Evaluating phylogenetic footprinting for human–rodent comparisons

Tilman Sauer1,* Ekaterina Shelest1 and Edgar Wingender1,2

1Department of Bioinformatics, UKG, Georg-August-University of Goettingen, Goldschmidtstrasse 1, 37077 Goettingen, Germany and 2BIOBASE GmbH, Halchtersche Strasse, 33, 38304 Wolfenbuettel, Germany

Received on September 27, 2005; revised on November 30, 2005; accepted on December 2, 2005
Advance Access publication December 6, 2005
Associate Editor: Nikolaus Rajewsky

ABSTRACT

Motivation: ‘Phylogenetic footprinting’ is a widely applied approach to identify regulatory regions and potential transcription factor binding sites (TFBSs) using alignments of non-coding orthologous regions from two or more organisms. A systematic evaluation of its validity and usability based on known TFBSs is needed to use phylogenetic footprinting most effectively in the identification of unknown TFBSs.

Results: In this paper we use 2678 human, mouse and rat TFBSs from the TRANSFAC® database for this evaluation. To ensure the retrieval of correct orthologous sequences, we combine gene annotation and sequence homology searches. Demanding a sequence identity of at least 65% is most effective in discriminating TFBSs from non-functional sequence parts, while different alignment algorithms only have a minor influence on TFBS identification by human–rodent comparisons. With this threshold ~72% of the known TFBSs are found conserved, a number which varies significantly between different transcription factors and also depends on the function of the regulated gene. TFBSs for certain transcription factors do not require strict sequence conservation but instead may show a high pattern conservation, limiting somewhat the validity of purely sequence-based phylogenetic footprinting.

Availability: Scripts are available from the authors upon request.

Contact: tsa@bioinf.med.uni-goettingen.de

Supplementary information: http://www.sybase.de/download

INTRODUCTION

Comparative genomics is a powerful approach for the prediction of gene structure and of cis-acting regulatory elements (Bulyk, 2003; Cooper and Sidow, 2003; Duret and Bucher, 1997; Frazer et al., 2003; Ureta-Vidal et al., 2003). The latter approach is also termed ‘phylogenetic footprinting’ and was first introduced by Tagle et al. (1988) who investigated primate \( \gamma \)- and \( \epsilon \)-globin genes. The basic assumption of phylogenetic footprinting is that regulatory elements in non-coding regions are under a higher selective pressure during evolution than non-functional regions. Deleterious mutations in regulatory regions are removed from a population owing to purifying selection and non-functional regions are therefore expected to accumulate more mutations over time than regulatory regions.

Aligning orthologous regions from two or more organisms hence should highlight the conserved parts which are expected to play a functional role. The term ‘functional’ is used in the following tantamount to an experimentally proven transcription factor–DNA interaction.

Owing to the availability of an increasing number of complete eukaryotic genomes, phylogenetic footprinting is an approach which is widely applied to identify regulatory regions and potential transcription factor binding sites (TFBSs) in different organisms. The inherent problem of comparative genomics is the question that which species should be compared with each other to most reliably identify functional regions in the genomic sequence.

Several studies dealt with this topic for Drosophila species. Bergman et al. (2002) indicated that large-scale comparative genomic sequence data from Drosophila pseudoobscura could greatly contribute to the functional annotation of the genome of Drosophila melanogaster. Species like Drosophila erecta are too closely related to differentiate whether conservation of non-coding sequence is due to functional constraint or shared ancestry, whereas the Anopheles gambiae genome may not substantially aid genome annotation as it lacks non-coding conservation in comparison with D.melanogaster. Embery et al. (2003) found that sequence conservation between D.melanogaster and D.pseudoobscura alone may not reliably discriminate regions containing TFBSs from those that do not. However, it has been shown that a combination of TFBS clustering and comparative sequence analysis is much more effective in identifying regulatory modules in D.melanogaster than each individual method on its own (Berman et al., 2004; Sinha et al., 2004).

A catalogue of regulatory motifs in Saccharomyces cerevisiae was discovered by a comparative analysis using high-quality draft sequences of three related yeast species (Kellis et al., 2003). These three species had sufficient sequence similarity to S.cerevisiae to reliably align orthologous regions, but sufficient sequence divergence to allow recognition of many functional elements by their greater degree of conservation. The same sequence data were also used to study the evolution of known TFBSs in S.cerevisiae, which show fewer substitutions than background DNA (Moses et al., 2003).

Human–mouse comparisons have been used extensively to identify potential regulatory regions which in many cases proved to be functional (Dermitsakis et al., 2002; Elnitski et al., 2003;
sequence retrieval of orthologous sequences

If a TFBS could be assigned to a gene, the Ensembl-database ensembl_mart_30 was queried for known orthologous genes (mouse orthologs for human TFBSs, human orthologs for rat or mouse TFBSs). The sequences obtained above were blasted against the second genome using WU-BLAST (W. Gish, http://blast.wustl.edu) with standard parameters. We preferred a whole genome approach to a WU-BLAST search just in regions near the known orthologous genes, to avoid an arbitrary choice of the size of these regions. Prior to blasting, potential repeats were species-specifically masked using RepeatMasker (A. F. A. Smit and P. Green, http://repeatmasker.org) in combination with MaskerAid (Bedell et al., 2000) and RepBaseUpdate of March 2004 (Jurka et al., 2000). The sequence around the WU-BLAST-hit having the highest possible score and being assigned to a known orthologous gene (as described earlier) was extracted (see also Supplementary Material). Start and end of the query sequence were extrapolated to retrieve a sequence of similar length, i.e. \( \sim 800 \) bp.

If the obtained orthologous sequence pairs (OSPs) were found to overlap, they were merged into one OSP.

All sequences containing the TFBSs were mapped to whole genome alignments (WGAs) made with BLASTZ, which were downloaded from the UCSC website (ftp://hgdownload.cse.ucsc.edu; hg17vsMm5, mm5vsHg17 and rn3vsHg17 in axtNet format).

alignment of orthologous sequences

The obtained OSPs were aligned with several global alignment algorithms (AVID, CLUSTAL W, CONREAL, DIALIGN, LAGAN and T-COFFEE) (Berezikov et al., 2004; Bray et al., 2003; Brudno et al., 2003; Morgenstern et al., 2000; Notredame et al., 2000; Thompson et al., 1994) and two local alignment algorithms (BLASTZ and LALIGN) (Huang and Miller, 1991; Schwartz et al., 2003). The alignment programs were run with default parameters. They are not specially designed for aligning non-coding sequences, except for CONREAL which uses a different approach by scanning both sequences with PSSMs for potential TFBSs to establish anchors between the orthologous sequences and to guide the sequence alignment.

conservation rate and background conservation

The conservation of the TFBSs is calculated using their sequence identity in the obtained pairwise alignments. The percentage of identity (PI) value of TFBS \( i \) (with \( i = 1, \ldots, N \)) is defined as the fraction of identical symbols in the columns \( j \) of the pairwise alignment over the length \( l \) of the TFBS

\[
PL = \frac{1}{N} \sum_{j=1}^{N} \delta_{ij}, \quad \text{where} \quad \delta_{ij} = \begin{cases} 1, & \text{if } X_j = Y_j \\ 0, & \text{if } X_j \neq Y_j \end{cases}
\]

\[
X_j, Y_j \in \{A,C,G,T,-\} \text{ are the human and rodent nucleotides or gaps (a nucleotide which is aligned to a gap is treated the same as a match) at position } j \text{ of TFBS } i \text{ in the pairwise alignment. A TFBS } i \text{ is considered as conserved, if its PI value is equal to or above a certain CT. The conservation rate } C_{\text{seq}} \text{ is defined as the fraction of all } N \text{ TFBSs that are considered as conserved:}
\]

\[
C_{\text{seq}} = \frac{1}{N} \sum_{i=1}^{N} b_i, \quad \text{where} \quad b_i = \begin{cases} 1, & \text{if } PL_i \geq CT \\ 0, & \text{if } PL_i < CT \end{cases}
\]

Analogously the conservation of the background is calculated. The background is defined as all columns of the obtained alignments which do not overlap with exons or known TFBSs. Every length \( l \), of all TFBSs (with \( i = 1, \ldots, N \)) is used as a sliding window to calculate PI values for all possible positions of the background. One obtains a distribution of PI values which is based on the same length distribution as the PI value distribution used for calculating \( C_{\text{seq}} \). The background conservation rate \( C_{\text{bg}} \) is defined as the fraction of these PI values that are equal to or above a certain CT.
Gene function dependency of TFBS conservation

The gene ontology (GO) annotation for each gene linked to a TFBS was replaced by the corresponding 'GO slim' terms (Harris et al., 2004). To construct 'GO slim' a set of high level terms was selected by the EBI to cover most aspects of each of the molecular function, biological process and cellular component ontologies. If several GO terms for one gene correspond to the same GO slim term, the GO slim term was counted only once. The conservation rate $C_{seq}$ for all TFBSs linked to a certain GO slim term was calculated individually.

Conservation rate on pattern level

The complete vertebrate PSSM library of TRANSFAC<sup>®</sup> was used to predict TFBSs in the OSPs using the MATCH program (Kel et al., 2003). All PSSMs related to TFs binding to an experimentally verified TFBS were used to scan both sequences. Matrix similarity score (mSS) thresholds from the predefined profile minFN were used to minimize false negative predictions.

If a PSSM hit in the orthologous sequence occurred in the alignment within ±2 bp of a PSSM hit for the same matrix in the original sequence, the ground conservation rate was used in the construction of the PSSM. In this case it is known whether the hit should occur. The conservation rate $C_{pat}$ is defined as the fraction of all predicted TFBSs which are pattern-conserved. The background conservation rate $C_{bg,pat}$ is defined as the fraction of all predicted TFBSs (for a particular TF) which are pattern-conserved.

Significance of conservation rates

If $K$ TFBSs out of a group of $N$ TFBSs are conserved ($C_{eq,pat} = K/N$), the probability to observe this conservation rate by chance (given a background conservative rate $C_{bg,pat}$) is calculated using a cumulative binomial distribution:

$$P_{eq,pat} = \sum_{j=0}^{K} \binom{N}{j} \left( C_{bg,pat} \right)^{j} \left( 1 - C_{bg,pat} \right)^{(N-j)}$$  (3)

RESULTS AND DISCUSSION

Orthologous sequence retrieval

Intuitively orthologous sequences from a second species would be retrieved according to the annotated TSS, e.g. Jareborg et al. (1999) or Liu et al. (2004) focussed in their studies on the sequences 1000 bp upstream of the TSS for both human and mouse. But it is generally not valid to retrieve an orthologous sequence from the second species just by the known distance from the TSS in the first genome, since the TSS in the second species may be different or annotated incorrectly in either genome. Moreover, when analyzing enhancer sequences that may be located at rather large distances from the TSS, the chances of insertions/deletions between them and the TSS increase with distance. Therefore, an anchor on the second genome is needed to retrieve the correct orthologous sequence. In our approach this anchor is provided by a WU-BLAST search.

Of the 3508 TFBSs 3383 could be mapped to their corresponding genome and linked to a gene (see Materials and Methods section and Supplementary Material), with 72 TFBSs no mapping result was obtained and 53 TFBSs could not be mapped on to a gene. Failures may be due to incomplete genome assemblies and annotation or to errors in the sequences obtained from EMBL. The vast majority of TFBSs are located in the region of −1000 to +250 bp around the TSS (see Supplementary Material). Of these 3383 TFBSs 2971 (88%) were found to be covered by the WGAs, whereas 412 (12%) occurred in gaps of the WGAs.

For 224 of the 3383 TFBSs there was no annotation of an orthologous gene, giving 3159 TFBSs where a WU-BLAST search was performed. For 481 of these 3159 TFBSs the WU-BLAST search failed to identify an anchor on the second genome (15%). Reasons for this could be as described above, very weak sequence homologies that were not detected by the WU-BLAST search or owing to repeat-masking. It is known that regulatory sequences occur in repetitive DNA (Jordan et al., 2003). Repeats were found more frequently in sequences where no corresponding orthologous sequence could be identified, indicating that the repeat-masking may have prevented the identification of an orthologous sequence: For 184 of the 481 sequences (38%) repeats were found with an average of 34% of masked base pairs. In the remaining 2678 cases (85%) an orthologous sequence was identified. Here the amount of repeat-masked sequence was significantly lower: For 572 (21%) of these cases repeats were found with an average of 18% of masked base pairs. This low repeat content is also reflected in the corresponding identified orthologous sequences, where 467 (17%) orthologous sequences contain repeats with an average of 19% of masked base pairs.

Merging overlapping OSPs gave 614 OSPs containing 1349 human and 812 mouse TFBSs and 161 OSPs containing 515 rat TFBSs. For all 775 clusters the TSS distances of the human sequences were determined and compared with the TSS distances of the rodent sequences (Fig. 1). In 54% of the cases the TSS distances differed by <100 bp, but in ~25% by >500 bp (in 18% by >1000 bp), demonstrating the need for an anchor on the second genome other than the annotated TSS: as the average length of the sequences which have been aligned is ~950 bp, a TSS distance difference of 1000 bp would render it impossible to align the known TFBSs to their presumed orthologous counterparts if the orthologous sequences would have been retrieved in relation to the TSS.

Sequence conservation of TFBS

The aim of this work was to evaluate and compare different alignment algorithms in their ability to identify experimentally known TFBSs. The obtained results should give hints on how to use phylogenetic footprinting to optimally identify unknown TFBSs by human–rodent comparisons. Two parameters are of importance for this evaluation: sensitivity and specificity. As not all TFBSs are expected to be conserved between human and rodents, the term sensitivity is misleading, because a sensitivity of 1 cannot be reached a priori. To avoid misconception the term 'conservation rate' is used instead (see Materials and Methods section).

It is difficult to determine specificity in phylogenetic footprinting directly, because erroneously identified conserved regions [false-positives (FPs)] cannot be distinguished from potentially functional conserved regions. To get an impression of how each algorithm and parameter combination can distinguish between real functional conservation and background conservation, the percentage of background sequences with a PI value equal to or above a chosen CT was determined (see Materials and Methods section). This percentage is an overestimation of the real FP rate, as there should be other functional TFBSs (true positives) in the vicinity of the known TFBSs. These true positives would be counted as FP, thus rendering the background conservation rate $C_{bg,seq}$ very conservative. The
functionality of the known TFBSs is reflected in a higher sequence conservation (average PI value of 75.6%) compared with the background sequences (52.0%) (Fig. 2).

Plotting the obtained conservation rate versus the background conservation rate as a receiver operating characteristic (ROC) curve shows no significant differences between the investigated alignment algorithms (Fig. 3). Global alignment programs perform slightly better than the local alignment programs and the BLASTZ WGAs. Local tools generally have lower sensitivity as the alignments produced do not cover the input sequences completely. This observation was also made by Pollard et al. (2004) who simulated alignments over a range of divergence times and compared them with alignments produced by pairwise alignment tools. The BLASTZ WGAs are a bit weaker in performance than the other alignment programs, but are better in coverage, as 2971 TFBSs could be mapped to the WGAs, whereas only 2678 TFBSs were covered by the OSPs obtained using WU-BLAST. CONREAL performs best in separating the known TFBSs from non-functional sequences, but it has to be emphasized that CONREAL utilizes known TFBSs in the form of PSSMs to create the alignment (using predicted TFBSs as anchor points) in contrast to the other alignment programs. The circularity in this approach may be a drawback as footprints which cannot yet be assigned to any known TF remain unaligned. This disadvantage does not hold true for the other alignment algorithms, whose runtimes are additionally at least 20 times shorter. Alignments obtained with AVID were selected for further investigations.

While human and rodent sequences are still similar enough to guide most alignment programs, the performance differences between the alignment algorithms may vary substantially for increasing evolutionary distances as was shown recently by Pollard et al. (2004). Very high evolutionary distances may even render it impossible to obtain an accurate alignment. Rosenberg (2005) found that if <50% of positions between two non-coding sequences are identical, CLUSTAL W creates pairwise alignments that do not differ from random data. Thus for sequence comparisons of species more diverged than human and rodent one particular alignment program may be more clearly preferable than others.

The CT which gives the optimal trade-off between conservation rate and the background conservation rate was determined using the ROC curves by calculating the distance of each point in the plot to the upper left corner of the graph (see Supplementary Material). At this point the conservation rate would equal 100%, while the background conservation rate would equal 0%. A CT of 65% gives the minimal distance and is optimal in separating the real TFBSs from the background sequences and is therefore used for further analyses. This CT is an average value determined using all TFBSs in the dataset, for individual TFs the optimal CT may differ slightly. It gives a high conservation rate of $C_{seq}^{bg} = 71.7\%$ and a background conservation rate of $C_{seq}^{bg} = 35.2\%$. With this threshold 28.3% of TFBSs
remain that are not considered as conserved. A fraction of these TFBSs may be species-specific to human or rodents and therefore not detectable by sequence comparisons. Also it may be that regulatory regions, owing to their inner variability, gain and lose TFBSs, change their relative positioning or the spacing between them. The latter has been shown for *Drosophila* by Ludwig et al. (2000).

A subset of 100 rodent TFBSs was found to be aligned with human TFBSs belonging to the same TF. Of these TFBSs 94 suffice the CT of 65%, proving that phylogenetic footprinting properly detects TFBSs existing in both species. These 100 rodent TFBSs were not included in the calculation of $C_{\text{seq}}$ to avoid counting the same information twice.

As for 481 TFBSs no orthologous sequence could be identified, the above mentioned $C_{\text{seq}}$ may be biased. If the orthologous sequence retrieval failed owing to low overall sequence homology and no other reasons, these TFBS would have a lower probability to be conserved and thus could decrease the percentage of conserved TFBSs. When assuming that all these 481 TFBSs are not conserved, a $C_{\text{seq}}$ of 60.4% would be obtained as an underestimate of the genome-wide conservation of TFBSs between human and rodents. Nevertheless, when using phylogenetic footprinting to detect unknown TFBSs, ~72% of real TFBSs would be detected because the orthologous sequence is needed as a prerequisite.

The fraction of conserved TFBSs observed in this study is in good agreement with observations made in earlier studies. Levy and Hannenhalli (2002) found 65% of experimentally defined TFBSs in regions conserved between human and mouse (having a minimal length of 50 bp and a minimal PI value of 70%). Lenhard et al. (2003) could detect ~68% of TFBSs in conserved regions defined by a sliding window of 50 bp and a CT of 70%. It has to be noted that the PI value of the individual TFBSs themselves may be <70%, as the average length of TFBSs is much smaller than the used window length of 50 bp. Liu et al. (2004) found ~60% of TFBSs and ~25% of background sequences to be conserved when using a window of 21 bp centered on the TFBSs and a CT of 70%. We did not focus on a certain window size to determine conserved regions but instead investigated the TFBSs themselves as did Dermitzakis and Clark (2002). They found ~62% of TFBSs to suffice a CT of 70%. Using the same CT we observe 66% of the TFBSs and 28% of the background sequences to be conserved.

**Table 1. TF dependency of the conservation rate**

<table>
<thead>
<tr>
<th>TF</th>
<th>$N^a$</th>
<th>$C_{\text{seq}}^b$ (%)</th>
<th>$p_{\text{seq}}^c$</th>
<th>$C_{\text{pat}}^d$ (%)</th>
<th>$p_{\text{pat}}^e$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MyoD</td>
<td>16</td>
<td>100.0</td>
<td>5.3E-06</td>
<td>100.0</td>
<td>21.21</td>
</tr>
<tr>
<td>MEF-2A</td>
<td>13</td>
<td>100.0</td>
<td>0.00012</td>
<td>92.3</td>
<td>19.91</td>
</tr>
<tr>
<td>SRF</td>
<td>33</td>
<td>96.9</td>
<td>6.4E-12</td>
<td>87.8</td>
<td>24.17</td>
</tr>
<tr>
<td>NF-AT1</td>
<td>31</td>
<td>96.7</td>
<td>4.9E-11</td>
<td>93.5</td>
<td>28.31</td>
</tr>
<tr>
<td>USF-1</td>
<td>22</td>
<td>86.3</td>
<td>0.0001</td>
<td>77.2</td>
<td>19.70</td>
</tr>
<tr>
<td>NF-κB</td>
<td>38</td>
<td>84.2</td>
<td>6.7E-08</td>
<td>71.0</td>
<td>21.98</td>
</tr>
<tr>
<td>CREB</td>
<td>49</td>
<td>83.6</td>
<td>3.8E-10</td>
<td>81.6</td>
<td>28.62</td>
</tr>
<tr>
<td>C-Ens</td>
<td>18</td>
<td>83.3</td>
<td>0.00368</td>
<td>66.6</td>
<td>26.59</td>
</tr>
<tr>
<td>ER-α</td>
<td>41</td>
<td>82.3</td>
<td>0.00876</td>
<td>52.9</td>
<td>22.08</td>
</tr>
<tr>
<td>MITF</td>
<td>24</td>
<td>79.1</td>
<td>0.00128</td>
<td>54.1</td>
<td>18.42</td>
</tr>
<tr>
<td>C/EBPβ</td>
<td>42</td>
<td>78.5</td>
<td>1.1E-06</td>
<td>90.4</td>
<td>33.40</td>
</tr>
<tr>
<td>Sp3</td>
<td>27</td>
<td>77.7</td>
<td>0.00073</td>
<td>92.5</td>
<td>30.26</td>
</tr>
<tr>
<td>RXR-α</td>
<td>54</td>
<td>75.9</td>
<td>1.2E-07</td>
<td>64.8</td>
<td>27.12</td>
</tr>
<tr>
<td>HNF-1α</td>
<td>20</td>
<td>75.0</td>
<td>0.03155</td>
<td>85.0</td>
<td>30.95</td>
</tr>
<tr>
<td>AP-1</td>
<td>77</td>
<td>74.0</td>
<td>4.2E-10</td>
<td>76.6</td>
<td>28.27</td>
</tr>
<tr>
<td>YY1</td>
<td>29</td>
<td>72.4</td>
<td>0.00469</td>
<td>68.9</td>
<td>28.87</td>
</tr>
<tr>
<td>POU2Fl1</td>
<td>32</td>
<td>65.6</td>
<td>0.04194</td>
<td>87.5</td>
<td>30.05</td>
</tr>
<tr>
<td>Sp1</td>
<td>229</td>
<td>63.7</td>
<td>1.5E-16</td>
<td>78.1</td>
<td>31.94</td>
</tr>
<tr>
<td>C/EBPα</td>
<td>87</td>
<td>63.2</td>
<td>9.3E-06</td>
<td>89.6</td>
<td>39.27</td>
</tr>
<tr>
<td>c-Myb</td>
<td>16</td>
<td>62.8</td>
<td>25.0</td>
<td>28.19</td>
<td>1</td>
</tr>
<tr>
<td>AP-2αA</td>
<td>29</td>
<td>48.2</td>
<td>1</td>
<td>89.6</td>
<td>33.59</td>
</tr>
<tr>
<td>GATA-1</td>
<td>50</td>
<td>44.0</td>
<td>1</td>
<td>50.0</td>
<td>21.71</td>
</tr>
</tbody>
</table>

*aNumber of TFBSs per TF.*

*bThe sequence background conservation $C_{\text{bg}}^d$ is 35.22%.

*cThe $p$-value to obtain the observed (or a bigger) difference between the conservation rate and the background conservation rate by chance have been calculated using Equation (3).
These highly conserved sequences are probably existing in all vertebrate species and are essential to vertebrate development. They are in most cases at least several kb apart from known genes and may act as distal enhancers or play a role in structuring the genomic architecture. In contrast to that, Iwama and Gojobori (2004) focussed in their study on the conservation of sequences directly upstream of 3055 orthologous human–mouse gene pairs. They found that TF genes and developmental process-related genes show the highest degree of upstream sequence conservation, while genes involved in metabolism and cell cycle show less upstream sequence conservation.

While Iwama and Gojobori (2004) investigated upstream sequence conservation in general, in this study we explicitly checked if there is a correlation between gene function and the conservation of individual TFBSs. GO slim annotation was retrieved for all genes linked to a TFBS from Ensembl (see Materials and Methods section). The results are shown in Table 2 (see also Supplementary Material). TFBSs regulating genes having transcription regulator activity \( (C_{\text{seq}} = 80.5\%) \) or which are involved in cell death (77.9%), the regulation of biological processes (77.3%), nucleic acid binding (77.2%), signal transduction (75.8%) and development (75.1%) have high conservation rates. TFBSs regulating genes involved in transport (62.1%) and having certain catalytic activities, e.g. like kinase (66.9%), transferase (61.9%) and oxidoreductase activity (59.9%), exhibit low conservation rates. These results are in very good agreement with the studies mentioned earlier and indicate that TFBSs regulating genes involved in processes essential to vertebrate evolution are under a high selective pressure and therefore exhibit a higher sequence conservation.

Because the investigated TFBSs are linked simultaneously to TFs and GO slim terms, we also checked if the results of the gene function dependency could be explained by the results of the TF dependency analysis, but we could not find that certain TFs act solely on genes in certain functional categories (data not shown).

**Pattern conservation of TFBS**

We have shown that the sequence conservation of TFBSs differs among different TFs. Since the DNA-binding specificity of some TFs is rather relaxed, the sequence of a TFBS can be poorly conserved, but the TF may still find a recognizable binding site in the orthologous sequence which may match with a ‘pattern’, e.g. a probabilistic representation of a binding sequence as it is presented by a PSSM. On the contrary it may be that even if the sequence of a TFBS is highly conserved, a single mutation, insertion or deletion may have rendered the orthologous TFBS non-functional. A corresponding PSSM could be retrieved for 1990 of 2678 TFBSs from the TRANSFAC\textsuperscript{B} database and pattern conservation was determined as described in the Materials and Methods section for AVID alignments.

TFBSs (93) of the subset of 100 rodent TFBSs which were found to be aligned with human TFBSs belonging to the same TF are pattern conserved and not included in the calculation of the pattern conservation rate, giving a \( C_{\text{pat}} \) of 69.5%.

On sequence level 72.3% of the 1890 TFBSs are found conserved (with a CT of 65%). Of the conserved sequences 58.4% are conserved by both sequence and pattern, which means that 13.9% are conserved by sequence only and 11.1% by pattern only. Also 16.6% are conserved neither on sequence nor on pattern level.

**Table 2. Gene function dependency of the conservation rate**

<table>
<thead>
<tr>
<th>GO slim term</th>
<th>( N^a )</th>
<th>( C_{\text{seq}} (%) )</th>
<th>( p_{\text{seq}}^b )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transcription regulator activity</td>
<td>262</td>
<td>80.5</td>
<td>3.1E-49</td>
</tr>
<tr>
<td>Hydrolase activity</td>
<td>199</td>
<td>79.4</td>
<td>2.1E-35</td>
</tr>
<tr>
<td>Cell death</td>
<td>195</td>
<td>77.9</td>
<td>2.4E-32</td>
</tr>
<tr>
<td>Regulation of biological process</td>
<td>745</td>
<td>77.3</td>
<td>5E-120</td>
</tr>
<tr>
<td>Nucleic acid binding</td>
<td>356</td>
<td>77.2</td>
<td>3.8E-57</td>
</tr>
<tr>
<td>Catabolism</td>
<td>187</td>
<td>77.0</td>
<td>1.4E-29</td>
</tr>
<tr>
<td>Signal transducer activity</td>
<td>714</td>
<td>75.8</td>
<td>2E-106</td>
</tr>
<tr>
<td>Development</td>
<td>526</td>
<td>75.1</td>
<td>9.5E-76</td>
</tr>
<tr>
<td>Protein binding</td>
<td>779</td>
<td>74.8</td>
<td>1E-110</td>
</tr>
<tr>
<td>Response to stimulus</td>
<td>703</td>
<td>74.8</td>
<td>8E-100</td>
</tr>
<tr>
<td>Cell communication</td>
<td>602</td>
<td>74.1</td>
<td>3.1E-82</td>
</tr>
<tr>
<td>Enzyme regulator activity</td>
<td>112</td>
<td>72.3</td>
<td>9.1E-14</td>
</tr>
<tr>
<td>Metabolism</td>
<td>1197</td>
<td>71.1</td>
<td>9E-139</td>
</tr>
<tr>
<td>Catalytic activity</td>
<td>646</td>
<td>67.5</td>
<td>1.5E-60</td>
</tr>
<tr>
<td>Kinase activity</td>
<td>136</td>
<td>66.9</td>
<td>3.6E-12</td>
</tr>
<tr>
<td>Cell motility</td>
<td>64</td>
<td>65.6</td>
<td>4.5E-05</td>
</tr>
<tr>
<td>Electron transport</td>
<td>95</td>
<td>65.3</td>
<td>1.6E-07</td>
</tr>
<tr>
<td>Transport</td>
<td>388</td>
<td>62.1</td>
<td>2.8E-25</td>
</tr>
<tr>
<td>Transferrase activity</td>
<td>223</td>
<td>61.9</td>
<td>3.2E-14</td>
</tr>
<tr>
<td>Oxidoreductase activity</td>
<td>187</td>
<td>59.9</td>
<td>3.6E-10</td>
</tr>
<tr>
<td>Transporter activity</td>
<td>350</td>
<td>59.7</td>
<td>6.2E-19</td>
</tr>
<tr>
<td>Electron transporter activity</td>
<td>62</td>
<td>56.5</td>
<td>0.03223</td>
</tr>
</tbody>
</table>

\( ^a \)Number of TFBSs linked to the GO slim term.

\( ^b \)See Table 1.

**TF dependency of TFBS pattern conservation**

Generally sequence and pattern conservation of TFBSs are positively correlated, but for several TFs they differ (Table 1 and Supplementary Material). This may basically have two reasons: (1) As the conservation on sequence level does not require 100% sequence identity, it may happen for some TFBSs that just a few essential nucleotides have mutated or the TFBS has lost its function owing to a single insertion or deletion and a matrix search would give no hit above the mSS threshold in the orthologous sequence. A similar case would be a sequence-conserved TFBS that is highly degenerate and therefore not detectable with a corresponding PSSM in the original sequence. This would appear as a sequence-, but not pattern-conserved TFBS. (2) Alternatively it may be that the sequence identity is generally low but the TF binds nevertheless to both the original and the orthologous sequence as it may have a low binding specificity. Such a case would appear as a pattern-, but not sequence-conserved site.

Examples for the first case are the TFBSs of ER-\( \alpha \) which are well conserved on the sequence level (82%), but the conservation rate on pattern level is just \( \sim \)53%. For example in one ER-\( \alpha \) site in the rat Calbindin D9K gene one essential nucleotide in the core of the binding site has mutated in the orthologous sequence, which decreases the mSS drastically, although the surrounding sequence is conserved. It is possible that the used PSSM is too stringent owing to a single insertion or deletion and a matrix search would give no hit above the mSS threshold in the orthologous sequence.
Examples for the second case are TFBSs for C/EBPα, AP-2αA and Sp1. C/EBPα shows low conservation on sequence level (63%), but high conservation on pattern level (90%), which may, at least in part, be due to the high degeneracy of C/EBPα binding sites. The corresponding PSSMs will match frequently, which is reflected in the high background pattern conservation rate \( C_{\text{pat}} \) of 39%. For all investigated TFs we found a weak positive correlation between the pattern conservation of the known TFBSs and the background (correlation coefficient \( R = 0.53 \)). Also AP-2αA and to a smaller extent Sp1 show high pattern and low sequence conservation, while having a lower background conservation rate than C/EBPα. The conservation rate of TFBSs for AP-2αA does not differ significantly from the background conservation on sequence level (Bonferroni corrected \( p \)-value of 1), whereas this difference is significant on the pattern level (Bonferroni corrected \( p \)-value of \( 5 \times 10^{-5} \)). The concept of pattern conservation therefore is useful for certain TFs where purely sequence based phylogenetic footprinting may fail.

**CONCLUSIONS**

In this study we evaluated the usefulness of phylogenetic footprinting when using human–mouse/rat comparisons. Starting from known TFBSs we tuned the approach to obtain the best discrimination between functional and non-functional elements. Retrieving the correct orthologous sequence is the first crucial point in this approach and in terms of performance it is recommendable to create an anchor on the orthologous genome by combining a WU-BLAST search and gene annotation, because retrieving orthologous sequences just by using an identical relative distance to the TSS in both genomes may not return the correct orthologous region, as in 25% of the investigated 775 OSPs the TSS distance difference exceeded 500 bp. However orthologous sequences could not be identified in all cases by this approach, which can be avoided by using WGAs which proved to be better in coverage by providing alignments for 2971 TFBSs compared with 2678 TFBSs in the WU-BLAST approach.

The optimal cut-off to determine conservation is the second crucial point and was evaluated in two points: first, the fraction of TFBSs which complied with the CT and second, what FP rates were obtained with this CT. Two measures for the FP rate have been introduced showing that a CT of 65% optimally discriminates real TFBSs from non-functional regions. About 72% of the known TFBSs suffice the CT of 65%, suggesting that ~72% of unknown TFBSs may be identified by human–rodent comparisons, if an orthologous sequence could be retrieved.

Alignment algorithms differ only marginally in their performance when comparing human and rodent sequences with global alignment programs performing slightly better than local alignment programs or the WGAs. For the comparison of more diverged species it is expected that the performance difference will increase, as indicated by the results of Pollard et al. (2004) and Rosenberg (2005).

The conservation of TFBSs shows some dependence on the corresponding TF, e.g. binding sites for MyoD, MEF-2, SRF or NF-AT1 are conserved very frequently, whereas binding sites for Sp1, C/EBPα, AP-2αA or GATA-1 show a lower conservation rate. Depending on the functionality of the regulated genes, the conservation rate for TFBSs also differs: genes involved in transcriptional regulation, development or signal transduction show a higher TFBS conservation rate than genes involved in catalytic activities and transport. Additionally, for some TFs sequence conservation may not be informative about the conservation of the TFBS in the orthologous sequence, as the constraints for binding may be relatively relaxed. A PSSM for a certain TF may still score above the mSS threshold although several positions have mutated in the orthologous sequence. This leads to significant differences in the conservation rates on sequence and pattern level for several factors, e.g. for AP-2αA binding sites one obtains a conservation rate of 90% on pattern level compared with 48% on sequence level. These findings have to be taken into account when assessing the reliability of phylogenetic footprints for certain TFs.

**ACKNOWLEDGEMENTS**

This work was supported by the Bioinformatics Competence Center ‘Intergenomics’ using a grant of the German Ministry of Education and Research (grant no. 031U210A). The authors are indebted to Martin Haubrock, Alexander Kel, Peter Meinicke and Anne-Kathrin Schultz for stimulating discussions and three anonymous reviewers who helped to improve this manuscript. Funding to pay the Open Access publication charges for this article was provided by the Georg-August-University of Goettingen.

Conflict of Interest: none declared.

**REFERENCES**


