BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 13 2005, pages 2933–2942
doi:10.1093/bioinformatics/bti473

Sequence analysis

MatInspector and beyond: promoter analysis based on transcription factor binding sites

K. Cartharius*, K. Frech, K. Grote, B. Klocke, M. Haltmeier, A. Klingenhoff, M. Frisch, M. Bayerlein and T. Werner

Genomatix Software GmbH, Landsberger Strasse. 6, 80339 München, Germany

Received on March 14, 2005; revised and accepted on April 25, 2005
Advance Access publication April 28, 2005

ABSTRACT

Motivation: Promoter analysis is an essential step on the way to identify regulatory networks. A prerequisite for successful promoter analysis is the prediction of potential transcription factor binding sites (TFBS) with reasonable accuracy. The next steps in promoter analysis can be tackled only with reliable predictions, e.g. finding phylogenetically conserved patterns or identifying higher order combinations of sites in promoters of co-regulated genes.

Results: We present a new version of the program MatInspector that identifies TFBS in nucleotide sequences using a large library of weight matrices. By introducing a matrix family concept, optimized thresholds, and comparative analysis, the enhanced program produces concise results avoiding redundant and false-positive matches. We describe a number of programs based on MatInspector allowing in-depth promoter analysis (DiAlignTF, FrameWorker) and targeted design of regulatory sequences (SequenceShaper).

Availability: MatInspector and the other programs described here can be used online at http://www.genomatix.de/matinspector.html. Access is free after registration within certain limitations (e.g. the number of analysis per month is currently limited to 20 analyses of arbitrary sequences).

Contact: cartharius@genomatix.de

Supplementary information: http://www.genomatix.de/matinspector.html

INTRODUCTION

Control of transcription initiation is a pivotal mechanism for determining whether or not a gene is expressed and how much mRNA—and consequently protein—is produced. A promoter is a sequence that initiates and regulates the transcription of a gene. Protein binding sites in a promoter represent the most crucial elements and the corresponding proteins are called transcription factors (TFs). There is a large variety of TFs in the cell. Currently, more than 1400 human TFs are known and a total of ∼1850 (Venter *et al*., 2001) to 3000 (Lander *et al*., 2001) was estimated for the human genome. To be able to predict potentially functional transcription factor binding sites (TFBS) is an important first step in promoter analysis.

One way to describe TFBS is by nucleotide or position weight matrices (NWM or PWM) (for review see Stormo, 2000). A weight matrix pattern definition is superior to a simple IUPAC consensus sequence as it represents the complete nucleotide distribution for each single position. It also allows the quantification of the similarity between the weight matrix and a potential TFBS detected in the sequence. The concept of PWMs was developed in the 1980s, but the widespread use of the concept in form of programs was delayed almost a decade since only a few special matrices had been defined (Bucher, 1990). MatInspector was one of the first programs to close this gap in 1995, offering an extensive precompiled library of 214 weight matrices (Quandt *et al*., 1995).

Other programs based on PWMs include SignalScan (Prestridge, 1996) and Matrix Search (Chen *et al*., 1995), which were based on TRANSFAC 2.5, TFD 7.4 and IMD, and are not updated anymore. TESS (http://www.cbil.upenn.edu/tess/) uses a library based on TRANSFAC 4.0 and applies a log-likelihood score or the MatInspector scoring scheme for matrix searches. The program Match (Kel *et al*., 2003) uses a similar scoring scheme as MatInspector, but lacks matrix families (see below), resulting in a large number of redundant matches. The freely available version of Match allows searching for matrices from TRANSFAC 6.0 public, containing 336 matrices. The commercial version of TRANSFAC (release 8.4) currently has 741 entries, but academic users cannot use it freely. ConSite (Sandelin *et al*., 2004b) is based on the JASPAR database (Sandelin *et al*., 2004a), which contains 111 entries. Today the MatInspector library contains 634 matrices representing the largest library available for public searches.

It is important to note that TFBS only carry the potential to bind their corresponding protein. However, they can occur everywhere in the genome and are by no means restricted to regulatory regions. Sites outside regulatory regions are known to bind their TFs (Kodadek, 1998), and it is the context that differentiates a functional binding site affecting gene regulation from a mere physical binding site (Elkon *et al*., 2003). TFBS prediction programs like MatInspector can infer the binding potential, although not the functionality of a site. Functionality can ultimately be proven only by a wet-lab experiment with defined settings, particularly since potential binding sites in a promoter can be functional in certain cells, tissues or developmental stages and non-functional under different conditions. Involving advanced bioinformatics strategies like detailed promoter analysis of, for instance, orthologous or co-regulated genes can significantly reduce the number of test candidates.

ALGORITHM

In our original paper (Quandt *et al*., 1995) a set of tools for the generation of matrices (MatInd) and the detection of potential

*To whom correspondence should be addressed.

Table 1. Sequences used for creation of the RUNX2 (AML3) matrix

Name	Reference	Sequence (core sequence bold)	Matrix similarity
NMP2	Alvarez *et al*. (1997)	TTTA **GTGG** TTTTTC	0.898
OSE2	Willis *et al*. (2002)	TGCT **GTGG** TTGGT.	0.961
mRANKL2	O'Brien *et al*. (2002)	GGCT **GTGG** GTTGGG	0.879
hMMP-13	Mengshol *et al*. (2001)	GAGT **GTGG** TTTGTG	0.994
rbtMMP13	Mengshol *et al*. (2001)	AAGT **GTGG** TTTGTG	1.000
mMMP13	Mengshol *et al*. (2001)	AAGT **GTGG** TTTGTG	1.000
hRANKL	O'Brien *et al*. (2002)	TCCA **GTGG** TTCCAG	0.869
Collagenase-3	D'Alonzo *et al*. (2002)	ACGT **GTGG** TTTGTG	1.000
OPGhuman	Thirunavukkarasu *et al*. (2001)	CTCT **GAGG** TTTCCC	0.832
Galectin10	Dyer and Rosenberg (2001)	AGGT **GTGG** TTGTGA	0.913
mRANKL1	O'Brien *et al*. (2002)	TCCA **GTGG** TTGGTT	0.932
Ameloblastin	Dhamija and Krebsbach (2001)	catt ttgg tgagct	Rejected

TFBS (MatInspector) was introduced. MatInd constructs a matrix description consisting of a PWM, a conservation profile (so-called conservation index vector, C_i-vector) and a core region for a set of training sequences. The C_i-value at each position i of the matrix is calculated by

$$C_i(i) = \left(\frac{100}{\ln 5}\right) \times \left[\sum_{b\in \mathrm{A,C,G,T,gap}} P(i,b) \times \ln P(i,b) + \ln 5\right]$$

where $P(i,b)$ is the relative frequency of nucleotide b at position i. MatInspector uses this information to scan nucleotide sequences for matches to this pattern by calculating a matrix similarity score which reaches 1 only if the test sequence corresponds to the most conserved nucleotide at each position of the matrix. The matrix similarity is calculated by

$$\mathrm{mat\_sim} = \frac{\left[\sum_{j=1}^{n} C_i(j) \times \mathrm{score}(b,j)\right]}{\left[\sum_{j=1}^{n} C_i(j) \times \mathrm{max\_score}(j)\right]}$$

where $C_i(j)$ is the C_i-value of position j, n is the length of the matrix, score(b,j) is the matrix value for base b at position j and max_score is the maximum score within a matrix column at position j (for details see Quandt *et al*., 1995).

Both programs MatInd and MatInspector were significantly enhanced.

IMPLEMENTATION

Definition of matrices (MatDefine)

The weight matrices of the MatInspector library are now generated with MatDefine, a tool for automatic definition and evaluation of weight matrices from a set of TFBS. MatDefine extends the original algorithm MatInd in several respects: (1) First, a short highly conserved core sequence that is common to the input sequences is defined. For this purpose the tuple search algorithm developed by Wolfertstetter *et al*. (1996) is used. The algorithm is based on a search for n-tuples (default $n = 4$), which occur at least in a minimum percentage of the sequences (default 90%) with no or one mismatch, which may be at any position of the tuple. Selection of tuples is carried out by maximization of the information content (C_i-value) of the tuples. The tuple with the highest C_i-value is selected as matrix core. (2) The alignment of the binding sites is then anchored at the first position of the identified core sequence. (3) The length of the weight matrix is automatically determined by cutting off low conserved positions ($C_i < 25$) at both matrix ends. (4) All training sequences are checked against the resulting matrix, and sequences with a matrix similarity <0.8 are rejected. (5) This process is repeated until the matrix recognizes all the remaining sequences. No matrix is generated if <5 sequences remain. Rejected sites may be weak binding sites that bind only in context with other TFs as shown in the example in Table 1 or sites not likely to bind the TF. (6) The final matrix is compared with all existing matrices of the library to check whether the new matrix is similar to an already available binding site description. For this purpose all binding sites used for generation of the matrix are searched for TFBS matches with the whole matrix library. The number of binding sites recognized by the same matrix family is used as a measure for the similarity to existing matrices and is displayed as information for the user.

Table 1 illustrates the definition of a matrix for the MatInspector library: 12 binding sites for the TF RUNX2 (runt-related TF 2, AML3) were selected from various papers as the initial training set. The alignment of the 11 sequences that passed the quality checks during matrix generation is also shown (Table 1). The final RUNX2 (AML3) matrix recognizes all sites with a matrix similarity >0.8.

Figure 1 shows the profile of the C_i-vector of the matrix description, i.e. the C_i-values representing the conservation at each position. The three conserved G nucleotides reflect direct protein–DNA contacts as seen in X-ray structure analysis (Tahirov *et al*., 2001). One of the 12 binding sites was rejected during matrix generation as no core sequence was found. The rejected ameloblastin site misses one of these critical contacts, making it unlikely that it is a strong RUNX2 binding site. This notion is further supported by the fact that RUNX2 binds to that site only as part of a larger complex (Dhamija and Krebsbach, 2001). Similar binding of a protein complex to a weak binding site was described in Werner *et al*. (2003).

Search for matrix matches (MatInspector)

MatInspector uses the information of core positions, nucleotide distribution matrix and C_i-vector to scan sequences of unlimited length for pattern matches as described in Quandt *et al*. (1995).

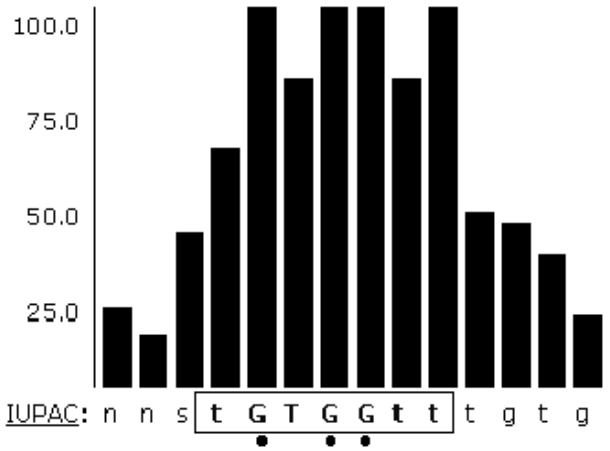

Fig. 1. C_i-vector profile and IUPAC representation for the RUNX2 (AML3) matrix. Nucleotides marked by a box show high information content, i.e. the matrix exhibits a high conservation (C_i-value >60) at these positions. Nucleotides in capital letters denote the core sequence. Circles indicate protein–DNA contacts identified in X-ray structure analysis.

We developed a so-called family concept to minimize redundant matches and optimized matrix thresholds were introduced to reduce false-positives. Both concepts are given in detail below.

The MatInspector family concept It is reasonable to keep different matrix descriptions for a factor in the library as these matrices might be based on different training data originating from independent publications. Similar matrices for one TF can lead to multiple matches at the same position or to matches that are only shifted by a few base pairs if the corresponding matrix descriptions are only partially overlapping or differ in length. A new feature of the matrix library is the assignment of each individual matrix to a family consisting of matrices that represent similar DNA patterns.

The assignment of matrices from the MatInspector library to families involves two steps. First, the matrices are grouped automatically by an unsupervised clustering approach based on a self-organizing map algorithm. For each matrix, a feature vector is calculated based on the summed probabilities of dinucleotides, trinucleotides and tetranucleotides occurring within the matrix. These feature vectors are then used to generate a standard self-organizing map (Kohonen, 1995). The map dimensions were 20 × 15 nodes, reflecting the ratio of the values of the first two principal components of the feature vector distribution. The map was then initialized with random vectors and a two-stage learning process was applied. For the first so-called pre-ordering stage we used an initial neighborhood radius of 8 nodes, a learning ratio of 0.2 and 10 000 adaptation steps. For the second fine-tuning stage, the initial radius was reduced to 3 nodes, the learning rate was set to 0.02 and 80 000 steps were used. The process was repeated with different random initializations. Groups were then selected based on the preservation of distances between the feature vectors of the trained map throughout this repetitive process.

Owing to their high C_i-values the conserved regions of a matrix provide prominent feature values, whereas the remaining positions add valuable secondary contributions, making the approach feasible for a distinctive grouping of the matrices. We found that larger areas of the map also tend to reflect a clustering of matrices of TFs with similar binding domains (bHLH, bZIP, etc.; for a review of binding domains see Garvie and Wolberger, 2001). For example, the zinc finger proteins EGR1, EGR2, EGR3, EGR4 and WT1 are represented by the V$EGRF family. Some protein domain families, such as bZip TFs, are further subdivided by function, e.g. V$AP1F and V$CREB. The functional context of the TF families differs: cAMP response for CREB and early growth/phorbol-ester response for AP1. This underlines the capability of the method to reproduce biologically relevant relations between the different matrices. In a second step, the resulting groups are checked for biological significance and correctness by evaluating the corresponding literature. If new matrices are added to the library, they are mapped onto the existing map to check whether they are similar to an existing family or if a new family has to be created.

After identifying all individual matches, MatInspector now applies a further step and compares the matches of matrices that belong to the same family. The program only lists the match with the highest score of overlapping matches from one family in the output. The family concept leads to a significantly condensed and comprehensive output because redundant matches are eliminated. The concept of matrix families does not compromise the specificity of the individual matrix approach, which would be the case if a combined matrix consisting of all individual matrices was generated.

The following example details the consequences of the introduction of matrix families. The family V$IRFF (interferon regulatory factors) comprises a total of six matrices representing different TFs with highly similar binding sites (IRF1, IRF2, IRF3, IRF4, IRF7 and ISRE). When searching IRF binding sites in the human IL10 promoter with all matrix families, two matches are found, the first being an experimentally verified IRF binding site (Ziegler-Heitbrock *et al*., 2003). A search with all six individual matrices results in four additional redundant matches (Table 2). The highest scoring IRF family match is IRF1 for both matches, but this does not mean that only IRF1 can bind to these sites; all TFs of the IRF family are able to bind to the TFBS identified. The DNA binding sites of the different interferon regulatory factors cannot be distinguished computationally; this is the reason why they are in the same family.

Optimized matrix thresholds With the original version of MatInspector it was apparent that with a fixed matrix similarity threshold some matrices matched very frequently, whereas others hardly appeared at all, even missing true positive matches in evaluation sequences. The reason is the different length and conservation profile of the matrices in the library. Matches to a long and highly conserved matrix have a lower probability to reach any fixed threshold, as compared with matches to short, less conserved matrices. The first case may result in false negative, the latter in false positive matches. Pickert *et al*. (1998) described a strategy to evaluate false positives and negatives. Since there are usually only a limited number of true positives, which are also used in the training process, we concentrated on reducing the number of false positives by introducing a so-called optimized matrix threshold for each individual matrix in the library.

We defined the optimized threshold of a weight matrix as the matrix similarity threshold that allows a maximum of three matches in 10 000 bp of non-regulatory test sequences (1.5 million bp of coding sequences, excluding first exons, and genomic repeats). The latest version of MatInspector uses the optimized matrix threshold for each matrix as default and allows adjustment of thresholds relative to this default. The optimized threshold balances the differences in match frequencies between highly specific or relatively long matrices and less specific or shorter matrices.

Table 2. Searching IRF sites with individual matrices in the human IL10 promoter

Family/matrix	Opt.	Position	Strain	Core sim.	Matrix sim.	Sequence
V$IRFF/IRF1.01	0.86	345–363	(+)	1.000	0.908	caaaaattGAAAactaagt
V$IRFF/IRF3.01	0.85	345–363	(+)	1.000	0.854	caaaaattGAAAactaagt
V$IRFF/IRF7.01	0.86	345–363	(+)	0.768	0.860	caAAAAttgaaaactaagt
V$IRFF/IRF2.01	0.80	345–363	(+)	1.000	0.825	caaaaattGAAAactaagt
V$IRFF/IRF1.01	0.86	521–539	(+)	0.765	0.861	tgcaaaacCAAAccacaag
V$IRFF/IRF2.01	0.80	521–539	(+)	0.750	0.812	tgcaaaacCAAAccacaag

Sites found using matrix families are marked in gray.

Table 3. Searching Sp1 sites with a fixed matrix threshold of 0.85 in the PHGPx promoter

Family/matrix	Further information	Opt.	Position	Strain	Core sim.	Matrix sim.	Sequence
V$SP1F/BTEB3.01	Basic transcription element (BTE) binding protein, BTEB3, FKLF-2	0.93	162–176	(−)	1.000	0.855	gtacaGGAGtctcct
V$SP1F/GC.01	GC box elements	0.88	385–399	(−)	0.764	0.875	ggcggGGCTgggctt
V$SP1F/GC.01	GC box elements	0.88	390–404	(−)	1.000	0.957	gcttgGGCGgggctg
V$SP1F/SP1.01	Stimulating protein 1 SP1, ubiquitous zinc finger TF	0.89	495–509	(+)	1.000	0.891	ccgagGGCGggcaag
V$SP1F/GC.01	GC box elements	0.88	546–560	(+)	0.876	0.921	tgaggGGAGgagccg

Sites exceeding the optimized matrix threshold are marked in gray.

The following example shows that the effect of optimized matrix thresholds is also biologically meaningful. When searching Sp1 binding sites in the murine PHGPx (Gpx4: glutathione peroxidase 4) promoter with optimized matrix similarity, three matches are found. All three Sp1 sites have been reported to be important for expression regulation of the *PHGPx* gene (Ufer *et al*., 2003). Searching with a fixed threshold of 0.85 results in two additional putative Sp1 matches without known function (Table 3).

Another example is the human RANTES promoter (Chr. 17, contig NT_0170799, 8944188–8944857, 670 bp), which has been extensively experimentally analyzed. To date at least 10 TFBS have been functionally verified (Fessele *et al*., 2002). The introduction of the family concept combined with optimized matrix thresholds reduced the number of total matches by 302 (75%) while missing only a single functional binding site (10%) as compared with individual matrices with a fixed threshold of 0.85 (Table 4).

Table 4. Number of MatInspector matches with different settings on 670 bp of the human RANTES (CCL5) promoter

Search mode	Number of matches	Number of functional TFBS (10)
Threshold = 0.85/individual matrices	402	8
Threshold = 0.85/families	291	8
Optimized matrix threshold/individual matrices	151	7
Optimized matrix threshold/families	100	7
Reduction in match number	302 (75%)	1 (10%)

Matrix library

The matrices in the MatInspector library are derived from single publications with either a nucleotide distribution matrix or a list of binding sites or from several papers where individual binding sites were published. In contrast to other approaches, the library is not supposed to represent all literature regarding a single TFBS but rather the best of current knowledge in terms of specificity and sensitivity of the resulting PWM. Thus, putatively erroneous binding sites are not represented in a matrix and the strand orientation for some binding sites might be inverted in comparison with the respective literature. Matrices that do not reach the quality thresholds are automatically removed from the MatInspector library. Quality thresholds are the number of binding sequences (at least 4) and the number of matrix matches expected in a random sequence of 1000 bp (<5).

The current library (version 5.0, February 2005) of MatInspector contains a total of 634 matrices in 279 families (Table 5), divided into the sections vertebrates, plants, fungi, insects and miscellaneous for TFs and a section of others, which contains patterns like PolyA signals. Matrices are not species-specific since TFs have been shown to bind cross-species, e.g. human promoters work very well in mouse as shown in Sarsero *et al.* (2004). The matrices are built from a minimum of 4 and a maximum of 389 binding sites (mean = 26.5). The length of the matrices is between 5 and 29 bp (mean = 16.6).

Until now 1454 human genes have been assigned a TF activity (LocusLink, GeneOntology). Of these, 322 TFs cannot be described by a nucleotide weight matrix as they either have no sequence-specific DNA binding site (e.g. HMG with non-specific DNA binding like chromatin proteins HMG14/17) or do not bind any DNA (e.g. the TAF family of cofactors of the TATA binding protein). A significant number of the remaining 1132 human TFs is represented by a matrix of the MatInspector library (590) or is in the pipeline for the creation

Table 5. Number of matrix families and matrices in the MatInspector library version 5.0

Library section	Number of matrices	Number of families	Families with 1 matrix	Families with 2 matrices	Families with 3 matrices	Families with 4 matrices	Families with >5 matrices
Vertebrates	409	150	67	33	18	6	26
Plants	126	58	35	12	2	3	6
Fungi	43	32	23	8	0	1	0
Insects	40	27	19	5	1	2	0
Miscellaneous	8	7	6	1	0	0	0
Others	8	5	4	0	0	1	0

of a new matrix. However, there is a large number of TFs (~300) for which no binding sites have yet been described in the literature (e.g. the repressor function of KRAB domain zinc finger proteins has been described in fusions with heterologous DNA binding domains yet in lack of natural targets). For the remaining TFs, the currently known bindings sites were insufficient for matrix generation. The MatInspector library will be continuously expanded until all known DNA-binding TFs are represented.

The vertebrate matrix families additionally include information on tissue associations of the TFs. They have been determined by semi-automatic analysis of all PubMed abstracts. A list of synonyms for each of the TFs in a family was generated, using LocusLink. PubMed abstracts were then scanned with these synonym lists and the associated MeSH terms representing tissues were recorded. Assignment of tissues to TFs followed the tree of MeSH terms (MeSH Browser NCBI), i.e. if a co-citation with a specific subtissue was found the parental tissue was also recorded (e.g. Langerhans islets as subtissue and pancreas as parental tissue). Tissue association was primarily done automatically by the following criteria: (1) the number of co-citations of a factor with a specific tissue MeSH term must exceed a fixed threshold, (2) the number of citations for that tissue must exceed a fixed threshold and (3) the co-citation frequency of factor and MeSH term must be overrepresented with regard to the calculated expected co-citation frequency within all abstracts with any synonym and tissue MeSH term.

The automatically generated list of TF–tissue associations was manually checked to eliminate obviously wrong associations arising from ambiguous usage of TF synonyms or extremely unequal distribution of examined tissues. It should be noted that the tissue associations derived statistically from literature do not necessarily reflect tissue-specific expression since gene expression may change drastically in development or disease. However, we found that our scheme of extracting tissue associations from literature works well as a first approximation of tissue-specific TF expression. MatInspector searches can be filtered for TFs associated with specific tissues.

TOOLS BASED ON MATINSPECTOR

Designing regulatory sequences

Regulatory sequences usually contain many sites capable of binding TFs, and the selection, which TFBS are functional for transcriptional control, depends on the biological context. Experimental expression analysis using expression vectors, therefore, may be affected by additional TFBS in the vector as the number and combination of relevant TFBS may be influenced by the experimental conditions. One way to prevent this is to remove all potential TFBS from the sequence of the expression vector except those necessary for the experiment. This is ideally achieved by point mutations, since the deletion of complete binding sites or even single nucleotides can change the distance between functional binding sites and thus influence the interactions of binding proteins.

In addition, potential side effects have to be considered; for example, deleting one binding site might remove additional overlapping binding sites from the sequence or might result in the generation of a new binding site. To make this kind of sequence design feasible even for larger sequences, the program SequenceShaper was developed (available within GEMS Launcher, www.genomatix.de). It systematically evaluates which and how many nucleotide substitutions are required to delete a specified set of TFBS. The modifications can be restricted to preserve other TFBS and to prevent the generation of new sites. Even the coding potential of a sequence (ORF) can be preserved. The number of nucleotides modified is minimized.

Figure 2 shows a representation of a fragment of the promoter of the human *HLA-B* gene. It contains 12 putative TFBS, two of which (NF-kappaB, IRF) are known to be functional in HeLa cells (Johnson and Pober, 1994). In other cell lines with a different composition of TFs the functionality of these sites may be affected owing to the overlapping character of the other potential binding sites. Analyzing the sequence with SequenceShaper results in the exchange of 17 nt. The modified promoter sequence only contains the two TFBSs for NF-kappaB and IRF. Thus, it is possible to modify known regulatory sequences or to design completely new sequences merging different functional elements into a new context.

Analyzing promoters for common TFBS

Although MatInspector can find most true positive TFBS matches and reduce the amount of false positive matches in a promoter region, not all sites found are necessarily functional in the particular biological context. A first step in examining functionality is a comparative promoter analysis. Promoters sharing a common function, e.g. promoters responsive to interferon or a set of actin promoters from different species can be compared. A binding site that occurs in most promoters—particularly at a similar relative position within the promoters—is presumably evolutionarily conserved and this represents supporting evidence that the site may be functional. As a first step to in-depth promoter analysis the 'common sites analysis' together with a concise graphical display of the results was added to MatInspector's capabilities. Displayed are only

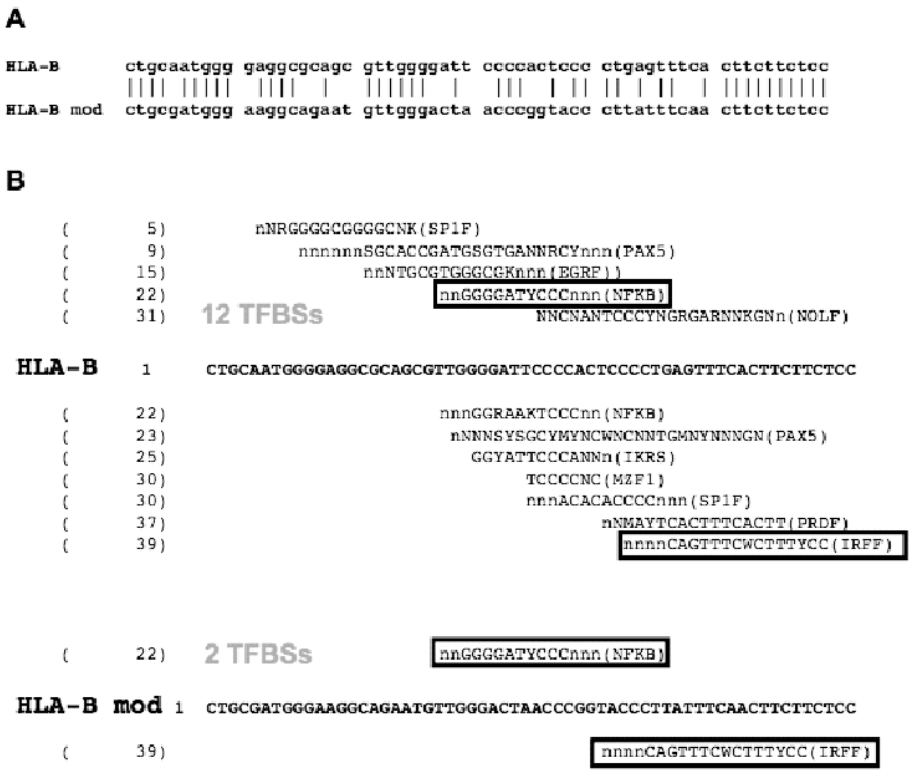

Fig. 2. Targeted elimination of TFBS. (**A**) Alignment of the 60 bp fragment of the HLA-B promoter containing the NFκB/IRF module. The modifications in HLA-B mod were determined by SequenceShaper to remove additional TFBS. (**B**) Overview of TFBS detected by MatInspector in each of the two sequence versions.

matches to matrices that are common to a user-defined percentage of input sequences (default 90%). In the following example a set of six promoter sequences for muscle actin genes from human and mouse was analyzed by MatInspector for TFBS common to all input sequences (Fig. 3).

This analysis for common TFBS in co-regulated promoters can reduce the number of relevant TFs dramatically. Figure 3 already reveals a recurring pattern of TFBS in muscle actin promoters, consisting of 2–4 SRF sites and a TATA box which is conserved throughout all sequences (Klingenhoff *et al*., 1999).

Detecting TFBS in aligned regions of orthologous promoters (DiAlignTF)

The functional conservation of TFBS can also be evident on sequence level. This has been shown recently for the αT-catenin gene (*CTNNA3*) (Vanpoucke *et al*., 2004) and for (Gfi1) growth factor independent 1 (Doan *et al*., 2004). Both publications show a multiple alignment of three orthologous promoters where the conserved TFBS are manually assigned. This task can now be performed automatically by DiAlignTF, a combination of MatInspector with the multiple alignment program DiAlign (Morgenstern *et al*., 1998). DiAlignTF is available from within the GEMS Launcher package at www.genomatix.de.

DiAlign is especially suited for alignment of regulatory sequences where only small fragments are conserved as it is able to detect local similarities in otherwise unrelated sequences. DiAlignTF displays TFBS located at the same position within the alignment as coloured boxes. Optionally, all TFBS identified by MatInspector or TFBS common to a user-defined percentage of the sequences can be displayed alongside the alignment. Similar tools for identification of evolutionarily conserved TFBS are rVista (Loots and Ovcharenko, 2004), ConSite (Sandelin and Wasserman, 2004) and CONREAL (Berezikov *et al*., 2004). However, all these tools are able to only compare two sequences but not multiple sequences like DiAlignTF.

The following example shows how DiAlignTF can be used to reduce the list of potential TFBS to the most likely functional matches. Six rbcS (ribulose bisphosphate carboxylase small subunit) promoters of different plant species are aligned and searched for plant TFBS conserved in the alignment. In each of the six promoter sequences (average length 310 bp) ~30 potential plant TFBS are found using default MatInspector parameters (data not shown). From these 30 TFBS, DiAlignTF identifies only 4 also conserved on sequence level: an I-box element, a G-box element, an ABA response element (ABRE) and the TATA box (Fig. 4). The I-box and the G-box elements are functional TFBS, they have been identified experimentally as light-responsive unit of the rbcS promoter for tomato (Baum *et al*., 1997) and tobacco (Martinez-Hernandez *et al*., 2002). ABRE that overlaps with the G-box is also known to be involved in light regulation.

Finding organizational promoter models

The regulation of genes is precisely controlled by the different combinations of TF-bound sites in various cell types (Werner *et al*.,

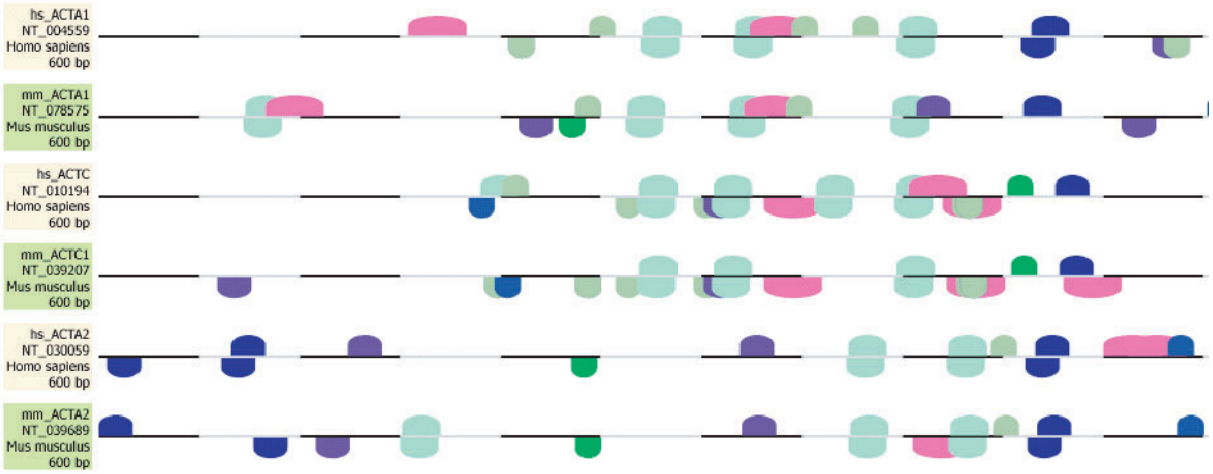

Fig. 3. TFBS common to all sequences in a set of six muscle actin promoters. Different colors denote different transcription factor families (turquoise, SRF; pink, PAX5; blue, TATA-binding protein factor/TBPF; dark green, CMYB; purple, ETS; light green, MAZF).

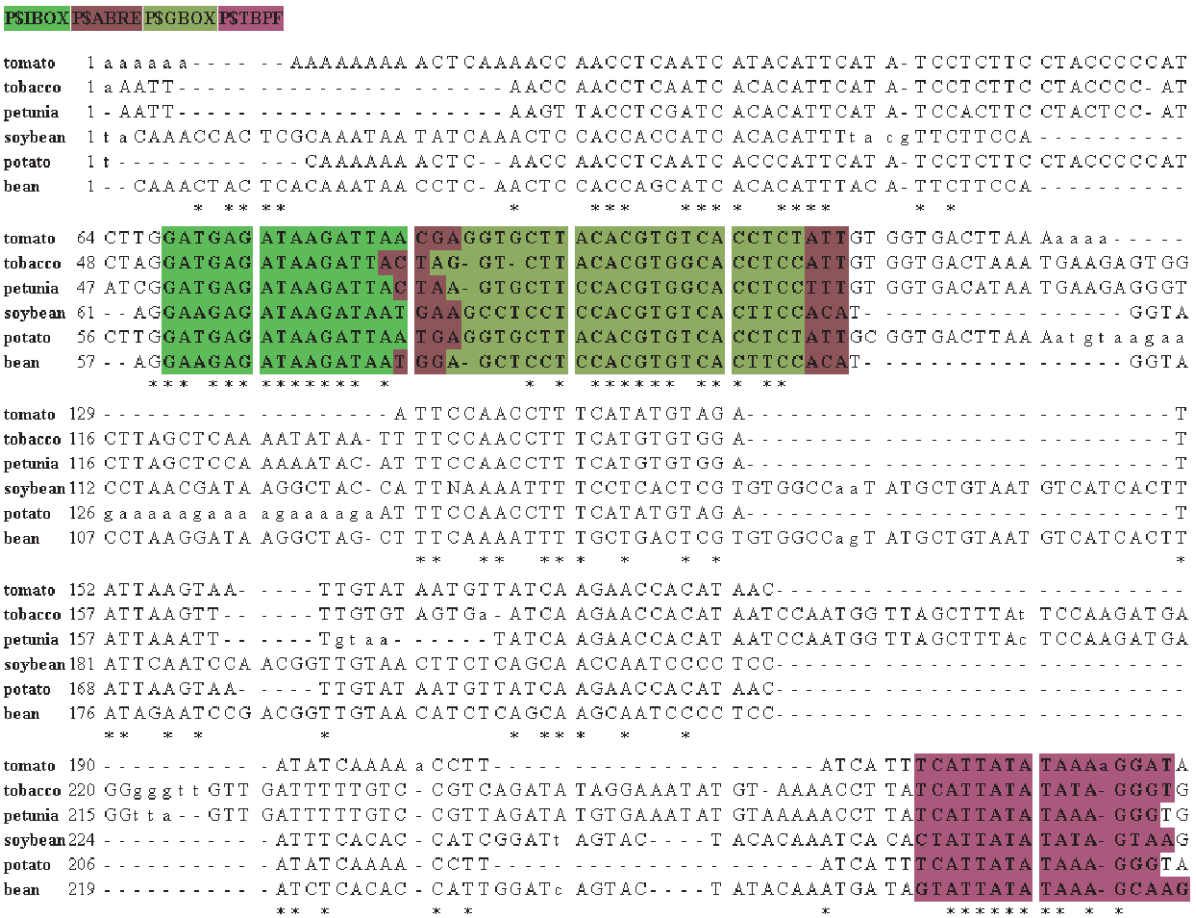

Fig. 4. TFBS conserved in the alignment of six rbcS promoters from different plant species. The color code of the matrix families is shown above the alignment. An asterisk below the alignment indicates identical nucleotides.

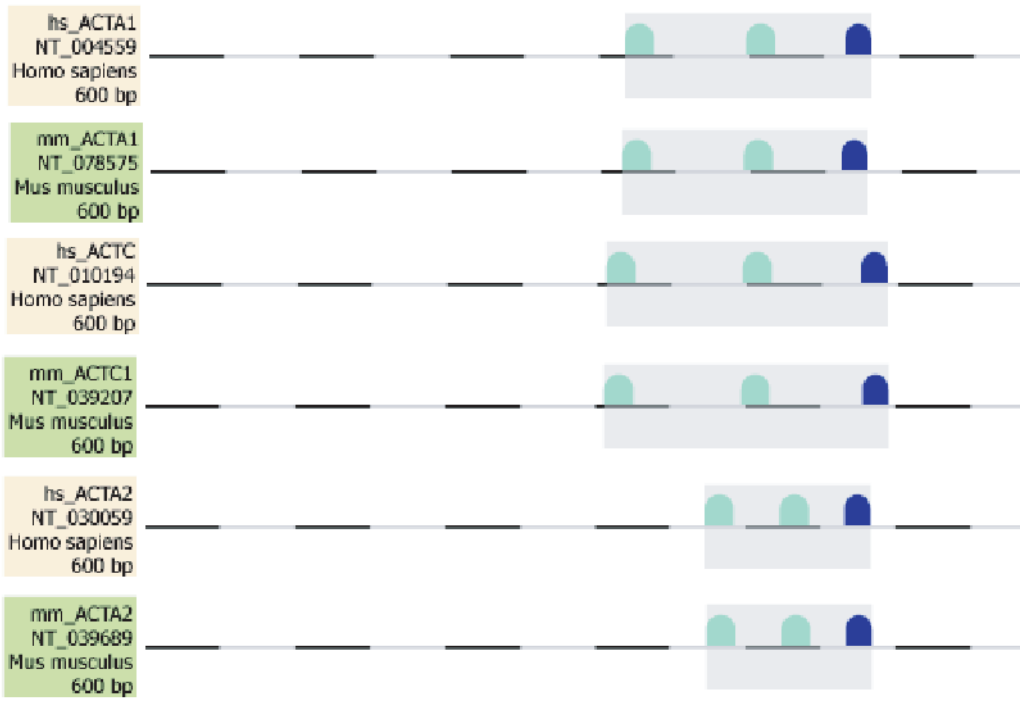

Fig. 5. Matches to a common promoter module found in the entire muscle actin promoter set. The module consists of two SRFs and a TATA site in a conserved distance. SRF and TATA sites are depicted in turquoise and blue, respectively.

2003; Boehlk *et al*., 2000). A prime example is the regulation of the *RANTES*/*CCL5* gene (Fessele *et al*., 2002). The promoter sequence of this chemokine contains six TF binding regions harboring 10 distinct TBFS that were assessed experimentally in five human cell types under both stimulated and unstimulated conditions. It was shown that various subsets of the six binding regions (each containing several TFBS) play a role in the regulation of the gene in different tissues. Each of the TFBS was relevant for some tissues but nonessential in others as was determined by mutation experiments.

The identification of individual matrix matches is, therefore, usually only the first step in promoter analysis; the subsequent aim will be the identification of more complex promoter models, i.e. functional units of a promoter consisting of at least two TFBS in conserved order. The so-called promoter modules form a functional unit, allowing synergistic or antagonistic effects for a specific activation or repression of a gene.

For the automatic detection of modules and more complex models only those combinations of TFBS are analyzed that are situated in a common order and distance from each other. This requires a program that analyzes the TFBS found by MatInspector with respect to their organization, i.e. their relative position. This program, called FrameWorker, is available at www.genomatix.de within the program package GEMS launcher. Constraints to be given by the user are the quorum (i.e. the minimum number of sequences to contain the model), a minimum and a maximum distance between the TF sites within a model. FrameWorker lists all models up to a given number of elements that are found to be common to the input sequences together with a graphical display. A model consists of the matrix families involved, the distance range between the elements and their sequential order.

The common pattern of TFBS automatically found in the muscle actin set consists of two SRF sites in a distance of 40–91 bp, followed by a TATA box in 42–80 bp distance (Fig. 5, parameters: quorum = 100%, 10–100 bp distance between elements). This promoter model is part of a more complex muscle-specific actin promoter model consisting of seven elements including SRF, TATA and SP1 as described and evaluated in Klingenhoff *et al*. (1999).

This promoter model can now be used to scan DNA sequences (e.g. the complete human genome) for the occurrence of pattern matches. When searching a database of 55 207 human promoters (Genomatix promoter database, 35 million basepairs), 49 of them contain the model, 23 of which are promoters of unknown genes corresponding to cDNAs generated by the oligocapping method (Ota *et al*., 2004), 4 are annotated as unknown or hypothetical proteins. The remaining 22 matches include the promoters for alpha1 actin, alpha2 actin, gamma2 actin, myosin, fibulin and tachykinin receptor 2. All these genes are associated with muscle development and are likely to be regulated by a common mechanism involving the TFBS of the model. Searching the promoter database for single matches to SRF or TATA results in 23 620 and 35 145 promoters, respectively with at least one site. Furthermore, 16 514 promoters contain both SRF as well as TATA. This demonstrates that frameworks with their inherent constraints are orders of magnitude more selective than simple co-occurences of matrix matches.

DISCUSSION

Similar to the well-known maps of metabolic pathways that describe pathways potentially used by a cell to accomplish metabolic processes, regulatory networks describe regulator–gene interactions that

show potential pathways a cell can use to control gene expression. Knowing at least parts of these networks is important to understand the molecular underpinnings of cell life, which may later on help to eliminate side effects when developing new drugs. Promoter modules provide the basis to understand regulatory networks involved in gene expression, because they represent the molecular basis for integration of several TF signals into one output (transcription or repression). The most important features of TFBS prediction programs are the highest possible coverage of known sites and the quality of matrix descriptions in the library. The MatInspector matrix library is designed to represent the best of current knowledge in terms of specificity and sensitivity regarding TFBS. The TRANSFAC database (Matys *et al*., 2003) focuses on representing all available publications. There is a subset of ~70% of the matrices in TRANSFAC that are marked as high quality matrices. The JASPAR database (Sandelin *et al*., 2004a) is mostly based on SELEX experiments, leading to very specific matrix descriptions not necessarily reflecting binding sites in their genomic context.

There have been several approaches to replace or extend the matrix model by methods taking into account dependencies between nucleotides at different positions within the binding site description (e.g. via hidden markov models). However, the number of sites required for training these methods ranges from at least several dozens (Locker *et al*., 2002; Ellrott *et al*., 2002) to several thousands (Roulet *et al*., 2002). For many TFs only far less samples are available, making it impossible to apply such methods to the majority of binding sites without extensive laboratory work to gain the required data. Moreover, even if there is a sufficient number of samples for training, methods using positional dependencies do not always lead to improved results (Barash *et al*., 2004). The matrix approach, therefore, still offers a feasible and valid method to detect binding sites for a maximum possible number of different TFs.

Sandelin and Wasserman (2004) introduced a concept for related families of TFs and constructed 11 familial binding profiles. Their main goal was to predict the structural class of TFs interacting with newly characterized binding sites and to enhance the sensitivity of *de novo* pattern discovery methods. Our family concept is intended to reduce redundancy within the matrix library. Schones *et al*. (2005) provided a methodology to compare frequency matrices allowing the grouping of matrices into families. They generated 145 representatives for families that were based on TRANSFAC (7.2) extended core matrices and 36 representatives for the JASPAR database. We preserve the original matrix information by assigning each matrix to a family because searching with a representative would result in a loss in specificity especially at flanking positions that do play a role in discriminating similar but different binding sites [like in the EBOX family where SREBP (sterol regulatory element binding protein) and Myc/Max factors are present].

MatInspector can find the potential binding sites of various activators and repressors that bind to specific DNA regulatory sequences. The concurrent expression of genes as observed in expression array analysis, in particular, is in part orchestrated by sets of common regulatory elements. These regulatory modules can be discovered as shown in this paper and used to identify additional target genes with similar regulatory properties in genomic sequence databases.

It should be stressed that such an analysis is complicated by the fact that co-expressed genes in an expression array experiment are not all necessarily co-regulated, as different regulation mechanisms can lead to the same expression pattern. Effects like a secondary cascade of transcription activation during the time course can divide a co-expressed cluster of genes in subsets regarding co-regulation. The careful selection of gene clusters is a crucial step for successful promoter analysis.

Although there is a consensus that the inspection of single TFBS in promoter sequences is not sufficient to fully understand gene regulation, it will remain one of the crucial first steps in the chain of analytical events. FrameWorker, DiAlignTF or SequenceShaper which are based on MatInspector are results of a consequent follow up leading to understanding the regulatory networks on a molecular level.

ACKNOWLEDGEMENTS

We thank Matthias Scherf and Martin Seifert for a critical reading of the manuscript and their helpful suggestions. Part of this work was supported by the BFAM ring funding project of the BMBF grant number 031U112B / 031U212B ‘Analysis of regulatory regions’.

REFERENCES

Alvarez,M. *et al*. (1997) Rat osteoblast and osteosarcoma nuclear matrix proteins bind with sequence specificity to the rat type I collagen promoter. *Endocrinology*, **138**, 482–489.

Barash,Y. *et al*. (2004) CIS: compound importance sampling method for protein–DNA binding site *P*-value estimation. *Bioinformatics*, **5**, 596–600.

Baum,K. *et al*. (1997) Improved ballistic transient transformation conditions for tomato fruit allow identification of organ-specific contributions of I-box and G-box to the RBCS2 promoter activity. *Plant J.*, **12**, 463–469.

Berezikov,E. *et al*. (2004) CONREAL: conserved regulatory elements anchored alignment algorithm for identification of transcription factor binding sites by phylogenetic footprinting. *Genome Res.*, **14**, 170–178.

Boehlk,S. *et al*. (2000) ATF and Jun transcription factors, acting through an Ets/CRE promoter module, mediate lipopolysaccharide inducibility of the chemokine RANTES in monocytic Mono Mac 6 cells. *Eur. J. Immunol.*, **30**, 1102–1112.

Bucher,P. (1990) Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol.*, **212**, 563–578.

Chen,Q.K. *et al*. (1995) MATRIX SEARCH 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices. *Comput. Appl. Biosci.*, **11**, 563–566.

D’Alonzo,R.C. *et al*. (2002) Physical interaction of the activator protein-1 factors c-Fos and c-Jun with Cbfa1 for collagenase-3 promoter activation. *J. Biol. Chem.*, **277**, 816–822.

Dhamija,S. and Krebsbach,P.H. (2001) Role of Cbfa1 in ameloblastin gene transcription. *J. Biol. Chem.*, **276**, 35159–35164.

Doan,L.L. *et al*. (2004) Targeted transcriptional repression of Gfi1 by GFI1 and GFI1B in lymphoid cells. *Nucleic Acids Res.*, **32**, 2508–2519.

Dyer,K.D. and Rosenberg,H.F. (2001) Transcriptional regulation of galectin-10 (eosinophil Charcot-Leyden crystal protein): a GC box (−44 to −50) controls butyric acid induction of gene expression. *Life Sci.*, **69**, 201–212.

Elkon,R. *et al*. (2003) Genome-wide *in silico* identification of transcriptional regulators controlling the cell cycle in human cells. *Genome Res.*, **13**, 773–780.

Ellrott,K. *et al*. (2002) Identifying transcription factor binding sites through Markov chain optimization. *Bioinformatics*, **18**(Suppl 2), S100–S109.

Fessele,S. *et al*. (2002) Regulatory context is a crucial part of gene function. *Trends Genet.*, **18**, 60–63.

Garvie,C.W. and Wolberger,C. (2001) Recognition of specific DNA sequences. *Mol. Cell*, **8**, 937–946.

Johnson,D.R. and Pober,J.S. (1994) HLA class I heavy-chain gene promoter elements mediating synergy between tumor necrosis factor and interferons. *Mol. Cell. Biol.*, **14**, 1322–1332.

Kel,A.E. *et al*. (2003) MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.*, **31**, 3576–3579.

Klingenhoff,A. *et al*. (1999) Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity. *Bioinformatics*, **15**, 180–186.

Kodadek,T. (1998) Mechanistic parallels between DNA replication, recombination and transcription. *Trends Biochem. Sci.*, **23**, 79–83.
Kohonen,T. (1995) *Self Organizing Maps*. Springer Verlag, Heidelberg.
Lander,E.S. *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
Locker,J. *et al*. (2002) Definition and prediction of the full range of transcription factor binding sites–the hepatocyte nuclear factor 1 dimeric site. *Nucleic Acids Res.*, **30**, 3809–3817.
Loots,G.G. and Ovcharenko,I. (2004) rVISTA 2.0: evolutionary analysis of transcription factor binding sites. *Nucleic Acids Res.*, **32**, W217–W221.
Martinez-Hernandez,A. *et al*. (2002) Functional properties and regulatory complexity of a minimal RBCS light-responsive unit activated by phytochrome, cryptochrome, and plastid signals. *Plant Physiol.*, **128**, 1223–1233.
Matys,V. *et al*. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
Mengshol,J.A. *et al*. (2001) IL-1 induces collagenase-3 (MMP-13) promoter activity in stably transfected chondrocytic cells: requirement for Runx-2 and activation by p38 MAPK and JNK pathways. *Nucleic Acids Res.*, **29**, 4391–4372.
Morgenstern,B. *et al*. (1998) DIALIGN: finding local similarities by multiple sequence alignment. *Bioinformatics*, **14**, 290–294.
O'Brien,C.A. *et al*. (2002) Cbfa1 does not regulate *RANKL* gene activity in stromal/osteoblastic cells. *Bone*, **30**, 453–462.
Ota,T. *et al*. (2004) Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nat Genet.*, **36**, 40–45.
Pickert,L. *et al*. (1998) Transcription regulatory region analysis using signal detection and fuzzy clustering. *Bioinformatics*, **14**, 244–251.
Prestridge,D.S. (1996) SIGNAL SCAN 4.0: additional databases and sequence formats. *Comput. Appl. Biosci.*, **12**, 157–160.
Quandt,K. *et al*. (1995) MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.*, **23**, 4878–4884.
Roulet,E. *et al*. (2002) High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. *Nat. Biotechnol.*, **20**, 831–835.
Sandelin,A. and Wasserman,W.W. (2004) Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *J. Mol. Biol.*, **338**, 207–215.
Sandelin,A. *et al*. (2004a) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**(Database issue), D91–D94.
Sandelin,A. *et al*. (2004b) ConSite: web-based prediction of regulatory elements using cross-species comparison. *Nucleic Acids Res.*, **32**, W249–W252.
Sarsero,J.P. *et al*. (2004) Human BAC-mediated rescue of the Friedreich ataxia knockout mutation in transgenic mice. *Mamm. Genome*, **15**, 370–382.
Schones,D.E. *et al*. (2005) Similarity of position frequency matrices for transcription factor binding sites. *Bioinformatics*, **21**, 307–313.
Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
Tahirov,T.H. *et al*. (2001) Structural analyses of DNA recognition by the AML1/Runx-1 Runt domain and its allosteric control by CBFbeta. *Cell*, **104**, 755–767.
Thirunavukkarasu,K. *et al*. (2001) Stimulation of osteoprotegerin (*OPG*) gene expression by transforming growth factor-beta (TGF-beta). Mapping of the OPG promoter region that mediates TGF-beta effects. *J. Biol. Chem.*, **276**, 39241–39250.
Ufer,C. *et al*. (2003) Functional characterization of *cis*- and *trans*-regulatory elements involved in expression of phospholipid hydroperoxide glutathione peroxidase. *Nucleic Acids Res.*, **31**, 4293–4303.
Vanpoucke,G. *et al*. (2004) GATA-4 and MEF2C transcription factors control the tissue-specific expression of the alphaT-catenin gene *CTNNA3*. *Nucleic Acids Res.*, **32**, 4155–4165.
Venter,J.C. *et al*. (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
Werner,T. *et al*. (2003) Computer modeling of promoter organization as a tool to study transcriptional coregulation. *FASEB J.*, **17**, 1228–1237.
Willis,D.M. *et al*. (2002) Regulation of osteocalcin gene expression by a novel Ku antigen transcription factor complex. *J. Biol. Chem.*, **277**, 37280–37291.
Wolfertstetter,F. *et al*. (1996) Identification of functional elements in unaligned nucleic acid sequences by a novel tuple search algorithm. *Comput. Appl. Biosci.*, **12**, 71–80.
Ziegler-Heitbrock,L. *et al*. (2003) IFN-alpha induces the human *IL-10* gene by recruiting both IFN regulatory factor 1 and Stat3. *J. Immunol.*, **171**, 285–290.