Comparative DNA Sequence Analysis for the Detection of Regulatory Elements

Ulrike Gausmann[1], Niels Jahn[1], Cornelia Platzer[2], Karol Szafranski[1], and Matthias Platzer[1]

ugau@imb-jena.de, cplatzer@mti-n.uni-jena.de, mplatzer@imb-jena.de

[1]Department of Genome Analysis, Institute of Molecular Biotechnology, Postfach 100 813, 07708 Jena;
[2]Institute of Anatomy II, Medical School, Friedrich Schiller University Jena, Teichgraben 7, 07740 Jena; Germany

SUMMARY

The most common method for the prediction of putative **regulatory regions in promoter sequences** is based on the comparison of the nucleotide sequence with known binding matrices for transcription factor binding sites (*MatInspector*). However, these predictions still appear to be insufficient in sensitivity and specificity. Several improvements are on the way namely data basis expansion for matrix computation (*TransFac*), building of higher order binding site models (e.g. *TransCompel*) and the consideration of comparative approaches (e.g. *rVISTA*).

Within the newly founded **Sonderforschungsbereich "Multifunctional Signaling Proteins"** we deal with groups of proteins which are physically interacting and may be potentially co-regulated (http://www.imb-jena.de/SFB604/) . Comparative analysis of the 5' upstream gene regions should reveal regulatory elements common for all members of one group. However, the interactions are not exactly known for all of these groups and post-transcriptional mechanisms may mask expression effects. To enhance the data basis for **comparative analysis** we include paralogous and orthologous gene sequences of interest.

The *in silico* analysis requires test- and training sets and the automated data transfer from data storage to data analysis programs. Hence, we started to build a **relational database** which will also provide the experimental researchers within the SFB with data they need for their experiments, i.e. sequences with corresponding annotations derived from public references and from own analyses.

Comparative analysis of non-coding regions is based on the fact that functional elements are more strictly conserved and can be identified by alignment procedures. Sensitive pairwise alignment algorithms are the first steps but will not give clear results in many cases. Therefore, methods based on multiple comparisons will enhance the significance of the identified regions. For some purposes specialized programs can be used, e.g. **'DNA block aligner'** (*DBA*) for the detection of conserved blocks in human and rodent orthologous non-coding sequences (1).

As an example for the applicability of this approach we present the analysis of the **Interleukin-10 (IL-10) promoter**. IL-10 is a cytokine with immunosuppressive and anti-inflammatory properties and its synthesis can be induced or enhanced by various stimuli, particularly *via* a cAMP-dependent signaling pathway. Two functionally active **CRE motifs**, CRE1 and CRE4, were identified in the promoter region (2). It was suspected that other, non-CRE sites play also a role in this pathway. Comparative analysis of three orthologous promoter sequences (*H. sapiens*, *M. musculus*, *M. monax*) revealed that two binding sites for the **CCAAT enhancer binding protein**, C/EBP3 and 5, and the previously identified CRE4 site, are located in conserved regions (*DBA*). Using EMSA and reporter gene assays the prominent functional role of these sites was experimentally confirmed. In contrast, the weakly conserved sites CRE1 and C/EBP1 were shown to be less important.

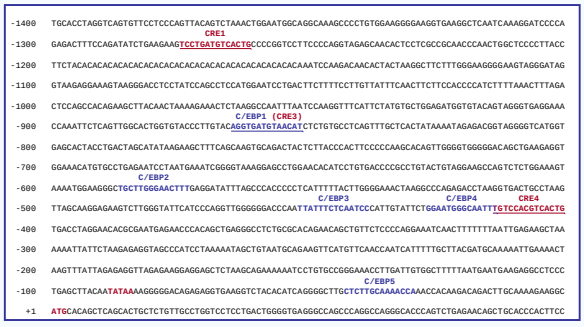
-1400 TGCACCTAGGTCAGTGTTCCTCCCAGTTACAGTCTAAACTGGAATGGCAGGCAAAGCCCCTGTGGAAGGGGAAGGTGAAGGCTCAATCAAAGGATCCCCA
CRE1
-1300 GAGACTTTCCAGATATCTGAAGAAGTCCTGATGTCACTGCCCCGGTCCTTCCCCAGGTAGAGCAACACTCCTCGCCGCAACCCAACTGGCTCCCCTTACC
-1200 TTCTACACACACACACACACACACACACACACACACACACACACAAATCCAAGACAACACTACTAAGGCTTCTTTGGGAAGGGGAAGTAGGGATAG
-1100 GTAAGAGGAAAGTAAGGGACCTCCTATCCAGCCTCCATGGAATCCTGACTTCTTTTCCTTGTTATTTCAACTTCTTCCACCCCATCTTTTAAACTTTAGA
-1000 CTCCAGCCACAGAAGCTTACAACTAAAAGAAACTCTAAGGCCAATTTAATCCAAGGTTTCATTCTATGTGCTGGAGATGGTGTACAGTAGGGTGAGGAAA
C/EBP1 (CRE3)
-900 CCAAATTCTCAGTTGGCACTGGTGTACCCTTGTACAGGTGATGTAACATCTCTGTGCCTCAGTTTGCTCACTATAAAATAGAGACGGTAGGGGTCATGGT
-800 GAGCACTACCTGACTAGCATATAAGAAGCTTTCAGCAAGTGCAGACTACTCTTACCCACTTCCCCCAAGCACAGTTGGGGTGGGGGACAGCTGAAGAGGT
-700 GGAAACATGTGCCTGAGAATCCTAATGAAATCGGGGTAAAGGAGCCTGGAACACATCCTGTGACCCCGCCTGTACTGTAGGAAGCCAGTCTCTGGAAAGT
C/EBP2
-600 AAAATGGAAGGGCTGCTTGGGAACTTTGAGGATATTTAGCCCACCCCCTCATTTTTACTTGGGGAAACTAAGGCCCAGAGACCTAAGGTGACTGCCTAAG
C/EBP3 C/EBP4 CRE4
-500 TTAGCAAGGAGAAGTCTTGGGTATTCATCCCAGGTTGGGGGGACCCAATTATTTCTCAATCCCATTGTATTCTGGAATGGGCAATTTGTCCACGTCACTG
-400 TGACCTAGGAACACGCGAATGAGAACCCACAGCTGAGGGCCTCTGCGCACAGAACAGCTGTTCTCCCCAGGAAATCAACTTTTTTTAATTGAGAAGCTAA
-300 AAAATTATTCTAAGAGAGGTAGCCCATCCTAAAAATAGCTGTAATGCAGAAGTTCATGTTCAACCAATCATTTTTGCTTACGATGCAAAAATTGAAAACT
-200 AAGTTTATTAGAGAGGTTAGAGAAGGAGGAGCTCTAAGCAGAAAAAATCCTGTGCCGGGAAACCTTGATTGTGGCTTTTAATGAATGAAGAGGCCTCCC
C/EBP5
-100 TGAGCTTACAATATAAAAGGGGGACAGAGAGGTGAAGGTCTACACATCAGGGGCTTGCTCTTGCAAAACCAAACCACAAGACAGACTTGCAAAAGAAGGC
+1 ATGCACAGCTCAGCACTGCTCTGTTGCCTGGTCCTCCTGACTGGGGTGAGGGCCAGCCCAGGCCAGGGCACCCAGTCTGAGAACAGCTGCACCCACTTCC

Fig. 1: Promoter region of the human IL-10 gene.
Assays were performed with a 1308bp fragment upstream of the start codon. Red CRE-sites are described in (2).

EXPERIMENTAL IL-10 PROMOTER ANALYSIS

In (2) it was shown that the regulation of the IL-10 promoter is dependent on other elements than CRE. AP-2 binding could be excluded through supershift experiments but for C/EBP (CCAAT/enhancer binding protein) the regulatory importance in cells of myelomonocytic origin could be demonstrated. C/EBP transcription factors belong to the same *TransFac* class as CREB proteins (C0008) . Both are involved in cAMP-regulated gene expression and contain a DNA-binding basic region followed by a leucine zipper. 5 candidate binding motifs could be detected in the 1308bp promoter fragment of IL-10 with *MatInspector* (**C/EBP1-5**). C/EBP1 is identical to the previously described putative CRE3 site which does not bind CRE. **Fig. 1** gives an overview for the human IL-10 promoter region.

EMSA and supershift assays (nuclear extracts of THP-1 cells, oligonucleotides containing consensus CRE & C/EBP motifs and antibodies against C/EBPα, β and δ from Santa Cruz) demonstrated that the C/EBP2 site does not compete with the C/EBP consensus (**Fig. 2A, lane 3**). Supershift complexes with C/EBP1, C/EBP3 and C/EBP5 contained C/EBPα and β but not C/EBPδ (shown for C/EBP5 in **Fig. 2B, lanes 9, 10, 11**). The C/EBP4 oligonucleotides were not shifted with anti-C/EBPα or β. Therefore, C/EBPα and β bind to three of the five candidate motifs (C/EBP1≡CRE3, C/EBP3 and C/EBP5). Supershift assays confirmed that the previously described CRE1 and CRE4 sites do not bind C/EBPα or β (**Fig. 2C**).

Reporter gene assays revealed that cAMP-stimulated activation depends on the integrity of the C/EBP binding sites (**Fig. 3**). In brief, these are the results of the other experiments performed (data will be published elsewhere):

- cAMP-stimulated IL-10 promoter activation in THP-1 cells involves C/EBPα
- C/EBPα influences basal IL-10 promoter activity in H-60 and THP-1 cells
- C/EBPα –dependent constitutive IL-10 promoter activity is mediated mainly through C/EBP5
- efficient cAMP-stimulated IL-10 promoter activation depends upon C/EBPα

COMPUTATIONAL IL-10 PROMOTER ANALYSIS

Human (**Fig. 1**) and murine genomic sequences were retrieved from the UCSC Genome Browser (http:// genome.ucsc.edu), datasets from April 2002 and February 2002, respectively. Genbank accession numbers for the other sequences are AF120030 (*Marmota monax* genomic IL-10 sequence), BG538741 (human EST), NM_000572 (human mRNA), Z30175 (human IL-10 promoter), NM_010548 (murine mRNA) and AF012909 (*M. monax* mRNA). Pairwise comparative analyses were performed with the program DNA Block Aligner (DBA). The DBA input files were masked with Repeatmasker (http://ftp.genome.washington.edu/ RM/ RepeatMasker.html). The prediction of transcription factor binding sites (TFBS) was done with the MatInspector Professional program (http:// www.genomatix.de). All promoter positions refer to the translation start site set to +1.

The trancription start site of the human IL-10 gene has not been demonstrated experimentally yet, but one human EST and 5' ends from mRNA of other species (*Sus scrofa* L20001, *Ovis aries* Z29362, *Trichosurus vulpecula* AF026277) suggest it at position -59. Three IL-10 promoter sequences were analysed: human (*hs*), murine (*mm*) and the *Marmota monax* (*ma*). All three contain a TATA-Box located at similar distances from the ATG: *hs* -89, *mm* -87, *ma* -86. Pairwise comparison with DBA reveals two blocks of similarity which appear in all three pairs: Block A at approximately -500 to -300 and Block B at -100 to -1 containing the TATA-box. These sequences were analysed by the MatInspector search program (core similarity 0.95, optimized matrix). Block A in human, mouse and *M. marmota* contains one C/EBP binding site (binding site matrix CEBP.01 in *hs* at -439 to -452, in *mm* at -450 to -463, in *ma* at -422 to -435) and one CREB protein binding site (binding site matrix ATF.01 in *hs* at -401 to -414, CREB.01 in *mm* at -415 to -422, CREB.04 in *ma* at -385 to -396). In Block B another C/EPB site was identifed: in *hs* at -30 to -43, and in *ma* at -39 to -52 (both match the C/EPB.01 matrix). Initially, in the murine sequence this matrix was not identified, but after lowering the search parameters, it could be detected at position -39 to -52. These results suggest that **Block A contains the elements C/EBP3 and CRE4** and **Block B harbors the C/EBP5** binding site (**Fig. 4**).

Considerating the correlation of biological significance and evolutionary conservation, it is not surprising to find that the regions containing the **functional relevant motifs C/EBP3, CRE4, and C/EBP5 are conserved** between human, mouse and *M. monax*. The *in silico* TFBS localization predictions are consistent with the experimental data, thus allowing to propose the exact binding sites for CRE4, C/EBP3 and C/EBP5 in the IL-10 promoter. The other functional motifs as CRE1, more distantly positioned, were not contained in all three DBA comparisons. Quality of local alignments (not shown) and TFBS matrix similarity suggest a closer relationship of the human and the *M. monax* IL-10 gene versus human and mouse or mouse and *M. monax*.

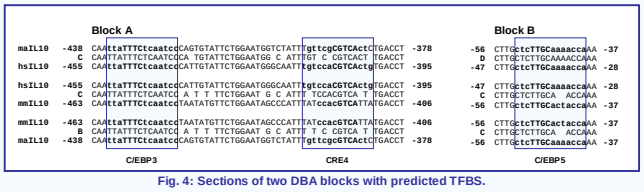

Fig. 4: Sections of two DBA blocks with predicted TFBS.
The IL-10 promoter sequences of *Homo sapiens* (hsIL10), *Mus musculus* (mmIL10) and *Marmota monax* (maIL10) were compared pairwise with the program DBA. 1.5kb upstream of the translation start were analyzed in each case. The categories B, C, and D mean whole DBA block sequence identity (B: 70-80%, C 80-90%, D: 90-100%). Bold letters: nucleotides belonging to the whole TFBS matrix region, bold capitalized letters: nucleotides of the TFBS core sequence (identified by MatInspector). Only the block sections containing the candidate TFBS are shown.

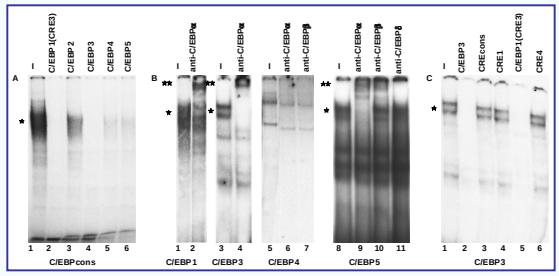

Fig. 2: Protein binding to the putative C/EBP motifs of the IL-10 promoter.
Radiolabeled oligonucleotides (as indicated at the bottom) were incubated with nuclear extracts from THP-1 cells and subjected to competition with a 100fold molar excess of unlabeled oligo (panels A and C) or supershifted with specific antisera (B).Competitor oligos and antibodies are shown at the top. Shifted DNA-C/EBP protein complex with one asterisk, supershifted complexes with two.

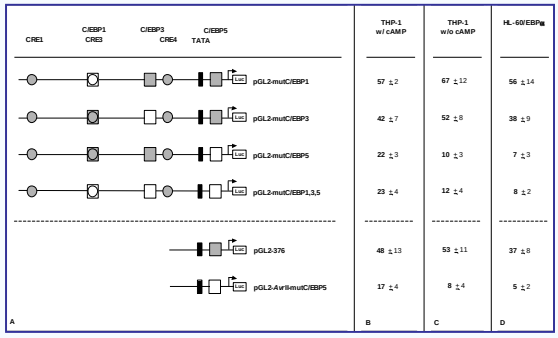

Fig. 3: Involvement of C/EBP binding sites in IL-10 promoter activation.
Assayed in THP-1 cells (monocytic) following cAMP stimulation, and in HL-60 cells (myeloid), which transiently express exogenous C/EBPC (HL-60/EBPα). Luciferase reporter gene plasmids containing mutated or deleted promoter fragments (A) were transfected. Promoter activity was expressed as percentage of cAMP-stimulated luciferase expression in THP-1 (B, C) or as a fraction of C/EBPα -induced basal luciferase expression obtained in HL-60/EBPα cells (D) with the complete unmutated reporter construct (representing 100%). Mean values of ten experiments each, circles: CRE, squares: C/EBP, grey: wildtype, blank: mutated.

(1) Jareborg N., E. Birney, R. Durbin (1999) Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. Genome Res. 9:815-824.
(2) Platzer C., E. Fritsch, T. Elsner, M.H. Lehmann, H.D. Volk, S. Prosch (1999) Cyclic adenosine monophosphate-responsive elements are involved in the transcriptional activation of the human IL-10 gene in monocytic cells. Eur. J. Immunol. 29:3098-3104.

seit 1558

JCB
JENA CENTRE FOR BIOINFORMATICS

SPONSORED BY THE Federal Ministry of Education and Research

NETWORK OF BIOINFORMATICS COMPETENCE CENTERS NBCC