZ was defined as the set of values of the scoring function for which we can expect to have some random patterns exhibiting such score values. Let us consider the notion of extraction parameters in a broad sense, including structural properties of the dataset (e.g., number of sequences, length of the sequences) and extraction constraints (e.g., selection threshold according to one or several measures, length of the patterns). Then, the TZ can be seen as a region (or set of regions) in the parameter space, where we are likely to obtain random patterns among the extracted patterns, these random pattern having scores as good (or even better) than the *true* patterns.

Having this view in mind, we define a Twilight Zone Indicator (TZI) to rank the patterns in the case of differential extractions. Let M be a pattern, occurring in $\textit{support}^+(M)$ sequences of the positive dataset, and in $\textit{support}^-(M)$ sequences of the negative dataset. Then, $\textit{TZI}(M)$ is an estimate of the number of random patterns, having the same length as M, that will be extracted using $\alpha_{\min} = \textit{support}^+(M)$ and $\alpha_{\max} = \textit{support}^-(M)$, i.e., using the most selective constraints that still permit to obtain M (since for larger $\alpha_{\min}$ and/or lower $\alpha_{\max}$ threshold values, M will not satisfy the constraints and will not be retained during the extraction). The higher is $\textit{TZI}(M)$, the *deeper* is M in the twilight zone.

We consider that all the sequences have the same length, denoted G. In this context, we want to estimate the number of SMP patterns of length L that will be extracted under the thresholds $\alpha_{\min}$, $\alpha_{\max}$ and α_{dist} (see section "The pattern discovery tool *Marguerite*", in Methods). Notice that estimating the number of EMPs is a particular case, where α_{dist} is set to 0. As in [4], we suppose that the data sequences are composed of independent and uniformly distributed symbols, having the same occurrence probability, and that the overlapping of the occurrences of the patterns has a negligible impact on the number of patterns extracted (since $L << G$). Additionally, we suppose that the two datasets are independent.

Occurrences at a given position

The data sequences are gene promoter sequences composed of 4 symbols. Then, there are 4^L different possible strings of length L. The hypotheses made on the distribution of the symbols imply that the probability that a pattern M of length L has an exact occurrence starting at a given position in a sequence [3] is $P(\textit{exact occ. of M at one position}) = 1/4^L$.

From an exact occurrence of M, one can construct the soft occurrences of M within a Hamming distance α_{dist} by placing k substitutions in $\binom{L}{k}$ possible ways, with $k \in \{0, \ldots, \alpha_{\text{dist}}\}$. Since we have 4 symbols, then for each position where we have a substitution, we have 3 different possible substitutions.

[3] Except the last $L - 1$ positions.

Thus, for a pattern M, there are $\Sigma_{k=0}^{\alpha_{\text{dist}}}\binom{L}{k} \times 3^k$ strings that are soft occurrences of M. Then, the probability that a pattern has a soft occurrence starting at a given position in a sequence is P (*soft occ. of M at one position*) $= \frac{\Sigma_{k=0}^{\alpha_{\text{dist}}}\binom{L}{k}\times 3^k}{4^L}$. In the following, we also need the probability that a pattern M has a *strict* soft occurrence starting at a given position (a *strict* soft occurrence of M, is a soft occurrence of M that is not an exact occurrence). In this case we have simply P(*strict soft occ. of M at one position*) $= \frac{\Sigma_{k=1}^{\alpha_{\text{dist}}}\binom{L}{k}\times 3^k}{4^L}$.

Occurrences in a random sequence

In a sequence there are $(G - L + 1)$ possible positions to place the beginning of an occurrence of M. Since $L << G$, for the sake of simplicity we approximate a number of possible positions by G. Then, the probability that there is no soft occurrence of M in a random sequence is P(*no soft occ. of M in a seq.*) $= (1 - P$(*soft occ. of M at one position*)$)^G$. Thus, the probability that there is at least one soft occurrence of M in a sequence is P(*exists soft occ. of M in a seq.*) $= 1 - (1 - P$(*soft occ. of M at one position*)$)^G$. Similarly, the probability that there is at least one strict soft occurrence of M is P(*exists strict soft occ. of M in a seq.*) $= 1 - (1 - P$(*strict soft occ. of M at one position*)$)^G$, and the probability that there is at least one exact occurrence is P(*exists exact occ. of M in a seq.*) $= 1 - (1 - 1/4^L)^G$.

Minimum support constraint

To determine P(*M sat. min. supp.*), i.e., the probability of M to satisfy the minimum support constraint, let us define X as the number of sequences, in the positive dataset, that contains at least one exact occurrence of M. The probability P(*M sat. min. supp.*) can be decomposed using the conditional probability of *M sat. min. supp.* given the value of X, as follows:

$$P(\textit{M sat. min. supp.}) = \sum_{i=1}^{N^+}(P(X = i) \times P\,(\textit{M sat. min. supp.}|X = i)) \quad (1)$$

where N^+ is the number of sequences in the positive dataset. Notice that the sum starts at $i = 1$, and not at $i = 0$, since the pattern must have at least one exact occurrence in the positive dataset (see section "The pattern discovery tool *Marguerite*", in Methods).

The variable X follows a binomial distribution $B(N^+, P$(*exists exact occ. of M in a seq.*)$)$, thus we have: $P(X = i) = \binom{N^+}{i} \times P$(*exists exact occ. of M in a seq.*)$^i\times$ (1 – P(*exists exact occ. of M in a seq.*))$^{N^+-i}$.

P(*M sat. min. supp.*$|\ X = i)$ is the probability that M satisfies the minimum support constraint, given that exactly i sequences contain at least one exact occurrence of M. This also means that $(N^+ - i)$ sequences do not have any exact occurrence of a pattern. Then according to i there are two cases:

1. If $i \geqslant \alpha_{\min}$ then P(*M sat. min. supp.*$|X = i)) = 1$ since the constraint is already satisfied by the i sequences that contain each at least one exact occurrence of M.
2. If $i < \alpha_{\min}$ then P(*M sat. min. supp.*$|X = i)$ is equal to the probability that at least $(\alpha_{\min} - i)$ of the $(N^+ - i)$ remaining sequences contain at least one strict soft occurrence. This number of sequences that contain at least one strict soft occurrence of M also follows a binomial distribution

$B(N^+ - i,$ P(*exists strict soft occ. of M in a seq.*)). Then we have:

$$P(\textit{M sat. min. supp.} \mid X = i) = \sum_{z=\alpha_{\min}-i}^{N^+-i} \left(\binom{N^+ - i}{z} \times P\,(\textit{exists strict soft occ. of M in a seq.})^z \times (1 - P\,(\textit{exists strict soft occ. of M in a seq.}))^{N^+-i-z} \right)$$

Thus, we can obtain P(*M sat. min. supp.*) by computing the sum in equation 1 and P(*M sat. min. supp.*$|X = i$) according to the two cases above.

Maximum Support constraint

Let Y be the number of sequences that support M in the negative dataset. A pattern M satisfies the maximum support constraint with threshold $\alpha_{\max}$ if $Y \leqslant \alpha_{\max}$. The variable Y follows a binomial distribution $B(N^-,$ P(*exists soft occ. of M in a seq.*)), where N^- is the number of sequences in the negative dataset. Then the probability that M satisfies the maximum support constraint is P(*M sat. max. supp.*) $= \sum_{z=0}^{\alpha_{\max}} \binom{N^-}{z} \times P$(*exists soft occ. of M in a seq.*)$^z \times (1 - P$(*exists soft occ. of M in a seq.*))$^{N^- - z}$.

Conjunction of Minimum Support and Maximum Support constraints

Given our hypothesis that the positive and negative datasets are independent, the probability that a pattern satisfies a conjunction of minimum support and maximum support constraints is P(*M sat. min. and max. supp.*) $= P$(*M sat. min. supp.*) $\times P$(*M sat. max. supp.*).

Number of expected patterns and Twilight Zone Indicator

Let *ENP* (L, $\alpha_{\min}$, $\alpha_{\max}$, α_{dist}) be the Expected Number of Patterns of length L that will be extracted under the thresholds $\alpha_{\min}$, $\alpha_{\max}$ and α_{dist}. Since there are 4^L possible patterns of length L, and from the hypothesis that the overlapping of the occurrences of the patterns has a negligible impact on the number of patterns extracted, we can approximate *ENP*(L, $\alpha_{\min}$, $\alpha_{\max}$, α_{dist}) by P(*M sat. min. and max. supp.*) $\times\, 4^L$.

Finally, let M be a pattern, occurring in *support*$^+(M)$ sequences of the positive dataset, and in *support*$^-(M)$ sequences of the negative dataset for a given α_{dist} threshold. Then, *TZI*(M) is defined as *ENP*(*length*(M), *support*$^+(M)$, *support*$^-(M)$, α_{dist}).

TRANSFAC®, Patchlike

Patch$^{\text{TM}}$ is a tool integrated in TRANSFAC® which identifies known TFBSs in a given sequence. One can verify whether the extracted patterns are known TFBSs by supplying them as input to Patch$^{\text{TM}}$. The database TRANSFAC® is distributed in plain text files altogether with a graphic interface written in Perl CGI and C programs that implements various functionalities, including Patch$^{\text{TM}}$.

To analyse the biological sequences we use a Macintosh platform. The programs coming with TRANSFAC® are platform dependent, and unfortunately there is neither compiled distribution for the Macintosh platform nor source code available. Thus, we developed a tool called Patchlike written in Perl which takes a collection of sequences as input (in our case, these sequences are the extracted patterns) and searches in these sequences for the TFBSs listed in the TRANSFAC® data files. It means that Patchlike

	Alignment	*TZI*
	.CGGCCGTT...	23.94
	.GCGCCGTT...	0.68
	...GCCGTTAT.	4.4
	....CCGTTCGT	4.4
	...GCCGTTCG.	23.75
	....CCGTTAGG	0.68
	TTGGCCGT....	23.75
	...GCCGTAAC.	107.37
	..TGCCGTAA..	0.58
Consensus	...gCCGTt...	
Transfac:	c-Myb-isoform1	
Mean of TZI:	21.06	

Fig. 1. Centroids of the clusters of SMPs and the consensus of these centroids computed by a multiple alignment.
The SMPs used are patterns of size 8 obtained by differential extraction on two sets of promoter sequences: 29 promoter sequences of genes repressed by the v-ErbA oncogene in the positive dataset, and 21 promoter sequences of genes activated by v-ErbA in the negative dataset. The 9 centroids (also of size 8) are the centroids of the clusters of these SMPs, obtained by hierarchical clustering. The consensus corresponds to a binding site of the TF c-Myb-isomorf1. A base is weakly conserved if it is shared by at least 50% of the patterns and highly conserved if it is shared by at least 90% of the patterns (out of 9 centroids, it means that the base must be conserved in all centroids). Bases that are highly conserved appear in red in the patterns and as uppercase letters in the consensus. Bases that are weakly conserved appear in green in the patterns and as lowercase letters in the consensus. Not conserved bases appear in black in the patterns. Positions with no conserved bases are indicated by dots in the consensus. Dots on the left and on the right of the centroids have no particular meaning (centroids are of size 8).

not only mimics but also serializes the search that can be made with Patch$^{\mathrm{TM}}$. For a given sequence M, Patchlike searches through the TFBSs of the TRANSFAC® data files, to find the ones that are contained in M (equal to M or that are substring of M). In Patchlike we only use the vertebrate data files, and do not allow mismatches between an input pattern and a TFBS. However, such searches can result in a quite large number of TFBSs that might burden the analysis, so we retrieve only the longest TFBS that are contained in M. Finally, Patchlike considers an input sequence and its reverse complement to look for the TFBSs in forward and reverse direction.

Hierarchical clustering of SMPs

The hierarchical clustering of the SMP patterns is performed using the *hclust* function of the package *stat* of the R environment [24]. The proximity between clusters is computed using the complete linkage method. To construct a distance matrix we estimate the dissimilarity of each pair of SMPs as follows. For each pair $< M_1, M_2 >$ we compute its optimal pairwise global alignment [25] with the following parameters: the score for a mismatch is 1, the score for a match is 0, the insertions and deletions inside an alignment are not allowed, the terminal gaps are not penalised, and the length of an alignment (terminal gaps are not included in the alignment length) must be at least a half of the shortest pattern in the pair (i.e., must be of size greater or equal to *min*(*length*(M_1), *length*(M_2))/2). Then, the dissimilarity of the pair of SMPs is computed as the score of the best alignment divided by its length. To ameliorate the quality and efficiency of the clustering we process the SMPs by groups of patterns having the same length.

Finding a consensus pattern of a cluster

To find the consensus pattern of each cluster of SMPs we align the patterns in each cluster using the multiple alignment tool MultAlin [26]. We use the following alignment scoring parameters: gap creation and extension penalty is -5, terminal gaps are not penalized, score for a match is 2, and score for any mismatch is 0. Once a consensus SMP is computed we can use Patchlike or consult TRANSFAC® to check whether it is a known TFBS. Figure 1 gives an example of a cluster whose consensus SMP is a binding site of the TF c-Myb-isomorf1.

Data selection

The promoter sequences were obtained as presented in [27]. We refer the reader to this previous work for the details. If the sequences are too short, or if we have a very small number of sequences then many TFBSs will be absent from the data, or poorly represented, and the random background itself will be underrepresented. As this can only degrade the result of differential extractions, we selected the largest datasets available. However, the datasets should not be too large, in the sense that they must remain specific to TFBSs locations and to biological situations. Thus the selected genes (29 in the positive dataset and 21 in the negative one, out of the 110 differentially expressed genes, for the main study reported in section "Results and discussion") are chosen because they are known to play a role in the corresponding biological situations. Concerning the parts of the genes retained to form the sequences, we should avoid to incorporate in the datasets portions that are not likely to contain binding sites, thus we selected sequences composed of 3000 bp upstream and 1000 bp downstream, a portion known to be rich in TFBSs [28].

RESULTS

Self-renewal is a characteristic property of stem cells, the molecular basis of which is still elusive. Deregulation of this process occurs frequently during cancer generation. We have decided to investigate this question through the discovery of differentially expressed genes by using the SAGE technique [29] on our primary model of T2ECs, that are normal chicken erythroid progenitors [30]. These cells can be induced to self-renew or to differentiate when normal. The expression of the v-*erbA* oncogene induces transformation by blocking their differentiation process [31]. We therefore decided to identify v-ErbA target genes responsible for the transformation process induced by v-ErbA. For this we have compared the transcriptome of T2ECs expressing an oncogenic form of v-ErbA with the transcriptome of T2ECs expressing the S61G mutant of v-ErbA. This mutant is defective in its ability to inhibit differentiation and to induce erythroid transformation [32]. Thus, the comparison between the transcriptome of cells expressing either the transforming form of v-ErbA or the S61G mutant of v-ErbA allowed us to generate a list of 110 differentially expressed genes between these two conditions [27]. We used this v-ErbA dataset throughout the present paper to exemplify the potential of our motif discovery method. In order to assess the generality of our approach, we also applied our method to a second data set, made from the promoters of genes showing differential expression, as assessed by SAGE, between self-renewing and differentiating erythroid progenitors [33].

Taking in account biological information in combinatorial pattern extraction

This work is based on the hypothesis that the transforming activity of v-ErbA arises from the repression of a set of genes and that at least some of these genes share some regulating TFs, which will be absent from most genes activated by v-ErbA. The motivation underlying the development of the method presented here is to help to discover the TFs that participate in the v-ErbA-induced transformation process. A classical approach would consist of identifying the genes repressed by v-ErbA and then extracting the putative TFBSs that are the patterns shared by the promoter sequences of these genes. The problem is that a combinatorial pattern extraction in such a context results in a large solution set containing many false positives. It is then very hard to pick out manually true positives from such a plethora of extracted patterns. Our approach is to first refine a pattern extraction task by introducing a negative dataset that represents the opposite biological situation and thereby reduce the number of false positives.

In order to find signature motifs for v-ErbA target genes, we first created two sets of promoter sequences of differentially expressed genes: (1) a dataset denoted R, for the genes *repressed* by the v-*erbA* oncogene, composed of 29 promoters, and (2) a dataset denoted A, for the genes *activated* by v-ErbA, composed of 21 promoters. Promoter sequences were extracted as previously described in [27], and are composed of 4000 bp (3000 bp upstream and 1000 bp downstream). Datasets R and A represent two biologically opposite situations of interest, and are used respectively as the positive and the negative dataset. *A priori* interesting patterns are strings that have many occurrences in the positive dataset but only a few in the negative dataset. We focus the search on putative TFBSs that could be used to regulate the transcription of the genes of the positive dataset while they are not likely to have an important impact on the regulation of the genes of the other set.[4] We extract the patterns by using a *minimum support* threshold $\alpha_{\min}$ (i.e., at least $\alpha_{\min}$ sequences must contain occurrences of the pattern) on the positive dataset and a *maximum support* threshold $\alpha_{\max}$ (i.e., at most $\alpha_{\max}$ sequences can contain occurrences of the pattern) on the negative dataset. We consider two kinds of patterns: Exact Matching Patterns (EMPs) and Soft Matching Patterns (SMPs). Both are strings of bases, but they differ in the way their supports are defined. The support of an EMP in a dataset is the number of sequences of the dataset that contain at least one exact occurrence of this EMP. Let α_{dist} be a given threshold, termed the *soft matching* threshold, then the support of a SMP is the number of sequences containing at least one soft occurrence of the pattern, where a soft occurrence is a part of the sequence different from the pattern in at most α_{dist} positions (i.e. the Hamming distance between this part of the sequence and the pattern is less or equal to α_{dist}). Both SMPs and EMPs are necessary: SMPs allow to gather the degenerated TFBSs while EMPs are dedicated to pick out the conserved ones. The two kinds of patterns are extracted using the tool *Marguerite* (see section "The pattern discovery tool *Marguerite*", in Methods). Figure 2 shows the reduction in the number of patterns when using two datasets (positive and negative) instead of using the positive one only.

Ranking patterns by a Twilight Zone Indicator

Even if refining the extraction context with a negative dataset, corresponding to the opposite biological situation, reduces the number of extracted patterns up to several orders of magnitude, there are still too

[4] However, we keep in mind that a single TF might be both an activator in the positive set and inhibitor in the negative one (the role of a TF can be determined by the molecular context around its binding site), but in this case we can detect its influence only if it has a TFBS in the positive dataset different from its TFBS in the negative dataset.

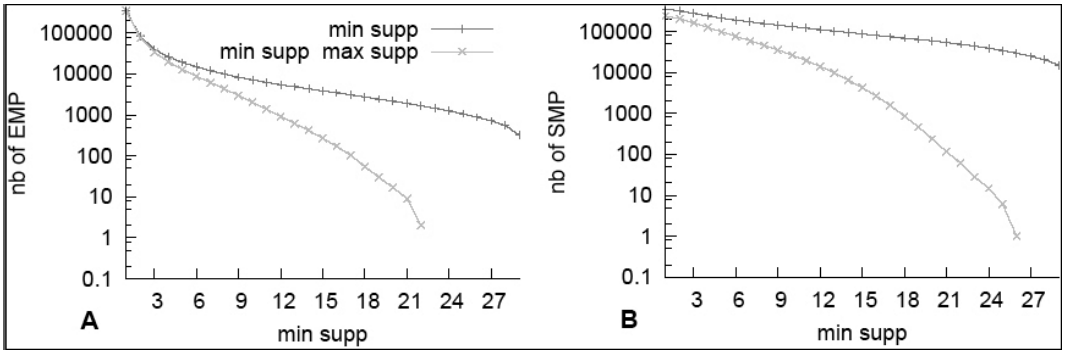

Fig. 2. Reduction in the number of patterns when using two datasets (positive and negative) instead of using the positive one only. Graph A (resp. B) gives the number of EMPs (resp. SMPs) satisfying only a minimum support constraint in the positive dataset R w.r.t. the number of EMPs (resp. SMPs) satisfying both a minimum support constraint in the positive dataset R and a maximum support constraint in the negative dataset A. The plots represent the number of patterns of length from 5 to 11, extracted when the minimum support varies from 1 to 29, and the maximum support is set to 7. In the case of SMPs the allowed Hamming distance is set to 1.

many of them to be verified manually. A classical approach is to associate to the patterns a measure of interest and select those having the most relevant measure value. In order to assess the significance of a pattern we used the notion of Twilight Zone (TZ) [4] to build a Twilight Zone Indicator (TZI). A zone TZ is a zone in a parameter space, where we are likely to obtain patterns produced by the random background. Let M be a pattern of length *length*(M), occurring in *support*$^+(M)$ sequences of the positive dataset and in *support*$^-(M)$ sequences of the negative dataset. Then *TZI*(M) is an estimate of the minimum number of patterns of length *length*(M) due to the random background, that will be extracted together with M. This minimum value is obtained in the most stringent conditions (i.e., with the strongest constraints) that still lead to the extraction of M. These conditions are obtained when we choose $\alpha_{\min} =$ *support*$^+(M)$ and $\alpha_{\max} =$ *support*$^-(M)$. The computation of the TZI is detailed in section "Twilight Zone Indicator", in Methods. It is based on the same hypothesis made in [4]: the data sequences are composed of independent and uniformly distributed nucleotides, and the possible overlapping of the occurrences of the patterns is considered to have a limited impact on the number of extracted patterns. In addition, we suppose that the positive and the negative datasets are independent.

To validate empirically the computation of the TZI, i.e., of the expected number of patterns, we compared it with the number of patterns extracted from random datasets and biological datasets R and A. Two pairs of random datasets that mimic the biological datasets R and A were constructed using the tool RanDNA [34]: (1) pair (R*,A*), where R* (resp. A*) have the same number of sequences and the same total number of nucleotides per sequence as R (resp. A), and are built from independent and uniformly distributed nucleotides with a homogeneous nucleotide composition (i.e., 25% of A, C, G and T); (2) pair (R**,A**), where R** and A** are generated using the same constraints as (R*,A*), except for the relative nucleotide composition. For R** (resp. A**) the relative nucleotide composition is the same as the one of R (resp. A). Moreover, the same sequencing uncertainties (N regions)[5] as in R (resp. A) have been implanted in R** (resp. A**). The graphs in Fig. 3 depict the number of patterns extracted with parameter settings chosen to be pertinent w.r.t. our biological problem.[6] The value of $\alpha_{\min}$ varies

[5]Notice that *Marguerite* does not require a predefined αbet, and can therefore handle sequences containing undefined bases, denoted by the symbol N.

[6]These settings are also representative, and other settings lead to a similar global behavior.

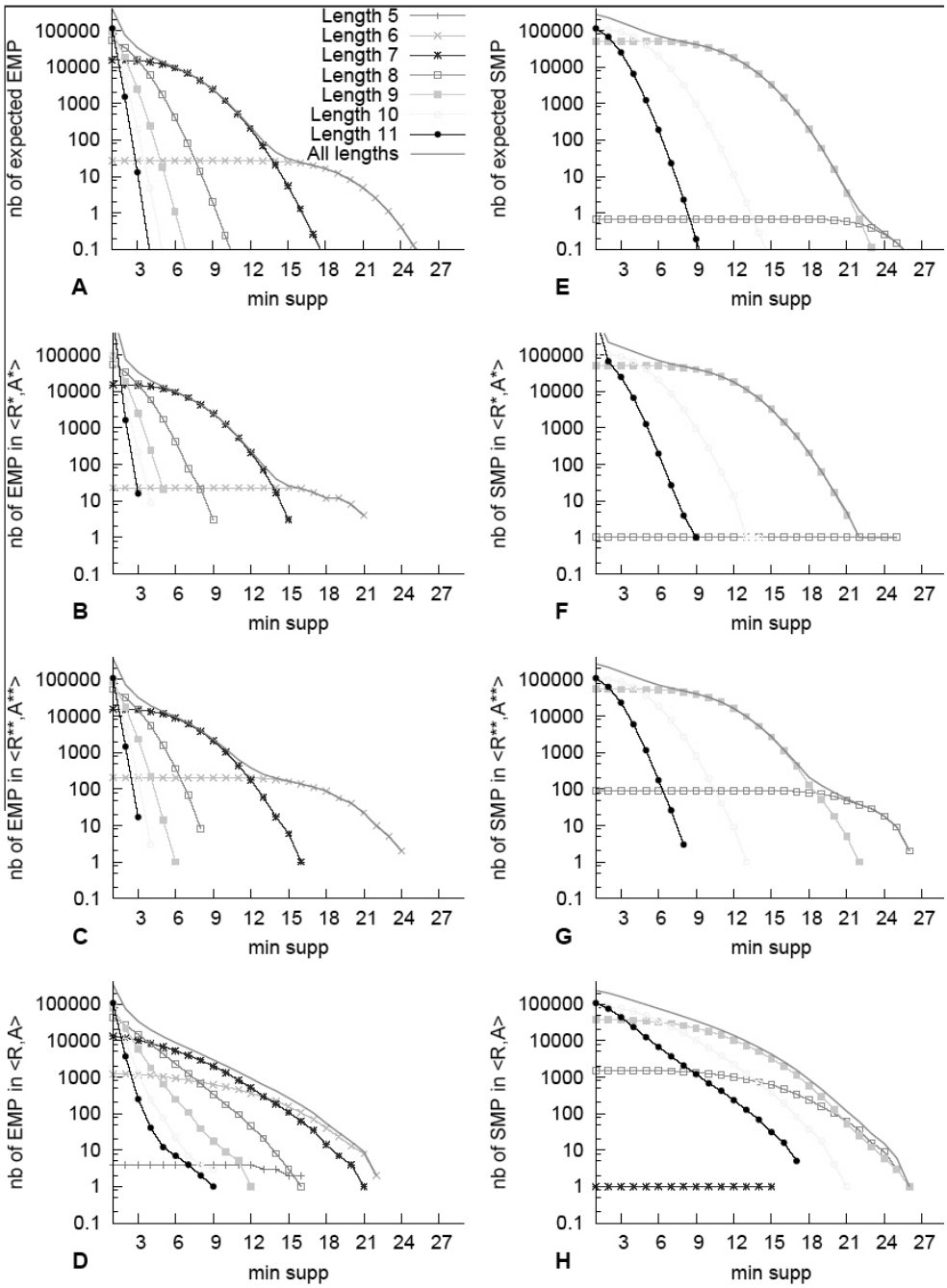

Fig. 3. Number of patterns satisfying both a minimum support constraint in the positive dataset and a maximum support constraint in the negative dataset. The plots represent the number of patterns, when the minimum support varies from 1 to 29, and the maximum support is set to 7. In the case of SMPs the allowed Hamming distance is set to 1. The numbers of patterns are given for pattern lengths from 5 to 11, and also accumulated for all these lengths. Graph A gives the number of expected EMPs computed using the TZI. Graph D (resp. B and C) gives the number of EMPs extracted from the pair of datasets (R, A) (resp. (R*, A*) and (R**, A**)). The corresponding plots in the case of SMPs are given in graphs E, H, F and G.

from 1 to 29, and $\alpha_{\max}$ is set to 7. The extractions were run using *Marguerite* that finds all patterns satisfying the α_{min}, α_{max} thresholds (and α_{dist} threshold for SMPs). Since *Marguerite* is complete, there is no need to run it several times with the same parameter values (contrarily to approaches based on incomplete heuristics). For each parameter setting, $\alpha_{\min}$, $\alpha_{\max}$ (and α_{dist} for SMPs), we compute the number of expected patterns due to the random background using the TZI formula.

The graphs depicted in Fig. 3 allow to compare the expected number of patterns with the number of patterns extracted in random and biological datasets (the extractions used for the illustration are representative, i.e., the behaviour remains the same also for other $\alpha_{\max}$ values). The number of patterns extracted in datasets (R*,A*) coincides with the expected number (graphs A vs. B and E vs. F in Fig. 3). This argues in favor of the correctness of the hypothesis made concerning the limited impact of the overlapping of the occurrences of the patterns on the number of extracted patterns. The number of patterns extracted in the datasets (R**,A**) (with the exceptions of the EMPs of length 6 and of the SMPs of length 8) is still well modeled by the computed number of expected patterns (graphs A vs. C and E vs. G in Fig. 3). This confirms that the simplification of considering an equiprobable nucleotide distribution and not taking into account the sequencing uncertainties do not modify the counts significantly. The number of the patterns extracted in the biological datasets R, A deviates more from the expected number and is greater than in the random datasets, but the estimations still model well the tendencies, especially in the range of parameters that are interesting to our problem, i.e., when $\alpha_{\min}$ is large and $\alpha_{\max}$ is small (graphs D and H in Fig. 3). This indicates that at least a part of the hidden structure of the biological dataset pair (R, A) (absent from the model of random background and absent from the random datasets) seems to be captured by the extracted patterns.

Setting parameters to focus on patterns satisfying the most stringent criteria

Even if we use some domain knowledge to construct a positive and a negative datasets, a poor constraint setting can lead to the extraction of many patterns likely to be due to the random background. Therefore we want to choose the parameter values to focus on the patterns satisfying the most stringent constraints (i.e., having a large support on the positive dataset and a small one on the negative dataset). Such patterns are interesting, since they are exceptionally conserved in the context of positive and negative datasets, and thus might be there to accomplish a biological function. To find these patterns, we use a parameter tuning method based on the following remark. When $\alpha_{\min}$ is very large and $\alpha_{\max}$ very small, we are likely to have no pattern satisfying the two constraints. Then if we decrease $\alpha_{\min}$ and/or increase $\alpha_{\max}$ (i.e., weaken the constraints) we go towards points in the parameter space for which we will start to obtain some patterns. Consistently, if from one of this point we go on decreasing $\alpha_{\min}$ and/or increasing $\alpha_{\max}$, we reach points in the parameter space for which we obtain more and more patterns satisfying the constraints. The parameter tuning method is based on the notion of *rising pattern* defined as follows. For a given value of $\alpha_{\min}$ we consider the minimal value of $\alpha_{\max}$ such that we have at least one pattern M satisfying *support*$^+(M) \geqslant \alpha_{\min}$ and *support*$^-(M) \leqslant \alpha_{\max}$. The patterns obtained for this $\alpha_{\min}$ and $\alpha_{\max}$ values are the rising patterns, i.e., there is no pattern for lower $\alpha_{\max}$ values, and for larger $\alpha_{\max}$ values we will have more patterns (or at least an equal number). Thus the rising patterns are located in the parameter space along a border that corresponds to the most stringent constraints that still lead to the extraction of at least one pattern. To find these rising patterns and corresponding parameter values we run an automated serialization of extractions (the algorithm is described in section "Procedure to find rising patterns", in Methods).

Graph A in Fig. 4 gives the numbers of rising EMPs of length 6 found in the datasets (R, A) (the exploration of the parameter is not made for $\alpha_{\min}$ values less than 15 since we are not likely to be

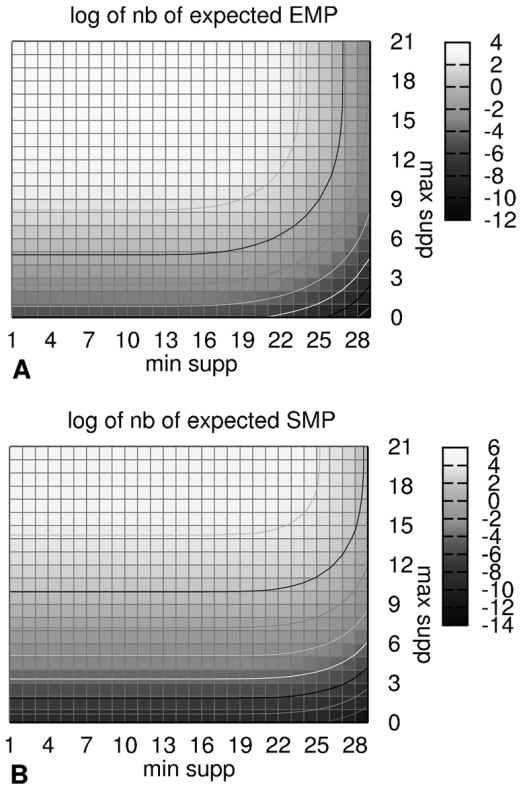

Fig. 4. Number of rising patterns and number of expected random patterns. For a point *MinSupp, MaxSupp* in the parameter space the color in the background corresponds to the $\log_{10}$ of the number of expected random patterns (i.e., the $\log_{10}$ of the TZI value for $\alpha_{\min} =$ *MinSupp* and $\alpha_{\max} =$ *MaxSupp*) in the dataset R, A. The values in the white circles give the number of rising patterns that were extracted for $\alpha_{\min} =$ *MinSupp* and $\alpha_{\max} =$ *MaxSupp*. The dashed line indicates the border of the TZ.

interested in patterns occurring in less than 15 sequences out of 29). Graph B in Fig. 4 gives the numbers of rising SMPs of length 8 with $\alpha_{\text{dist}} = 1$ in the same datasets. The colors of the background of the graphs correspond to the TZI values and indicates the number of expected random patterns. We observe that, consistently with the notion of TZI, the rising patterns are situated outside or in the very beginning of the estimated TZ.

It should be noticed that for the SMPs the increase of execution time[7] can, in some cases, prevent the possibility to run the extractions in a systematic way to explore the parameter space. In this situation we can still compute very efficiently the expected number of random patterns using the TZI and thus locate the TZ border. Then, the points along this border can be used as initial guess to find the rising patterns and their corresponding parameter settings. Moreover, the rising patterns are located along a rather regular contour, thus if the extraction time for SMPs is really too high, we can compute them only for a few $\alpha_{\min}$ values, and still have an idea of the whole curve.

The graphs in Fig. 4 give a global picture of the parameter space and are used to guide the setting of the parameters. For instance, if we are looking for an EMP with a high support in the positive dataset (e.g., $\alpha_{\min} = 27$), we have to accept a rather high support in the negative dataset ($\alpha_{\max} = 10$). Or, on

[7]Due to the soft occurrences handling and to the increase of the number of patterns satisfying the $\alpha_{\min}$ threshold [22].

the contrary, we can use the graph to choose a moderate support in the positive dataset (e.g., $\alpha_{\min} = 17$), and in this case we know that we can get some patterns having a low support ($\alpha_{\max} = 4$) on the negative dataset. Moreover, since this point ($\alpha_{\min} = 17$, $\alpha_{\max} = 4$) in the parameter space is not in the TZ, we can decide to increase a little the $\alpha_{\max}$ value to run an extraction to try to get a few more patterns. Of course, in this case we will enter the beginning of the TZ, and thus, among the patterns that will be obtained, several of them are likely to be due to the random background.

The length of the patterns is also a parameter in itself, and in order to ease its setting we use for the pattern extraction a tool that performs an exhaustive extraction within a range of length (from $L_{\min}$ to $L_{\max}$, see section "The pattern discovery tool *Marguerite*", in Methods). In practice, no pattern, or no interpretable pattern was found for a length out of the range $L_{\min} = 6$ and $L_{\max} = 10$. Finally, the last parameter α_{dist} (used for SMPs) should be kept as low as possible. When α_{dist} increases, a pattern matches occurrences that are more degenerated, and then is likely to be less specific to one of the two datasets and/or to be harder to interpret. In this study, the reasonable choice for α_{dist} is limited to $\alpha_{\text{dist}} = 1$ or 2, and patterns that we could interpret, in a useful way, were found only for $\alpha_{\text{dist}} = 1$.

Workflow of the motif-discovery process

Finding signature motifs characteristic of a given positive promoter set w.r.t. a negative set is only the first step of the process. The diagram given in Fig. 5 depicts the complete workflow ultimately designed to find putative TFBSs specific for a given promoter set. Using the tool *Marguerite* we extract patterns (EMPs or SMPs) specific to the positive dataset, i.e., all patterns satisfying a minimum support constraint in the positive dataset and a maximum support constraint in the negative one.

Then, the measure TZI (see section "Ranking patterns by a Twilight Zone Indicator", in Results and Discussion) is computed for every extracted pattern. For EMPs, we also compute as an additional measure the following ratio: the pattern support in the positive dataset divided by the pattern support in the negative dataset. The higher is the value of this ratio, the more specific to the positive dataset is the pattern. In the case of SMPs, the number of extracted patterns is much larger than the number of EMPs (see graphs H and D in Fig. 3), and in the result of an extraction many SMPs are similar to other SMPs obtained at the same time (due to the similarity based matching). Thus, we grouped the similar SMPs by performing a hierarchical clustering (see section "Hierarchical clustering of SMPs", in Methods) of the patterns. For each cluster we compute the average of the TZI of the patterns in the cluster. In each cluster, we also align the patterns with the multiple alignment tool MultAlin [26] to build a consensus SMP of each cluster (see section "Finding a consensus pattern of a cluster", in Methods). Finally, for both SMPs and EMPs, we use the tool Patchlike (see section "TRANSFAC®, Patchlike", in Methods) to check w.r.t. the TRANSFAC® database, which patterns are known TFBSs.[8] We are particularly interested in patterns that are the TFBSs of the TFs, involved in v-ErbA transforming activity. Untill this point, the extraction process does not rely upon any collection of known motifs. It is therefore obvious that some of the extracted motifs will not correspond to any known TFBS, either in TRANSFAC® or in any other database. Those motifs can be qualified as putative unknown binding sites, and are candidates for experimental analysis.

[8]Note, that Patchlike also considers the patterns originating from the reverse complemented sequence and thus search for TFBSs in both directions.

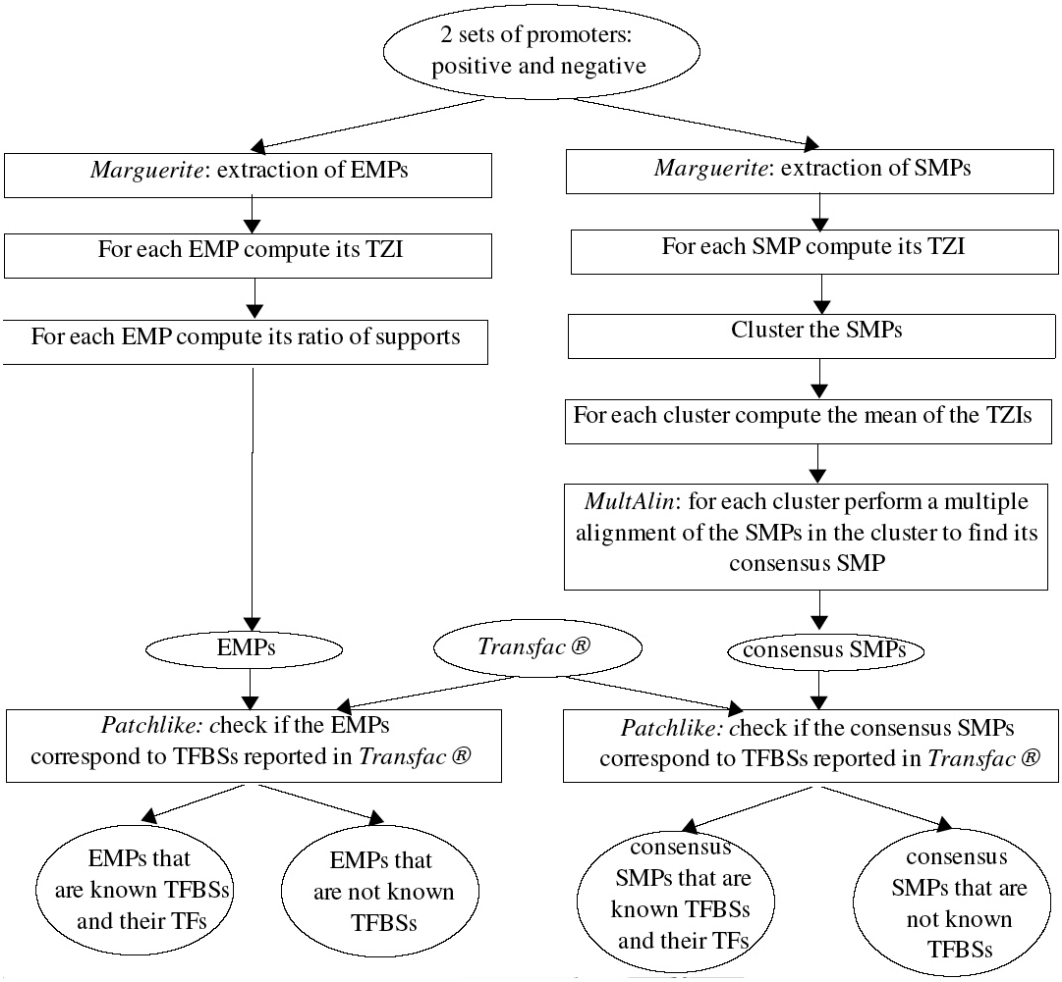

Fig. 5. Diagram depicting the steps of the whole motif-discovery process.

Patterns that are putative binding sites of TFs involved in v-ErbA transforming activity

The particularity of our method to use both a positive and a negative dataset, representing opposite biological situations, is a major reason why it would not be relevant to test our algorithm on a classical benchmark, such as [2] or [35], which uses only one set of sequences. Furthermore, as pointed out by [3], we lack an absolute standard against which to measure the correctness of any motif-finding tool. Therefore, to assess our approach, we will concentrate on the biological interpretation of some of the extracted patterns.

Exact Matching Patterns (*EMPs*)

We first extracted rising EMPs (see section “Procedure to find rising patterns”, in Methods) of length from 5 to 10 within an interval of thresholds for minimum support from 15 to 29 (corresponding to a relative support ranging from 51.7% to 100% of the sequences) in dataset R and an interval of thresholds for maximum support from 0 to 11 (corresponding to a relative support ranging from 0% to 52.4%

Table 1
EMPs that are putative TFBSs bound by TFs involved in the v-ErbA transforming activity

EMP	Supp in R	Ratio Supp in R Supp in A	TZI	TFBS	TF
GGAAACA	18	6	0.02	GGAAAC (+)	Net
				TGTTTC (−)	AR, GR-α
CGCTGCG	17	5.67	0.09	GCTGC (+)	CTCF
TGCAAAC	17	5.67	0.09	GTTTG (−)	ZEB (1124 AA)
CAGTTA	19	4.75	0.1	CAGTTA (+)	c-Myb, c-Myb-isoform1
				TAACTG (−)	c-Myb
				TAACT (−)	RXR-α
AGATAT	17	4.25	0.2	AGATAT (+)	GATA-3/3 isoform 1/ 4/5A/5B/6A/6B
				ATATCT (−)	GATA-1/1 isoform 1/3/3 isoform-1
				AGATA (+)	GATA-1/3/4/6
				TATCT (−)	GATA-1

Rising EMPs that are known binding sites of TFs implicated into self-renewal process. The sign indicates the direction of the match: (+) forward and (−) reverse complement.

of the sequences) in dataset A. These intervals are rather large, and the worst case ($\alpha_{\min} = 15$ and $\alpha_{\max} = 11$) is likely to lead to uninteresting patterns (not biologically founded). However on datasets containing an underlying structure, the procedure to find rising patterns does not reach such extreme cases, since rising patterns are obtained before, for more interesting $\alpha_{\min}$ and $\alpha_{\max}$ values. This is actually the case for datasets R and A, where, for the rising patterns, $\alpha_{\min}$ was always greater than $2.4 \times \alpha_{\max}$. In these extractions we obtained 33 rising EMPs, for each of them we computed its TZI measure and its support ratio (support in dataset R divided by its support in dataset A), and looked at putative binding TFs with Patchlike. After visual inspection of this information, we selected five rising EMPs as candidates for further biological exploration, because they had a high support in the positive datasets R, a high support ratio, an interesting TZI value (i.e., low value) and meaningful putative binding TFs (Table 1). These patterns were extracted for the following ($\alpha_{\min}$, $\alpha_{\max}$) pairs: (17,3); (18,3); (19,4). Some other EMPs have quite high support ratio, high support in dataset R, and low TZI value, but are not known TFBS in TRANSFAC® (not shown). This is one of the benefit of such an unsupervised approach to allow such unknown motif discovery. Since our knowledge of TFs-TFBSs relationship is still very incomplete, the best rated of those motifs could be used for functional assay using reporter gene transfection, and may lead to the discovery of new TFs, relevant for v-ErbA-induced transformation. Among the EMPs displayed in Table 1, one of the most interesting is CAGTTA, that is a known binding site for the transcription factor c-Myb. This pattern has a quite high support ratio (4.75), a high support in dataset R (19 out of 29 promoters), and an interesting TZI value of 0.1. Since this pattern appeared in a previous exploration of the same set of promoters, we had the opportunity to assess its putative relevance for the v-ErbA-induced transformation process. We indeed could demonstrate the existence of a functional interaction between v-ErbA and c-Myb [27], thereby demonstrating the biological relevance of this approach. Some other patterns are also of interest, and can be expected knowing the molecular action of v-ErbA. Indeed this oncogene is known to interfere with nuclear receptors, including RXR and GR [36]. Since v-ErbA is known to block the erythroid differentiation process [31], it is also expected to see some overlap between v-ErbA target genes and genes containing binding sites for the erythroid specific TF GATA-1 [37]. A similar reasoning can apply for the chromatin modifier CTCF [38,39] and the trancriptional repressor ZEB [40] that are both involved in the erythroid differentiation process. It would be of interest to verify the functional interaction between those TFs and v-ErbA by a reporter assay similar to the one used for c-Myb.

Table 2
SMPs that are putative TFBSs bound by TFs involved in the v-ErbA transforming activity

Consensus SMP	Nb of SMPs in cluster	Mean of TZI	TFBS	**TF**
ta.cTaTg	9	16.7	TAACT (+)	RXR-α
			TATCT (+)	GATA-1
			AGATA (−)	GATA-1/4/6
TaGttag	11	24	TAACT (−)	RXR-α
aTagTg.t	13	34.9	AGTGGT (+)	GR, GR-α
t.TCAACt	6	35.8	TCAACT (+)	CAR2:RXR-α
				CAR/PXR:RXR
			AGTTGA (−)	RAR-α1, RXR-α
			CAACT (+)	c-Myb-isoform1
aCgTt.a	17	36.9	TGAACG (−)	HOXA9
			TAACG (−)	c-Myb-isoform1
			GTTCA (+)	RAR-α1, T3R-α

Consensus SMPs that are known binding sites of TFs implicated into self-renewal process. The sign indicates the direction of the match: (+) forward and (−) reverse complement.

A similar search can of course be performed using the A set of promoter sequences as the positive set. In this case, one should note that, although patterns are, by construction, specific of a given promoter set, TFs binding those patterns can appear on both datasets (of course, in this case, the *same* TF will bind *different* patterns in the two sets). The user can perform both searches using sequentially both promoter sets as the positive set, and then either focus only on set-specific TFs or on TFs shared by the two sets.

Soft Matching Patterns (*SMPs*)

In the case of SMPs, we estimated analytically (as described in section "Setting parameters to focus on patterns satisfying the most stringent criteria", in Results and Discussion) that with minimum support threshold equal to 17 (it correspond to 58.6% in relative support) on a dataset R* and maximum support threshold equal to 10 (it correspond to 47.6% in relative support) on a dataset A* the SMPs of length between 7 and 11 are outside or in the beginning of the TZ. Thus we extracted the SMPs satisfying the length and support constraints with these parameters in datasets R and A. Table 2 gives five consensus SMPs (see section "Finding a consensus pattern of a cluster", in Methods) that have the best mean of TZI and are known TFBSs. These consensus SMPs are issued from hierarchical clustering of SMPs of length 8 using complete linkage method and a cut-off level of 50% (see section "Hierarchical clustering of SMPs", in Methods).

One can note an extensive similarity between the TFs binding to the SMPs and those binding to the EMPs. This concerns both the sites bound by nuclear receptors (RXR, RAR and GR) and those bound by GATA-1. Among the unexpected factors, one finds the HoxA9 homeobox factor. It is very interesting to note the mounting evidence suggesting that HoxA9 plays an important role in both normal hematopoiesis [41] as well as in leukemogenesis [42]. It would be of interest to examine the possibility that v-ErbA interferes with HoxA9 during the transformation process of erythroid progenitors.

Patterns that are putative binding sites of TFs involved in the self-renewal of eryhroid progenitors

In order to assess the generality of our approach, we also applied our method to a second data set, made from the promoters of genes showing differential expression, as assessed by SAGE, between self-renewing and differentiating erythroid progenitors [33]. In this case we only applied the EMP-based strategy, in order to isolate TFBSs specific for the promoters of genes significantly more expressed in

Table 3
EMPs that are putative TFBSs bound by TFs involved in the self-renewal of normal erythroid progenitors

EMP	Supp in AR	Ratio Supp in AR Supp in Diff	TZI	TF
CAGTTCT	16	5.3	0.41	c-Myb
CTGCTGG	21	3.5	0.000042	c-Maf (long form)
ATGCAGC	17	5.7	0.078	CTCF
CACCCAC	15	7.5	1.05	EKLF

Rising EMPs that are known binding sites of TFs implicated into self-renewal process.

the self-renewal condition than after inducing differentiation. The data set was made of 28 promoter sequences in the positive set, denoted *AR*, and of 16 promoter sequences in the negative set, denoted *Diff* (promoters of genes the expression of which is up-regulated during the first 24 hours of differentiation [33]). An exhaustive search for rising SMPs returned 55 different motifs. As previously described, motifs were selected for further analysis based upon: 1. their TZ value, 2: their min to max ratio, and 3. their TRANSFAC identification. This left us with 4 motifs (Table 3), the biological significance of which was further assessed.

We first notice the presence of a c-Myb-binding site. Since the transcription of c-myc is typically down-regulated during the erythroid cell maturation process [43] and since its constitutive expression blocks erythroid differentiation [44], the presence of a c-Myb-binding motif was expected among the self-renewal-specific set of promoters. One should note that a c-Myb binding motif has been revealed by our approach both in the case of normal and deregulated self-renewal induced by the v-*erbA* oncogene (see upper), although the exact sequence of the two motifs was subtly different. This nevertheless reinforce the biological involvement of c-Myb in the self-renewal process in our erythroid progenitors.

The case of the second motif illustrates one difficulty of the *in silico* approaches. Indeed there is a large number of members of the Maf family of basic region/leucine zipper (bZIP) transcription factors that play an essential role in growth and development by regulating tissue-specific gene expression. These proteins activate or repress transcription depending on their particular protein partner and the context of the target promoter [45] (and references therein). For example c-Maf has been shown to bind to c-myb resulting in the inhibition of Myb-dependent gene transcription [45]. Therefore those *in silico* approaches would have to be complemented by experimental approaches, like ChIP (chromatin immunoprecipitation [46]) for identifying the nature of the Maf members bound to the putative c-Maf-binding sites *in vivo*. Furthermore reporter assay should help understand the role of the context of the promoter.

Similarly, the last two motifs illustrate a fundamental limitation of any *in silico* method for identifying TFBSs, that is that the method does not provide the functional result of the putative binding of a given transcription factor. Indeed the CTCF factor has been shown to be a "dual functionality" protein that has divergent effects on different gene regulation systems [47] (and references therein). It is mainly involved in silencing the regions outside the active globin genes in erythroid cells, so its function in the self-renewal maintenance is still elusive. Similarly, the EKLF transcription factor is a known factor involved positively in the erythroid differentiation process [48]. One may imagine that the finding of such a site in the set of self-renewal specific genes points toward the possibility that EKLF promotes differentiation by, on one hand, activating differentiation-related genes, but also, on the other hand, by repressing a subset of self-renewal specific genes. But of course such an hypothesis still awaits an experimental validation that is beyond the goal of the present work.

DISCUSSION

An important research effort has been dedicated to the extraction of motifs to find putative TFBSs, but even the best today techniques report limited results in practice [2,3,35]. These techniques, based on combinations of efficient extraction strategies together with dedicated statistical measures, often still suffer from high false positive rates and/or from the difficulty to select appropriate parameters. We have used a new method that incorporates as a central aspect the use of background knowledge, in the extraction algorithm itself, to reduce the number of false positives, and that makes use of an effective parameter tuning strategy. To help the user to pick up the most promising patterns among the ones extracted, the whole process also includes the two following additional steps: (1) the computation of an interestingness measure based on the notion of Twilight Zone [4], and (2) the automated retrieval of the TFBSs known in TRANSFAC® (together with the corresponding TFs), for the TFBSs that are similar to the extracted patterns. The current knowledge about TFs and TFBSs does not permit to determine the set of true positives (the set of TFBSs in a gene promoter sequence involved in the regulation of this gene), and thus we do not provide an estimate of the number of true positive TBFSs that may be missed by the method.

The major strength of the whole approach is that it does not only rely on a model of the random background to assess the interestingness of the patterns, but uses one dataset in which the patterns are searched, and incorporates as background knowledge a second dataset in which the patterns of interest are not likely to appear. Although it seems simple, this strategy is not supported by the state of the art techniques to find putative TFBSs. In fact, putting at work this strategy is not straightforward in practice, in particular we have to face a large parameter space (one frequency threshold for each of the two datasets, the size of the patterns, the degree of approximated matching allowed) and selecting appropriated parameter values is a difficult task that may even turn to be prohibitive. Thus, the second main point of the proposed method is to provide an explicit parameter tuning technique, that turns out to be effective in practice.

This power is exemplified by using two pairs of promoter sets, each pair consisting of positive examples, and negative ones, derived from differential gene expression experiments. The first analyzed pair resulted from a SAGE-based analysis of v-ErbA-induced gene expression regulation. The v-*erbA* oncogene, carried by the Avian Erythroblastosis Virus, derives from the c-*erbA* proto-oncogene that encodes the nuclear receptor for triiodothyronine (T3R). v-ErbA transforms erythroid progenitors *in vitro* by blocking their differentiation [49], supposedly by interference with T3R and RAR (retinoic acid receptor) [36]. However, when v-ErbA target genes involved in its transforming activity were identified using a SAGE-based approach, it turned out that a vast majority of them were not regulated neither by T3R nor RAR, suggesting the involvement of new unanticipated mechanisms of v-ErbA action [27]. One open question therefore concerns the mechanisms through which v-ErbA represses the expression of those genes. The present study points toward promising leads in this regard:

- We re-identify a TFBS bound by the c-Myb proto-oncogene product. The importance of this TF in the v-ErbA-induced phenotype has been validated by experimental evidence showing that v-ErbA can indeed functionally interact directly or indirectly with the transcriptional activity of endogenous c-Myb in chicken erythroid progenitors [27].
- A second lead has yet to be functionally investigated. It concerns the putative role of the *HoxA9* gene in the v-ErbA-induced phenotype. Testing this hypothesis would require overexpression or sh-RNA mediated repression of expression, using a previously described strategy [50].

The second pair of sets of promoters was derived from a SAGE experiment aiming at finding genes the expression of which was modulated during the decision-making process between selfrenewal and differentiation [33]. One should note that functional analysis of one of this gene revealed that it was indeed functionally involved in the self-renewal process [50]. This analysis reinforced the putative role of c-myb in the control of the self-renewal of erythroid progenitors, and also pointed into directions that will require additional experimental work.

CONCLUSION

We demonstrate the power of using relevant biological information, in the form of a set of differentially expressed genes that is a classical outcome in most of transcriptomics studies. This allows to severely reduce the search space and to design an adapted statistical indicator. Taken together this allows the biologist to concentrate on a small number of putatively interesting TFs.

AUTHORS CONTRIBUTIONS

IM designed the *Marguerite* software. SS designed the rising pattern extractor and the Patchlike program. LM, IM and CR designed the TZI and the clustering of SMPs. LM, SS and IM performed the described experiments. OG and JFB drafted the initial version of the project. OG provided the promoter sequences. CR, JFB and OG provided guidance throughout the project. IM, CR and OG wrote the manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGEMENTS

We are grateful to Yann Letrillard for performing a preliminary tentative sketch of this study. This work has mainly been done within the Bingo project framework. The authors thank all members of the Bingo project for fruitful discussions. This work has been partially funded by the ANR (French Research National Agency) project BINGO2 ANR-07-MDCO-014 which is a follow-up of the first BINGO project (2004–2007) and by EU contract IST-FET IQ FP6-516169 (FET arm of the IST programme). SS was funded by a “Pôle Régional de bioinformatique CPER 2000–2006” research grant to OG.

REFERENCES

[1] Vanet, A., Marsan, L. and Sagot, M.-F. (1999). Promoter sequences and algorithmical methods for identifying them. Res. Microbiol. **150**, 779-799.
[2] Tompa, M., Li, N., Bailey, T. L., Church, G. M., De Moor, B., Eskin, E., Favorov, A. V., Frith, M. C., Fu, Y., Kent, W. J., Makeev, V. J., Mironov, A. A., Noble, W. S., Pavesi, G., Pesole, G., Régnier, M., Simonis, N., Sinha, S., Thijs, G., van Helden, J., Vandenbogaert, M., Weng, Z., Workman, C., Ye, C. and Zhu, Z. (2005). Assessing computational tools for the discovery of transciption factor binding sites. Nat. Biotechnol. **23**, 137-144.
[3] Das, M. K. and Dai, H.-K. (2007). A survey of DNA motif finding algorithms. BMC Bioinformatics **8 Suppl 7**, S21.
[4] Keich, U. and Pevzner, P. A. (2002). Subtle motifs: defining the limits of motif finding algorithms. Bioinformatics **18**, 1382-1390.
[5] Keich, U. and Pevzner, P. A. (2002). Finding motifs in the twilight zone. Bioinformatics **18**, 1374-1381.
[6] Bulyk, M. L. (2003). Computational prediction of transcription-factor binding site locations. Genome Biol. **5**, 201.

[7] Matys, V., Fricke, E., Geffers, R., Gößling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A. E., Kel-Margoulis, O. V., Kloos, D.-U., Land, S., Lewicki-Potapov, B., Michael, H., Münch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S. and Wingender, E. (2003). TRANSFAC®: transcriptional regulation, from patterns to profiles. Nucleic Acids Res. **31**, 374-378.

[8] Fogel, G. B., Weekes, D. G., Varga, G., Dow, E. R., Craven, A. M., Harlow, H. B., Su, E. W., Onyia, J. E. and Su, C. (2005). A statistical analysis of the TRANSFAC database. Biosystems **81**, 137-154.

[9] Queen, C., Wegman, M. N. and Korn, L. J. (1982). Improvements to a program for DNA analysis: a procedure to find homologies among many sequences. Nucleic Acids Res. **10**, 449-456.

[10] Waterman, M. S., Arratia, R. and Galas, D. J. (1984). Pattern recognition in several sequences: Consensus and alignment. Bull. Math. Biol. **46**, 515-527.

[11] Staden, R. (1989). Methods for discovering novel motifs in nucleic acid sequences. Comput. Appl. Biosci. **5**, 293-298.

[12] Sagot, M.-F., Escalier, V., Viari, A. and Soldano, H. (1995). Searching for repeated words in a text allowing for mismatches and gaps. *In*: Proc. of the 2nd South American Workshop on String Processing (WSP'95), Vinas del Mar, Chile, pp. 87-100.

[13] Sagot, M.-F. and Viari, A. (1996). A double combinatorial approach to discovering patterns in biological sequences. *In*: Proc. of the 7th Annual Symposium on Combinatorial Pattern Matching (CPM'96), Laguna Beach, California, Springer-Verlag, pp. 186-208.

[14] Brazma, A., Jonassen, I., Vilo, J. and Ukkonen, E. (1998). Predicting gene regulatory elements in silico on a genomic scale. Genome Res. **8**, 1202-1215.

[15] Rigoutsos, I. and Floratos, A. (1998). Combinatorial pattern discovery in biological sequences: the TEIRESIAS algorithm. Bioinformatics **14**, 55-67.

[16] Brazma, A., Jonassen, I., Eidhammer, I. and Gilbert, D. (1998). Approaches to the automatic discovery of patterns in biosequences. J. Comput. Biol. **5**, 279-305.

[17] Wolfertstetter, F., Frech, K., Herrmann, G. and Werner, T. (1996). Identification of functional elements in unaligned nucleic acid sequences by a novel tuple search algorithm. Comput. Appl. Biosci. **12**, 71-80.

[18] Eden, E., Lipson, D., Yogev, S. and Yakhini, Z. (2007). Discovering motifs in ranked lists of DNA sequences. PLoS Comput. Biol. **3**, e39.

[19] Dan Lee, S. and De Raedt, L. (2004). An efficient algorithm for mining string databases under constraints. Proc. of the 3rd Int. Workshop on Knowledge Discovery in Inductive Databases (KDID'04), Pisa, Italy, Springer-Verlag, pp. 108-129.

[20] Raedt, L. D., Jaeger, M., Lee, S. D., and Mannila, H. (2002). A theory of inductive query answering. *In*: Proc. of the 2nd IEEE International Conference on Data Mining (ICDM'02), Maebashi City, Japan, IEEE Computer Society, pp. 123-130.

[21] Mitasiunaite, I. and Boulicaut, J.-F. (2006). Looking for monotonicity properties of a similarity constraint on sequences. *In*: Proc. of the 21st Annual ACM Symposium on Applied Computing (SAC'06), Special Track on Data Mining, Dijon, France, ACM Press, pp. 546-552.

[22] Mitasiunaite, I. and Boulicaut, J.-F. (2007). Introducing softness into inductive queries on string databases. *In*: Databases and Information Systems IV, vol. 155 of Frontiers in Artificial Intelligence and Applications, Vasilecas, O., Eder, J., and Caplinskas, A. (eds.), IOS Press, Amsterdam, The Netherlands pp. 117-132.

[23] Steuer, R. (1985). Multiple Criteria Optimization: Theory, Computation and Application. John Wiley & Sons.

[24] The R project for statistical computing, http://www.r-project.org/.

[25] Needleman, S. and Wunsch, C. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. **48**, 443-453.

[26] Corpet, F. (1988). Multiple sequence alignment with hierarchical clustering. Nucleic Acids Res. **16**, 10881-10890.

[27] Bresson, C., Keime, C., Faure, C., Letrillard, Y., Barbado, M., Sanfilippo, S., Benhra, N., Gandrillon, O. and Gonin-Giraud, S. (2007). Large-scale analysis by SAGE revealed new mechanisms of *v-erbA* oncogene action. BMC Genomics **8**, 390.

[28] Wray, G. A., Hahn, M. W., Abouheif, E., Balhoff, J. P., Pizer, M., Rockman, M. V. and Romano, L. A. (2003). The evolution of transcriptional regulation in eukaryotes. Mol. Biol. Evol. **20**, 1377-1419.

[29] Velculescu, V. E., Zhang, L., Vogelstein, B. and Kinzler, K. W. (1995). Serial analysis of gene expression. Science **270**, 484-487.

[30] Gandrillon, O., Schmidt, U., Beug, H. and Samarut, J. (1999). TGF-β cooperates with TGF-β to induce the self-renewal of normal erythrocytic progenitors: evidence for an autocrine mechanism. EMBO J. **18**, 2764-2781.

[31] Gandrillon, O., Jurdic, P., Pain, B., Desbois, C., Madjar, J. J., Moscovici, M. G., Moscovici, C. and Samarut, J. (1989). Expression of the v-*erbA* product, an altered nuclear hormone receptor, is sufficient to transform erythrocytic cells *in vitro*. Cell **58**, 115-121.

[32] Sharif, M. and Privalsky, M. L. (1991). v-*erbA* oncogene function in neoplasia correlates with its ability to repress retinoic acid receptor action. Cell **66**, 885-893.

[33] Damiola, F., Keime, C., Gonin-Giraud, S., Dazy, S. and Gandrillon, O. (2004). Global transcription analysis of immature avian erythrocytic progenitors: from self-renewal to differentiation. Oncogene **23**, 7628-7643.
[34] Piva, F. and Principato, G. (2006). RanDNA: a random DNA sequence generator. In Silico Biol. **6**, 253-258.
[35] Hu, J., Li, B. and Kihara, D. (2005). Limitations and potentials of current motif discovery algorithms. Nucleic Acids Res. **33**, 4899-4913.
[36] Gandrillon, O. and Samarut, J. (1998). Role of the different RAR isoforms in controlling the erythrocytic differentiation sequence. Interference with the v-*erb*A and $p135^{gag-myb-ets}$ nuclear oncogenes. Oncogene **16**, 563-574.
[37] Wickrema, A. and Crispino, J. D. (2007). Erythroid and megakaryocytic transformation. Oncogene **26**, 6803-6815.
[38] Torrano, V., Chernukhin, I., Docquier, F., D'Arcy, V., León, J., Klenova, E. and Delgado, M. D. (2005). CTCF regulates growth and erythroid differentiation of human myeloid leukemia cells. J. Biol. Chem. **280**, 28152-28161.
[39] Splinter, E., Heath, H., Kooren, J., Palstra, R.-J., Klous, P., Grosveld, F., Galjart, N. and de Laat, W. (2006). CTCF mediates long-range chromatin looping and local histone modification in the β-globin locus. Genes Dev. **20**, 2349-2354.
[40] Turner, J. and Crossley, M. (1999). Basic Krüppel-like factor functions within a network of interacting haematopoietic transcription factors. Int. J. Biochem. Cell Biol. **31**, 1169-1174.
[41] Chung, K. Y., Morrone, G., Schuringa, J. J., Plasilova, M., Shieh, J.-H., Zhang, Y., Zhou, P. and Moore, M. A. (2006). Enforced expression of *NUP98-HOXA9* in human $CD34^+$ cells enhances stem cell proliferation. Cancer Res. **66**, 11781-11791.
[42] Wang, G. G., Pasillas, M. P. and Kamps, M. P. (2006). Persistent transactivation by Meis1 replaces Hox function in myeloid leukemogenesis models: Evidence for co-occupancy of Meis1-Pbx and Hox-Pbx complexes on promoters of leukemia-associated genes. Mol. Cell. Biol. **26**, 3902-3916.
[43] Kuehl, W. M., Bender, T. P., Stafford, J., McClinton, D., Segal, S. and Dmitrovsky, E. (1988). Expression and function of the c-myb oncogene during hematopoietic differentiation. Curr. Top. Microbiol. Immunol. **141**, 318-323.
[44] Clarke, M. F., Kukowska-Latallo, J. F., Westin, E., Smith, M. and Prochownik, E. V. (1988). Constitutive expression of a c-*myb* cDNA blocks Friend murine erythroleukemia cell differentiation. Mol. Cell. Biol. **8**, 884-892.
[45] Hedge, S. P., Kumar, A., Kurschner, C. and Shapiro, L. H. (1998). c-Maf interacts with c-Myb to regulate transcription of an early myeloid gene during differentiation. Mol. Cell. Biol. **18**, 2729-2737.
[46] Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007). Genome-wide mapping of *in vivo* protein-DNA interactions. Science **316**, 1497-1502.
[47] Chan, P. K., Wai, A., Philipsen, S. and Tan-Un, K.-C. (2008). 5'HS5 of the human β-globin locus control region is dispensable for the formation of the β-globin active chromatin hub. PLoS ONE **3**, e2134.
[48] Bieker, J. J. and Southwood, C. M. (1995). The erythroid Krüppel-like factor transactivation domain is a critical component for cell-specific inducibility of a β-globin promoter. Mol. Cell. Biol. **15**, 852-860.
[49] Gandrillon, O., Rascle, A. and Samarut, J. (1995). The v-*erbA* oncogene: a superb tool for dissecting the involvement of nuclear hormone receptors in differentiation and neoplasia. Int. J. Oncol. **6**, 215-231.
[50] Bresson-Mazet, C., Gandrillon, O. and Gonin-Giraud, S. (2008). Stem cell antigen 2: a new gene involved in the self-renewal of erythroid progenitors. Cell Prolif. **41**, 726-738.