DOI: 10.1093/nar/gkg548

PatSearch: a program for the detection of patterns and structural motifs in nucleotide sequences

Giorgio Grillo, Flavio Licciulli, Sabino Liuni, Elisabetta Sbisà and Graziano Pesole[1],*

Sezione di Bioinformatica e Genomica di Bari, Istituto Tecnologie Biomediche CNR, via Amendola 168/5,
70125 Bari, Italy and [1]Dipartimento di Fisiologia e Biochimica Generali, Università di Milano, via Celoria 26,
20133 Milano, Italy

Received February 12, 2003; Revised and Accepted March 27, 2003

ABSTRACT

Regulation of gene expression at transcriptional and post-transcriptional level involves the interaction between short DNA or RNA tracts and the corresponding trans-acting protein factors. Detection of such *cis*-acting elements in genome-wide screenings may significantly contribute to genome annotation and comparative analysis as well as to target functional characterization experiments. We present here PatSearch, a flexible and fast pattern matcher able to search for specific combinations of oligonucleotide consensus sequences, secondary structure elements and position-weight matrices. It can also allow for mismatches/mispairings below a user fixed threshold. We report three different applications of the program in the search of complex patterns such as those of the iron responsive element hairpin-loop structure, the p53 responsive element and a promoter module containing CAAT-, TATA- and cap-boxes. PatSearch is available on the web at http://bighost.area.ba.cnr.it/BIG/PatSearch/.

INTRODUCTION

The detection of regulatory elements controlling gene expression at the transcriptional and post-transcriptional level, generally embedded in the non-coding part of the genome, represents a major challenge in post-genomic biology. Transcriptional or post-transcriptional control of gene expression generally involves short DNA or RNA tracts, respectively interacting with transcription factors or RNA-binding proteins.

Genetic information for transcriptional control is mostly defined by a series of *cis*-acting DNA elements, such as promoters, enhancers, silencers and locus control regions, organized in a modular structure whose specific arrangement determines the expression specificity in a spatio-temporal framework.

Post-transcriptional regulation of gene expression in eukaryotes is mostly exerted by 5′ and 3′ untranslated regions (5′ UTR, 3′ UTR) and includes modulation of mRNA nucleo-cytoplasmic transport, translation efficiency, subcellular localization and stability. UTR-mediated regulatory activity generally involves specific interactions between RNA-binding proteins and specific RNA elements whose biological activity relies on a combination of primary and secondary structure patterns. A large number of oligonucleotide patterns involved in transcriptional and post-transcriptional regulation have been experimentally characterized and this information collected in a number of specialized databases such as TRANSFAC (1) for promoters and UTRsite (2) for UTR-specific functional motifs.

Transcription factor (TF) binding sites are typically 5–15 nt long and their consensus is generally described by position weight matrices (PWMs) that assign a weight to each possible nucleotide in each position of the putative binding site based on the observed occurrence of that nucleotide in the known promoter elements. Several publicly available programs to locate promoter elements in DNA sequences have been developed (3) but they do not show a satisfactory level of accuracy because of the remarkable level of degeneracy of single promoter elements that contaminate prediction with a huge number of false positives. A remarkable increase in selectivity can be obtained by considering two or more TF-binding sites with a defined order, orientation and spacing. UTR-localized structural elements whose biological activity has been demonstrated experimentally include the iron responsive element (IRE) (4), the histone 3′ UTR structure (5) and many others which play crucial functional roles (6).

Thus, it is of utmost importance to develop specific software tools that are able to identify the above described functional elements in genomic or cDNA sequences and significantly contribute to their functional characterization. We present here the PatSearch software which is able to scan user submitted sequences for any combination of PWMs, primary sequence patterns and structural motifs also allowing mismatches and/or mispairings below a user fixed threshold.

SYSTEM AND METHODS

The PatSearch program is written in C++ language and runs under different Unix operating systems. Accepted input formats include FASTA, EMBL, GenBank and others.

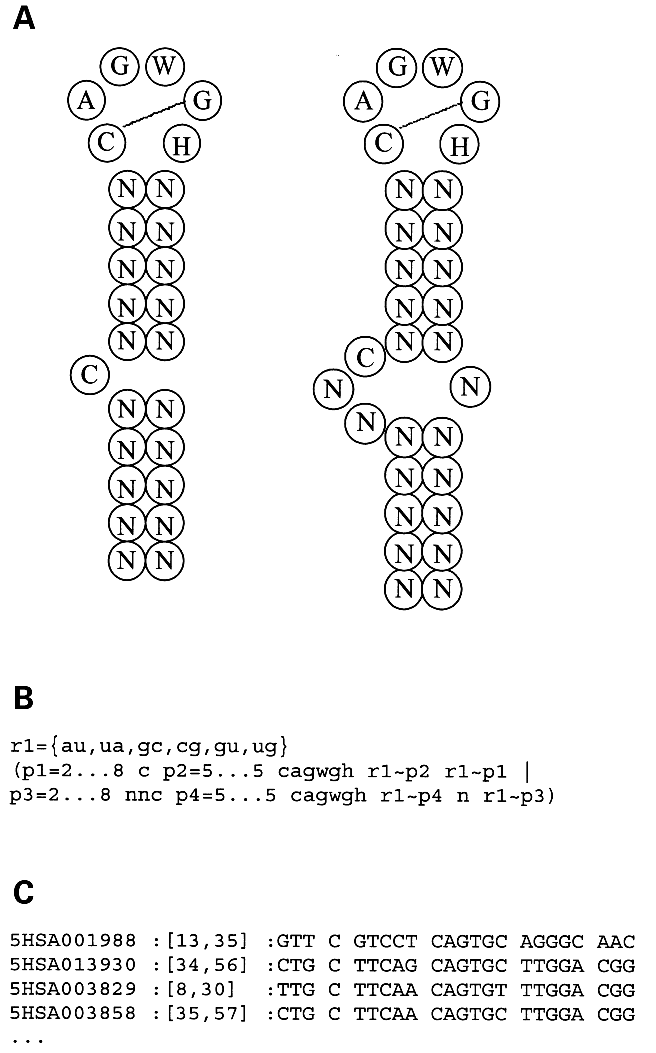
*To whom correspondence should be addressed. Tel: +39 02 50314915; Fax: +39 02 50314912; Email: graziano.pesole@unimi.it

Standard IUB codes for ambiguous nucleotides (e.g. R for A or G) are also recognized.

The pattern description has been inspired by 'regular expression' rules, although both the syntax and the semantics are different, especially for the inclusion of specific operators for finding PWMs, complementary helices and palindromes. Below we clarify what we mean by a pattern and how the program locates the subsequences matched by it. A pattern is made by a combination of pattern units that may or may not be named. Names used for pattern unit are p1, p2, p3, etc., using lower case letters and need to be assigned if one or more reprocessing steps of already matched sequences have to be carried out.

We describe below possible pattern units.

1. *String pattern unit.* This is a string of characters that may include ambiguous characters. Two examples of string pattern units are AAUAAA, the polyadenylation site; or TTTSSCGS (S = C or G) the consensus site for E2F transcription factor (TRANSFAC site ID E2F$CONS_01). A pattern unit qualifier [*mismatches, deletions, insertions*] can be added with a specification of the number of mismatches, deletions and insertions allowed in the matched sequence. For example, TTATTTATT[1,0,0] would match any sequence with up to one mismatch with the sequence TTATTTATT. Loosely defined motifs, simultaneously allowing several mismatches, deletions and insertions may considerably slow down the pattern searching with a very large number of different sequences.
2. *Range pattern unit.* This has the form *min . . . max* (e.g. 1 . . . 200: match any subsequence from 1 to 200 characters). The range pattern unit will first match the minimum number of residues.
3. *Start of sequence.* The symbol ^ matches only at the start of a sequence (e.g. ^ATG matches ATGACCT but not ACCTATG).
4. *End of sequence.* The symbol $ matches only at the end of the sequence (e.g. ATG$ matches ACCTATG but not ATGACCT).
5. *Palindrome pattern unit.* A palindrome of a subsequence previously matched as p1 is matched as <p1 (e.g. p1 = 4 . . . 4 <p1 matches the subsequence ACGTTGCA).
6. *Hairpin loop pattern unit.* The pattern unit of the form ~p1 matches the reverse complement of the subsequence recorded in p1 (e.g. the pattern p1 = 6 . . . 7 3 . . . 6 ~p1 matches a stem-loop structure with a stem 6–7 nt long and a loop 3–6 nt long). The user can define non-standard pairing rules, allowing for example GU/UG pairings. One or more pairing rules can be defined, named r1, r2, etc., using the format: r1 = {au, ua, cg, gc, gu, ug}. To use a specific pairing rule (e.g. r1) the format for the reverse complement pattern is r1~p1.



```
r1={au,ua,gc,cg,gu,ug}
(p1=2...8 c p2=5...5 cagwgh r1~p2 r1~p1 |
p3=2...8 nnc p4=5...5 cagwgh r1~p4 n r1~p3)
```

```
5HSA001988 :[13,35]  :GTT C GTCCT CAGTGC AGGGC AAC
5HSA013930 :[34,56]  :CTG C TTCAG CAGTGC TTGGA CGG
5HSA003829 :[8,30]   :TTG C TTCAA CAGTGT TTGGA CGG
5HSA003858 :[35,57]  :CTG C TTCAA CAGTGC TTGGA CGG
...
```

Figure 1. The consensus structures of the iron responsive elements (IRE) (**A**); the relevant PatSearch pattern syntax (**B**); part of PatSearch output obtained on a set of 5′ UTRs collected in UTRdb (**C**).

7. *Position weight matrix (PWM) pattern unit.* A PWM can be used as a pattern unit. PWMs are used in the following format, similar to that adopted in the TRANSFAC database (1), setting a similarity threshold ≤1. The degree of similarity between the PWM and the scanned sequence is calculated by the method described by Werner (7). It is also possible to define a higher similarity threshold for a core portion of the PWM. In this case core positions are marked by * and a second threshold value is defined:

   ```
   {
   01       12     22     17     199     T
   02      210     12      9      19  *  A
   03       46     32     36     136  *  T
   04      159     33     29      29  *  A
   05      158     37     31      24  *  A
   06       11     21      9     209     T
   } > 0.70, 0.90
   ```

   Alternatively, a normalized log-odds weight matrix can be used as described in Bucher (8) providing as a threshold the optimized cutoff score. This has the format {* . . . *} > cutoff. The tilde ~ qualifier before the matrix (e.g. ~ {PWM}) allows the user to look for PWMs in the reverse orientation.
8. *Repeat pattern unit.* This has the form : *max > repeat (pattern) range > min*, where *min* and *max* are integers

A

```
p1=rrrcwwgyyy[3,0,0] p2=0...13 p3=rrrcwwgyyy[3,0,0]
p4=0...13 p5=rrrcwwgyyy[3,0,0] p6=0...13 p7=rrrcwwgyyy[3,0,0]
length(p2+p4+p6)<20
p1/p3/p5/p7:(p8=nnncnngnnnnnncnngnnnnnncnngnnnnnncnngnnn)
p1/p3/p5/p7:(p9=rrrcwwgyyyrrrcwwgyyy[3,0,0])
```

B

```
9EP11070   : [117,175] : GAGCAGGCGG T             GCACTCGGCC CACGGGGAACTGG AGACCGGCCC TAGAA        GAGCGAGTCT
31EP11083  : [358,410] : AGGCTCGTAA A             AATCTTGTAT GGCTGC        AGGCAAGCCA AACCCT       TGACAGGCAC
46EP30021  : [400,456] : GGACCAGTGA GCAG          CAACAGGGCC G             GGGCTGGGCT TATCAGCCTCCC AGCCCAGACC
61EP11091  : [86,144]  : GGCCAGGACT GTCCTG        GGGCCAGCCG GGGCACCTGGT   GGCCAAGCTT AG           AAACATGACA
68EP07077  : [367,413] : TGGCCAGCCT               TGCCTTGACC AAT           AGCCTTGACA AGGC         AAACTTGACC
86EP47007  : [173,225] : ACCCTTGGCC TT            ATTCTGGTCT ACTGAGCTGG    GAGCTTGTCT G            AGGCTGGAGC
126EP70008 : [132,184] : GATCTAGCTA TGG           GGTCAAGCCT GGAGGGG       ACGCTGGTTC TCC          AGACATGGTC
134EP17067 : [104,157] : AGACTTGTAA GAACCTCAAATGA GGACATGCAC A             AAACAGGGAT              GGCCATGGGC
173EP11139 : [53,111]  : TGACTTGCCC AAGGTG        ACCCAAGCTC CCGAGTGCCA    GGGCAGGATC TGA          ATTCAGGCTC
190EP25039 : [522,580] : GAACTGGCAG GCA           CCGCGAGCCC CTAGCACC      CGACAAGCTG AGTGTGCA     GGACGAGTCC
196EP48004 : [176,227] : GTGCAGGTGT GT            GTGCAGGTGT GTGTGC        AAACATGCAC ACGC         GTGCAAGCAT
207EP25010 : [535,583] : CGACGAGCGC CGGGGCA       AGGCAAGCCC T             GGACGGGATT G            CGACGTGCGC
238EP25050 : [394,440] : AAACAGGCTT CA            AAGCAAGGGC T             TGGCTGGCAC ACAG         GGGCTTGGTC
245EP14076 : [256,308] : AGGCCTGCCC AG            AAACAAGTGA TGA           GGGCCTGGGC AGCCAATG     GATCGTGCTG
274EP16071 : [306,361] : CGGCAAGGTC ACA           AGACATGCTT AAGTAAGATAG   GGTCATGTTG CA           AATCCTGTTG
```

Figure 2. PatSearch syntax for the p53 responsive element including two post-processing steps and a constraint on the total length of decamer spacers (**A**); partial output obtained searching the pattern in the EPD database (13) (**B**).

defining the minimum and maximum number of pattern repeats and *range* is a range pattern unit (see above) fixing the accepted distance between two repeats. In this case the repeat pattern (e.g. a stem-loop structure) has, in general, a different nucleotide sequence. To search for exact repeats, i.e. with a sequence identical to that of the first matching repeat element, the operator *frepeat* has to be used instead. For example the syntax: *20 > frepeat (p1 = NNN) 0...0 > 10* defines a sequence string of 11–19 identical tandem trinucleotides.

9. *Either/or pattern unit*. This has the form (pi|pj) which matches either pi or pj (i, j integers). The alternatives may be themselves complex patterns made up of more than one pattern unit.
10. *Length constraints*. Length constraints can be settled for specific combinations of pattern units. These may have the form:

$$\min < \text{length}(\text{pi} + \text{pj} + \text{pk} + \cdots) < \max$$
$$\text{length}(\text{pi} + \text{pj} + \cdots)/\text{length}(\text{pk} + \text{pn} + \cdots) > \min$$
$$\text{length}(\text{pi} + \text{pj} + \cdots)\ \text{MOD value}$$

with *i*, *j*, *k*, *n*, *value*, *min* and *max* integer numbers. For example the pattern:

$$\text{AUG p1} = 0\ldots300\ ((\text{UAA}|\text{UAG})|\text{UGA})$$
$$\text{length (p1) MOD 3}$$

can be used to search open reading frames <303 nt long in genomic sequences.

Post-processing

It may be very convenient to be able to reprocess a section of a sequence that has been already matched. For example consider the following pattern:

$$\text{p1} = \text{ATGCCGTA}[1, 0, 0]$$

where up to one mismatch is tolerated. To prevent mutation in the two Gs in position 3 and 6 a post-processing step can be carried out in which sequence hits from the first pass is processed adding further constraints. The pattern syntax to be used in this case is:

$$\text{p1} = \text{ATGCCGTA}[1, 0, 0]\ \text{p1:(p2} = \text{nnnGnnGnn)}$$

In general the post-processing syntax has the form *list:(subpattern)* where 'list' is a list of a concatenated named pattern units in the form pi/pj/pk etc. More than one post-processing step can be carried out.

APPLICATION

We report below three typical applications of PatSearch. The first one shows the search for specific *cis*-acting elements located in the 5′ UTR or 3′ UTR of eukaryotic mRNAs, which may play crucial roles in the post-transcriptional regulation of gene expression. These elements are collected in the UTRsite database (2) and annotated through PatSearch analysis of UTRdb, a specialized database collecting eukaryotic mRNA untranslated regions. The present release of UTRsite (March 2003) collects 25 regulatory elements that are annotated in a total of 61 177 UTRdb entries. The 5′ or 3′ UTR elements usually correspond to short oligonucleotide tracts, which generally fold into conserved secondary structures and are the binding site for specific regulatory

A

```
p1={
01      56      55      12      52      N
02      32      52      43      48      N
03      25      47      24      79      N
04      102     1       70      2       R
05      51      6       99      19      R
06      0       173     1       1   *   C
07      0       174     0       1   *   C
08      175     0       0       0   *   A
09      119     8       21      27  *   A
10      17      0       15      143 *   T
11      23      90      59      3       S
12      116     6       52      1       A
}>0.75,0.90
0...100
p2={
01     61    145    152     31      S
02     16     46     18    309    * T
03    352      0      2     35    * A
04      3     10      2    374    * T
05    354      0      5     30    * A
06    268      0      0    121      A
07    360      3     20      6      A
08    222      2     44    121      W
09    155     44    157     33      R
10     56    135    150     48      N
11     83    147    128     31      N
12     82    127    128     52      N
13     82    118    128     61      N
14     68    107    139     75      N
15     77    101    140     71      N
} > 0.75,0.90
10...40
p3={
01    49    48    69   137    N
02     0   303     0     0    C
03   288     0     0    15    A
04    26    81   116    80    N
05    77    95     0   131    H
06    67   118    46    72    N
07    45    85    73   100    N
08    50    96    56   101    N
}>0.90
```

B

```
6EP49001  : [442,506] : ATGCACCAATCA ... CTATATAAGGCCCCG ... TCATGCTT
7EP30042  : [443,502] : CTCCACCAATCA ... TTATATAAGCCCGGG ... GCAGCAAC
8EP11068  : [421,506] : ATTAGCCAATGG ... GTATAAATACTTCTC ... TCATTACA
9EP11070  : [413,503] : AACTACCAATCA ... CTATATAAAAGCGCC ... TCACGCTG
10EP11073 : [396,502] : ATGGACCAATCC ... CTATAAAAGAAGAGT ... CCACAGAC
15EP31009 : [408,509] : CTGGGCCAATGA ... CTATAAAAACTTTAT ... GCAGTGTG
26EP14031 : [388,504] : CAAGGCCAATCA ... GTATATAAGCGTTGG ... GCACTCTG
29EP17045 : [407,504] : CGGGGCCAATCG ... CTATAAAACCCAGCG ... CCACCNNN
...
-----------------------------------------------------------------------
 Match Total = 23                          Matched Sequences = 23 / 260
-----------------------------------------------------------------------
```

Figure 3. (**A**) PatSearch syntax for a promoter module made of a CAAT-box, a TATA-box and a cap-box (TRANSFAC IDs V$CAAT_01, V$TATA_01, V$CAP_01). The * denotes core promoter positions to which a higher similarity cutoff is assigned. (**B**) Partial output obtained searching the pattern in human EPD sequences (13).

proteins. The pattern description syntax of PatSearch is particularly suitable for modeling the consensus structure of such functional elements.

Among the *cis*-acting oligonucleotide patterns located in mRNA UTRs the IRE is among the more extensively studied and better characterized (4). The IRE is a conserved hairpin loop structure located in the 5′ or 3′ UTR of various mRNAs involved in cellular iron homeostasis that function regulating mRNA translation or stability through its specific interaction with iron regulatory proteins (IRPs).

Figure 1 shows the derived IRE consensus structure (Fig. 1A) and the corresponding PatSearch pattern (Fig. 1B). Two alternative IRE consensuses have been proposed both showing a bipartite stem-loop interrupted by a bulged C or by a small internal loop. Some evidence also suggests a structured loop with an interaction between the first (C1) and the fifth (G5) nucleotide. Figure 1C shows the output obtained by the application of PatSearch to a set of 5′ UTRs in the human division of UTRdb.

The second example regards the search for the recognition site of transcription factors belonging to the family of oncosuppressor p53 homologs, including p63 and p73 (9–11), which play a central role in the control of cell cycle and apoptosis. The consensus sequence of the p53 responsive element (p53RE) is composed of two or more decamers having the consensus RRRCWWGYYY (R = A, G; W = A, T; Y = C, T) separated by 0–13 bp and absolutely conserving positions 4 (C4) and 7 (G7). A maximum of three mismatches are allowed in two contiguous decamers. There are often other decamers, each tolerating up to three mismatches (12).

Figure 2A shows a PatSearch syntax devised for p53RE where two post-processing steps are required to force conservation of C4 and G7 and the maximum number of three mismatches for the core decamer pair. In the example shown a length constraint is settled for spacer nucleotides. A sample output of the search of p53RE on a set of human promoter regions is shown in Figure 2B.

The third application example regards the search for PWMs. To understand mechanisms of transcription regulation it is essential to characterize promoters in large scale genome analyses. The remarkable degree of degeneracy of single promoter elements severely hampers their prediction and results in the recovery of a large number of false positives. For this reason it is advisable to search instead for promoter modules made of two or more promoter elements in the correct arrangement and orientation. This task can be accomplished by using specific combinations of PWMs in PatSearch analyses. Figure 3 shows the PatSearch syntax devised to describe a promoter module composed of a CAAT-box, a TATA-box and a cap-box, and the resulting output on a sample of human promoters collected in the EPD database (13).

DISCUSSION

The flood of genomic data produced in recent years requires the development of suitable bioinformatics tools for analysis and annotation. The availability of a suitable pattern matching tool may greatly help this task through the indication of specific experimental work to validate *in silico* predictions. To this aim we developed PatSearch (14) which in the present version implements many new features including the possibility of simultaneously searching for several kinds of pattern units including oligonucleotide consensi, secondary structure motifs and PWMs. The use of a combination of pattern units may greatly increase search selectivity by greatly reducing the rate of false positives. Furthermore, one or more post-processing steps can be carried out on matching patterns allowing the addition of further constraints and accelerating the search process.

Other tools such as FindPatterns (GCG package), RNAmot, RNAbob or RNAmotifs (15,16) have been developed to carry out pattern matching but none of them offers the wealth of options and flexibility provided by PatSearch. PatSearch may represent a powerful tool to carry out genome-wide annotation of transcriptional and post-transcriptional regulatory elements already described in the literature.

In particular, searches for combinations of PWMs in the proper arrangement and orientation in the search of known promoter modules, such as the actin-like promoter, can be carried out on genome draft sequences thus providing a putative set of co-regulated genes with a satisfactory selectivity rate. Analogously, the annotation of known RNA motifs involved in the control of mRNA translation efficiency, stability and subcellular localization may significantly contribute to the transcriptome annotation. PatSearch has been used to provide UTR annotation for the mouse transcriptome (17).

The possibility of simultaneously searching for specific combinations of motifs greatly reduces the rate of false positives that could represent the great majority of single motif hits due to their high level of degeneracy. In all cases, depending on pattern selectivity, a given fraction of false positives is expected. Thus, the assessment of statistical significance of pattern matching analyses is mandatory.

A simple way to assess the statistical significance of PatSearch analyses consists of the reiteration of the same search on a shuffled dataset maintaining the same nucleotide, or possibly dinucleotide, composition of the sequences under analysis. The number of expected hits, determined in the analysis of the shuffled dataset, as compared to that of observed ones, makes it possible to assess the statistical significance of results through a simple chi-square test.

ACKNOWLEDGEMENTS

The PatSearch software is essentially based on the pattern matching program 'scan_for_matches' which was written by Ross Overbeek. The authors would like to thank Carmela Gissi, David Horner, Michele Iacono and Flavio Mignone for advice in the development of PatSearch software. This work has been supported by Telethon, AIRC and MIUR Cluster C03/2000 (CEGBA).

REFERENCES

1. Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
2. Pesole,G., Liuni,S., Grillo,G., Licciulli,F., Mignone,F., Gissi,C. and Saccone,C. (2002) UTRdb and UTRsite: specialized databases of sequences and functional elements of 5′ and 3′ untranslated regions of eukaryotic mRNAs. Update 2002. *Nucleic Acids Res.*, **30**, 335–340.
3. Fickett,J.W. and Wasserman,W.W. (2000) Discovery and modeling of transcriptional regulatory regions. *Curr. Opin. Biotechnol.*, **11**, 19–24.
4. Hentze,M.W. and Kuhn,L.C. (1996) Molecular control of vertebrate iron metabolism: mRNA-based regulatory circuits operated by iron, nitric oxide, and oxidative stress. *Proc. Natl Acad. Sci. USA*, **93**, 8175–8182.
5. Williams,A.S. and Marzluff,W.F. (1995) The sequence of the stem and flanking sequences at the 3′ end of histone mRNA are critical determinants for the binding of the stem-loop binding protein. *Nucleic Acids Res.*, **23**, 654–662.
6. Mignone,F., Gissi,C., Liuni,S. and Pesole,G. (2002) Untranslated regions of mRNAs. *Genome Biol.*, **3**, 0004.1–0004.10.
7. Werner,T. (2000) Computer-assisted analysis of transcription control regions. Matinspector and other programs. *Methods Mol. Biol.*, **132**, 337–349.
8. 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.
9. Yang,A., Kaghad,M., Caput,D. and McKeon,F. (2002) On the shoulders of giants: p63, p73 and the rise of p53. *Trends Genet.*, **18**, 90–95.
10. Levrero,M., De Laurenzi,V., Costanzo,A., Gong,J., Wang,J.Y. and Melino,G. (2000) The p53/p63/p73 family of transcription factors: overlapping and distinct functions. *J. Cell. Sci.*, **113** (Pt 10), 1661–1670.
11. D'Erchia,A.M., Tullo,A., Pesole,G., Saccone,C. and Sbisà,E. (2003) p53 gene family: structural functional and evolutionary features. *Curr. Genom.*, **4**, 13–26.
12. Bourdon,J.C., Deguin-Chambon,V., Lelong,J.C., Dessen,P., May,P., Debuire,B. and May,E. (1997) Further characterisation of the p53 responsive element—identification of new candidate genes for trans-activation by p53. *Oncogene*, **14**, 85–94.
13. Perier,R.C., Junier,T., Bonnard,C. and Bucher,P. (1999) The Eukaryotic Promoter Database (EPD): recent developments. *Nucleic Acids Res.*, **27**, 307–309.
14. Pesole,G., Liuni,S. and D'Souza,M. (2000) PatSearch: a pattern matcher software that finds functional elements in nucleotide and protein sequences and assesses their statistical significance. *Bioinformatics*, **16**, 439–450.
15. Macke,T.J., Ecker,D.J., Gutell,R.R., Gautheret,D., Case,D.A. and Sampath,R. (2001) RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res.*, **29**, 4724–4735.
16. Laferriere,A., Gautheret,D. and Cedergren,R. (1994) An RNA pattern matching program with enhanced performance and portability. *Comput. Appl. Biosci.*, **10**, 211–212.
17. Okazaki,Y., Furuno,M., Kasukawa,T., Adachi,J., Bono,H., Kondo,S., Nikaido,I., Osato,N., Saito,R., Suzuki,H. *et al.* (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature*, **420**, 563–573.