Footer: A quantitative comparative genomics method for efficient recognition of *cis*-regulatory elements

David L. Corcoran,[1,2] Eleanor Feingold,[1,2] Jessica Dominick,[2] Marietta Wright,[4] Jo Harnaha,[4] Massimo Trucco,[4] Nick Giannoukakis,[4] and Panayiotis V. Benos[2,3,5]

[1]Department of Biostatistics, Graduate School of Public Health, [2]Department of Human Genetics, GSPH, and [3]Department of Computational Biology and University of Pittsburgh Cancer Institute, School of Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania 15621, USA; [4]Children's Hospital of Pittsburgh, Pittsburgh 15213, USA

The search for mammalian DNA regulatory regions poses a challenging problem in computational biology. The short length of the DNA patterns compared with the size of the promoter regions and the degeneracy of the patterns makes their identification difficult. One way to overcome this problem is to use evolutionary information to reduce the number of false-positive predictions. We developed a novel method for pattern identification that compares a pair of putative binding sites in two species (e.g., human and mouse) and assigns two probability scores based on the relative position of the sites in the promoter and their agreement with a known model of binding preferences. We tested the algorithm's ability to predict known binding sites on various promoters. Overall, it exhibited 83% *sensitivity* and the *specificity* was 72%, which is a clear improvement over existing methods. Our algorithm also successfully predicted two novel NF-κB binding sites in the promoter region of the mouse autotaxin gene (*ATX, ENPP2*), which we were able to verify by using chromatin immunoprecipitation assay coupled with quantitative real-time PCR.

[Supplemental material is available online at www.genome.org.]

The advent of the genomic era initially raised hopes that the cellular mechanisms would be easily deciphered once the set of proteins of an organism had been identified. We now know that more complex phenotypes do not necessarily result from a larger number of genes, but could be the result of fine-tuning of their regulation. Thus, the identification of the DNA regulatory elements that control gene expression is the next necessary step in discovering regulatory pathways and understanding the basis of many diseases.

Each transcription factor (TF) protein usually recognizes a small set of TF binding sites (TFBSs) with high affinity. These sites can be identified by biochemical studies or by in vitro DNA selection binding experiments, such as SELEX (Choo and Klug 1994). The recent advances in chromatin immunoprecipitation (ChIP) (Orlando 2000) have enhanced our ability to identify multiple in vivo targets of a particular TF under certain cellular conditions. In any case, the information on the binding preferences of a TF needs to be organized in a form that will allow one to search the genome for new binding sites in the promoters of other genes. A lot of progress has been made during the last two decades on the development of such organization/representation methods (for a review, see Stormo 2000). The most widely used method is based on Position-Specific Scoring Matrices (PSSM), which represent the binding preferences of a TF to the DNA as a $4 \times L$ weight matrix (L is the length of the pattern). Typically, the weights in the matrix constitute some form of log-probability of the binding frequencies, and in some cases, it has been shown that they correspond to the binding energies (Benos et al. 2001, 2002a,b). In other cases, position dependencies on the DNA targets make simple PSSM models less accurate (Barash et al. 2003; Zhou and Liu 2004). However, in the case of complex modeling, the much larger sampling space combined with the lack of data might cause lower performance due to data overfitting.

If a PSSM model exists for a TF, it can be used to scan the promoters of other genes or the genome for high-scoring sites/patterns (Peters et al. 2002). However, the pattern matching on a genomic scale is far from accurate, especially in complex eukaryotic genomes. Additional properties that affect binding, such as DNA structure and the state of the chromatin cannot be modeled by these methods. Furthermore, the TFBS "signal" sometimes stands only slightly above the "background noise" of the genome. Protein cooperative effects might allow or require a functional binding site to be of suboptimal specificity. Thus, even with a perfect representation of the binding site preferences of a TF, the prediction of the real sites in mammalian promoters with simple pattern matching algorithms is still inaccurate.

One way to increase the signal is to use evolutionary information in identifying the conserved sites in multiple species. This is known as *phylogenetic footprinting*, a term coined by Tagle et al. (1988). A number of algorithms have been developed that use prior information (such as PSSM models) and comparative genomics to address the issue of accurate detection of TFBSs, including rVista (Loots et al. 2002) and ConSite (Lenhard et al. 2003). Given an alignment between two homologous promoter sequences, rVista (Loots et al. 2002) scans one of the promoters by using TRANSFAC (Wingender 2004) PSSM models for putative

[5]Corresponding author.
E-mail benos@pitt.edu; fax (412) 624-3020.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.2952005.

15:840–847 ©2005 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/05; www.genome.org

TF binding sites. The binding sites are then evaluated based on the degree of conservation of the interval in which they are located. Percent conservation is calculated in a 21-bp dynamically sliding window. Sites that are conserved on both promoters are reported. More recently, Lenhard et al. (2003) developed a flexible suite of methods for the identification and visualization of the conserved regulatory motifs in homologous sequences. Their method scans both promoter regions for putative binding sites and reports those pairs that are situated in equivalent positions in the conserved regions. Conserved regions are calculated over a sliding window of 50 bases (default value). In addition, other automated or semi-automated methods and tools have been recently developed around similar ideas of detection of *cis*-regulatory motifs in conserved DNA regions (Jegga et al. 2002; Schwartz et al. 2003).

Our method, Footer, differs from existing pattern matching methods in that it combines two types of information into a single scoring scheme. To our knowledge, it is the first method that combines multiple quantitative criteria in deciding about the micro-homology of the TFBSs. Overall, it surpasses existing methods in performance and utilizes efficiently information from the evolution of the *cis*-regulatory regions.

Methods

Phylogenetic conservation in the DNA regulatory regions

Not much is known about the evolution of DNA regulatory regions. This is due to the complexity of the constraints applied on them and the limited amount of available biological data. Constraints can be associated with the distance of the TFBS from the transcription start site (TSS) or from other sites (distance constraint) or with deviations from the TFs "preferred site" pattern (model score constraint). The distance constraint can be attributed to the need of the TFs to "communicate" with other proteins or protein complexes via protein–protein interactions. This is probably the most important of the constraints, and other phylogenetic footprinting algorithms have heavily depended on that for reducing the number of false-positive predictions (Jegga et al. 2002; Loots et al. 2002; Lenhard et al. 2003; Schwartz et al. 2003). The PSSM model score constraint reflects to the affinity of a TF to its DNA targets. Although it has not been proved in general, certain examples exist that show that the PSSM model scores could be considered as an approximation of the TF-DNA binding energies (Benos et al. 2001, 2002a,b). On theoretical grounds, if one assumes that the TF interacts with the DNA in equilibrium, then the protein–DNA specificity will follow a derivative of the Boltzmann distribution:

$$P(D|P) = \frac{P_{ref}(D) \cdot e^{-H(D,P)/RT}}{Z} \quad (1)$$

where D and P are the specific DNA target and the TF protein, respectively; $H(D,P)$ is the binding energy of the interaction; $P_{ref}(D)$ is the frequency of target D among the accessible genomic sites; and Z is the *partition function*, summed over all DNA targets (for a given protein):

$$Z = \sum_d P_{ref}(d) \cdot e^{-E(d,P)/RT}$$

PSSM scores are calculated as the negative logarithms of the frequencies of the base pair targets (typically normalized by the background frequencies). Thus, to the extent that these frequencies constitute an approximation of the base probabilities, the PSSM scores can be viewed as an approximation to the difference in the average specific binding energy with respect to the background. The protein–DNA interaction energy is one of the factors that determine the biological role of a binding site, although there is not a linear correlation between binding energy and transcription efficiency. Likewise, there is no linear correlation between the distance of a site from the TSS and transcriptional activation or repression. However, a model that utilizes both information has the potential on performing better in predicting the true binding sites.

Overview of the Footer algorithm

Two promoters are scanned for high-scoring patterns with a user-selected set of species-specific or mammalian-specific models. We use the program Consensus (Hertz and Stormo 1999) to calculate the PSSM models from sets of binding sites derived from the TRANSFAC database (Wingender 2004). For a given TF, a user-specified number of the top-scoring patterns is selected in each of the two promoters ("seed" patterns) and compared pairwise. The comparison is based on (1) the relative location of the patterns in the promoter region and (2) their agreement to the corresponding PSSM models. A probabilistic scoring scheme has been adopted for each of these criteria, under the *null hypothesis* that the two patterns are unrelated; thus the observed position and score values are due to chance alone. The two probabilities are combined in a single metric, and the pairs that score below a user-specified average *P*-value threshold are reported as the most likely true TF targets. For the matching, we apply a greedy algorithm according to which in each round the best matching pair is reported and the corresponding patterns are excluded from subsequent rounds. The algorithm stops when all seed patterns have paired or when the *P*-value threshold has been exceeded. Below, we describe some aspects of the method in more detail.

PSSM models

We constructed species-specific (i.e., human and mouse/rat) matrices from known target sequences recorded in the TRANSFAC database (Wingender 2004). The sites are selected so that both the protein and the DNA targets belong to the same species. The model is constructed from the base frequencies in each column with the addition of pseudocounts equal to 10% of the number of sequences and distributed according to the background model (i.e., equiprobable base distribution for mammals). If the number of available target sequences in each of the human and mouse/rat is greater than six, then species-specific models are created; otherwise sequences are joined into a mammalian-specific model. Traditionally, the pattern searching algorithms were using "all-species" matrices for the searches, although more recently, class-specific (i.e., mammalian-specific for human–mouse comparisons) matrices have been used (Lenhard et al. 2003). In our recent study, we found that although in general the evolution of the binding profiles follows the evolution of the protein families, there are notable exceptions to this rule (Mahony et al. 2005). An example of the effect of TFBS heterogeneity in different species is presented in Figure 1. The LOGO of the binding preferences of the AP-1 TF from sites from multiple species is plotted (TRANSFAC score matrix M00199) (Fig. 1A). Currently, this is the most general AP-1 weight matrix in the TRANSFAC database and is derived from sites from a variety of species, including human,

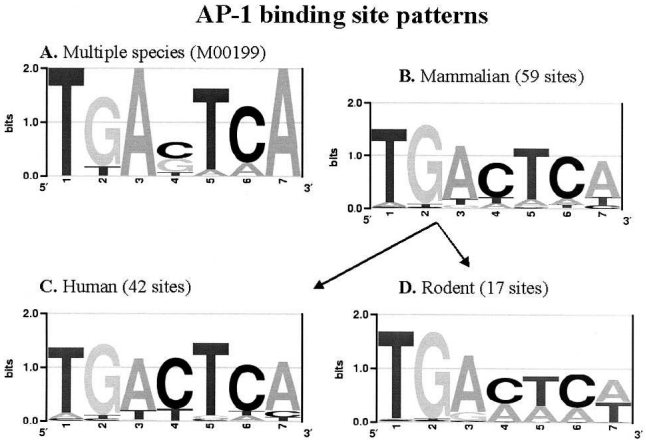

Figure 1. AP-1 binding site preferences in different species clusters. The LOGOs for the AP-1 binding preferences from (*A*) TRANSFAC weight matrix M00199 (includes sites from various organisms like human, mouse, rat, chicken and frog) and (*B*) human, mouse, and rat. Most of current algorithms are using one of the two variants of weight matrices to scan the promoter regions. (*C*, *D*) Further partition of the mammalian sequences into human and rodent, revealing differences in the suboptimal patterns. LOGOs are created by using the program enoLOGOS (Workman et al. 2005) available on the Web http://biodev.hgen.pitt.edu/cgi-bin/enologos/enologos.cgi.

mouse, rat, chicken, and frog. Figure 1, C and D, shows the collections of human and rodent sites, respectively, that Footer uses. In Figure 1B the human and rodent sets have been joined to create a mammalian-specific set. These LOGOs show clearly that although the optimal pattern is conserved in all data sets, species-specific differences emerge in suboptimal patterns. This might not affect the pattern matching algorithms that use a PSSM score cutoff to predict sites, but it can be important for the Footer algorithm that uses the PSSM score in a quantitative way (see below). For example, assuming that the PSSM models are indicative to the binding affinity of the TF to the DNA through equation 1, the pair of AP-1 sites TGA**T**TCA (human) and TGA**A**TCA (mouse) will be closer in affinity to the corresponding species-specific optimal sites than will the TGA**A**TCA (human) and TGAT**T**CA (mouse) pair. Note that both sites may score above a PSSM score threshold, regardless of the PSSM model and the organism they are found in. In addition, algorithms that use the multiple-species or the mammalian model will consider the two pairs equivalent. In fact, according to the multiple-species model of Figure 1A, site TGAATCA will also be even less favorable than site TGAGTCA, regardless of the organism it was found in. The quantitative use of species-specific PSSM models, however, allows Footer to distinguish between the above two pairs in comparison. Currently, we have constructed 19 species-specific and 108 mammalian-specific matrices, but we expect that more species-specific matrices will be used in the future, once more data becomes available.

Site comparison

Each pair of high-scoring patterns (e.g., human and mouse) is evaluated according to the *distance* and *PSSM score* criteria.

Distance probability

Assume that a pair of patterns is found to be d bases apart in the two promoters of length N^{tot}. For a pattern of length L we define the effective promoter length, $N = N^{tot} - L + 1$. Assuming a uniform distribution of randomly occurring patterns, the "distance probability," PF_D, is

$$PF_D = P(D_{XY} \le d) = \frac{1}{N} + \sum_{k=1}^{d} \frac{2 \cdot (N-k)}{N^2} \quad (2)$$

This is derived from the fact that the probability of observing two patterns at a distance of zero is $1/N$ and the probability of observing any other distance $d > 0$ is $2(N - d)/N^2$. Equation 2 can be further simplified to $(2 \times d + 1)/N - d \times (d + 1)/N^2$, which we use for the PF_D calculations. The distance between two putative sites is calculated in relation to their closest 3′ conserved region boundary (instead of the TSS) in order to allow for corrections of local insertions/deletions that frequently occur in the promoter regions. This applies to patterns that are located in both conserved and nonconserved regions, as they are determined by the output of the program DNA Block Aligner (DBA; part of the Wise2 software package) (Jareborg et al. 1999).

PSSM similarity score probability

Assume that we have the PSSM models M_1 and M_2 for the two species, and that S and T are random variables following the models' score distributions. Ideally, we would like to calculate the probability that we will observe scores s and t (or better) by chance given M_1 and M_2. We formulate this in the following way, which allows for more flexibility in the choice of the scores:

$$PF_S = P((S + T) \le (s + t) | M_1, M_2) \quad (3)$$

To reduce the computational complexity, we approximate the tail probability given in equation 3 by using the Gaussian distribution. In this way, only the mean and variance need to be calculated. On a theoretical basis, the Gaussian approximation (Hertz and Stormo 1999) is justifiable under the central limit theorem for reasonably long TFs and for PSSMs that are not extremely skewed. We checked the approximation for three TFs (including one of the most skewed) using Q–Q plots, and found it to be quite good except at the extreme tails. This is more than adequate for distinguishing between "moderate" and "small" *P*-values, the primary function of the PF_S score. We calculate the mean and standard deviation for the Gaussian approximation by Monte Carlo sampling of 1 million sequences from each pair of models.

Composite score and weighted average P*–value*

The composite score of Footer is the weighted negative sum of the logarithms of the individual tail probabilities from equations 2 and 3:

$$PF = -w_D \cdot \ln(PF_D) - w_s \cdot \ln(PF_S) \quad (4)$$

The weights w_D and w_S are positive numbers that sum to one. Summation of the logarithms is valid, since the two tail probabilities are based on the null hypothesis that the human and the mouse patterns are not true binding sites, and hence the individual tail probabilities are independent. According to equation 4, higher *PF* values correspond to higher probability that the human and mouse patterns are true sites. There is no objective way to assign values to these weights that will be valid for all TFs. In fact, one would expect that for each TF, the distance and the score probability scores have different importance. Besides, the quality of the PSSM model will be a factor that will affect the w_s parameter. In the current study, we used weights of 0.85 and 0.15

for the w_D and w_S, respectively. *PF* is the weighted sum of negative logarithm *P*-values (equation 4). Thus, exponentiation of $-PF$ will give the weighted average *P*-value (WAP) of Footer score. In the present study, we used a WAP threshold of 0.05%. All Footer parameters we used in this study, including the weights and the WAP threshold, were determined empirically, based on an initial analysis we performed on the 18 TFBSs in the promoters of the genes *PEPCK, G6Pase,* and *Leptin*. Exclusion of these 18 sites from the evaluation set does not alter the results (see below).

ChIP assays

ChIP (Orlando et al. 1997; Orlando 2000) is a biochemical procedure for capturing in vivo TFBSs. In our case, we used ChIP assays coupled with real-time PCR with primers specific for the NF-κB sites (known and predicted) in the promoters of the *iNOS* and autotaxin genes (*ATX, ENPP2*), in order to confirm Footer predictions.

About 5×10^6 D2SC-1 cells (dendritic cell [DC] line) were propagated in DMEM/F12, 10% fetal calf serum. NF-κB nuclear translocation was induced by LPS (25 μg/mL final). Cells were collected prior induction and at 30-min, 1-h, and 2-h time-points post-induction. Proteins were cross-linked onto the DNA with formaldehyde (1% final 15 min at 37°C). Cells were lysed and sonicated (1-min total pulse at 20% amplitude on a COLE PALMER 750-W sonicator). The conditions for the sonication had been previously determined empirically to yield fragments of 500-bp average size (data not shown). The complexes precipitated for 18 h at 4°C with an anti-NF-κB (p65) monoclonal antibody sc-8008 (Santa Cruz Biotechnology). Cross-links were reversed by heating for 4 h at 65°C, and proteinase K treatment was used to purify the DNA fragments. As a control experiment, we performed a "mock ChIP" (i.e., precipitation without the primary antibody) in identical conditions with the LPS-induced cells for 1 h.

Real-time PCR on ChIP-precipitated DNA

Real-time PCR is a technique used to quantify the number of template DNA (or RNA) molecules that are in a sample. We used real-time PCR on the ChIP-precipitated DNA in order to identify how many NF-κB targets were captured for each of the four sites. The real-time PCR data correspond to the amount of DNA recovered in each precipitation (induced and non-induced). This percentage does not reflect the in vitro binding affinity of the NF-κB to its corresponding DNA sequence target(s). It is rather related to the percentage of the cells that NF-κB was bound to at the time on the particular target. Thus, it can be considered as the in vivo affinity (VIVA) of NF-κB to the corresponding DNA target, taking into consideration protein–protein interactions, chromatin accessibility status, etc. (Fernandez et al. 2003). This method for data collection is designed to overcome problems related to predicted targets, where the prediction score of the NF-κB target sequences is not directly associated with the binding strength (Hoffmann et al. 2003).

One tenth of each ChIP DNA precipitate was used as template in real-time PCR reactions under standard buffer conditions. PCR ran for 40 cycles in a BioRad iCycler. The PCR conditions were 30 sec at 95°C, 30 sec at 56°C, and 30 sec at 72°C. The primers were designed with the IDTDNA publicly available software (http://www.idtdna.com/), so that they would amplify a region of ~150 bp around each of the four NF-κB binding sites we tested. The sequences of the primers are as follows: iNOS1-for 5′-ATGGCCTTGCATGAGGATACACCA, iNOS1-rev 5′-GGTGG CTGAGAAGTTTCAAACCAG, iNOS2-for 5′-TCCTGTCAGGGA CAGATCCACTTT, iNOS2-rev 5′-TCTGATGATGGATGTGGCAG GTGA, autotaxin1-for 5′-TTGGAAGCTCCCATTGTGTGAAGC, autotaxin1-rev 5′-TCTGGCAGTTGGAATGACCCTGTA, autotaxin2-for 5′-GTAAACGCTTCGAGCTGATGGGAA, and autotaxin2-rev 5′-GCTGTGGCCAATAACAGTGCAT.

Results

Sensitivity and specificity

We measured the efficiency of Footer in terms of sensitivity and specificity. Sensitivity (S_N) is defined as the percent of successful predictions (compared with the total number of sites) and specificity (S_P) is defined as the percentage of predictions that are correct (compared with the total number of predictions). Naturally, in the case of prediction of TFBSs, the number of false negatives cannot be calculated accurately, since many TFBSs are yet to be discovered. Hence the S_P value should be considered as the lower limit of the true specificity.

Analysis of known binding sites

We tested the prediction efficiency of our algorithm on 72 confirmed TFBSs of 19 TFs in 24 promoter regions. In all cases, we analyzed the 3-kb region upstream and 50 bp downstream of the TSS. In each run, we "blindly" retained and analyzed the top 10 predictions per TF in each promoter, or one prediction per 300 bp (default value for our server). We chose this criterion instead of selecting the sites that score in the (e.g.) top 10% of the PSSM scores because we would like also to consider some of the lower scoring sites (if available). Lower scoring sites can be functional sites if the corresponding TF participates in a protein complex, which as a whole shows high affinity for the DNA. This has been observed in the case of homodimer proteins, where one of the two half sites might score low against a monomer PSSM model (Wu et al. 1998; Hollenbeck and Oakley 2000). Likewise, this can also happen in heterodimeric proteins. A summary of the results of our analysis is presented in Table 1. Detailed presentation of the results is provided in the Supplemental material.

Overall, Footer was able to correctly identify 60 of the 72 binding sites, hence exhibiting a *sensitivity* of 83.3% (Table 1). If we consider all additional (unverified) predictions to be false

Table 1. Summary of the results of predictions of programs Footer, ConSite and rVista

	Footer	ConSite (def)	ConSite (70%)	rVista
No. of sites	72	49	49	69
TP	60	23	34	54
FP	23	15	28	189
S_N	83.3%	46.9%	69.4%	78.2%
S_P	72.3%	60.5%	54.8%	22.2%

Footer ran with its default parameters (one "seed" site per 300 bp, WAP cutoff 0.05%). ConSite ran with its default parameters and also with 70% for score threshold and the minimum between 70% or the default for identity percentage. rVista ran with the option "conserved". Based on the results of the optimal runs of these programs on the same promoter set, their sensitivity and specificity values were measured to be 69% and 55% for ConSite (on 49 sites) and 78% and 22% for the rVista (on 69 sites), respectively. By comparison, Footer achieved a sensitivity of 83% and specificity of 72% (on 72 sites).

positives, then Footer exhibited a specificity of 72.3% (i.e., 60 of the 83 predictions are confirmed). However, six of these additional predictions are located outside of the regions that were biochemically analyzed, and some of them differ from a true site by a single base. Note that if we exclude from the analysis the 18 TFBSs we used to assign the weight values in equation 4 (i.e., the TFBSs at the promoters of the genes PEPCK, G6Pase, and Leptin), the results are consistent (S_N = 87% and S_P = 70%). This indicates that the initial 18 selected sites constitute a representative sample.

Prediction of sites in nonconserved regions

We found that most of the known sites were located in conserved genomic regions as these were defined by the alignment program DBA (Jareborg et al. 1999). Three out of the 72 confirmed sites were located in nonconserved regions: one NFY site (*ACDC* promoter at −117), one Sp1 site (*MMP9* promoter at −560), and one HNF-1α site (Pdx-1 promoter at −2114). In the first two cases, both human and mouse sites were in a nonconserved region, whereas in the case of the HNF-1α site, the human site was located in a conserved region. Footer predicted correctly the NFY and the HNF-1α site.

Performance of Footer with respect to the WAP threshold

The WAP parameter is the most important parameter of our method. Thus, we tested Footer performance on multiple WAP thresholds. The results are presented in Figure 2. Footer performs best at a WAP threshold of 0.05%, which is the value we used for the analysis of all genes in this article.

Performance of Footer with respect to the promoter length

The distance probability depends on the size of the promoter (equation 2). Therefore, we tested Footer on various promoter lengths, in order to better evaluate it. The results on the sensitivity and specificity are shown in Figure 3. Specificity shows a maximum value at 2.5kb (74%), but it remains essentially constant over all promoter lengths up to 3.5 kb. Sensitivity, on the other hand, increases with the promoter length up to the size of 3 kb. This can be attributed to the fact that a considerable number of confirmed TFBSs (~12%) are located in the region beyond 1 kb from the TSS; but it is also due to the fact that the number of "seed" sites that Footer retains and analyzes is proportional to the promoter length. We did not test Footer in promoter lengths >3.5 kb, since all the confirmed sites are within that region (Table 1).

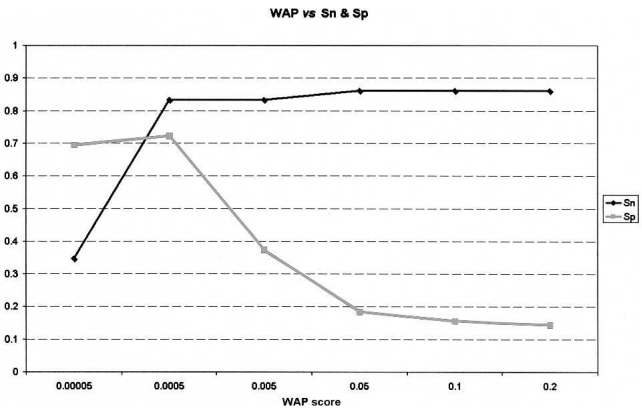

Figure 2. Performance of Footer in relation to the WAP threshold. This graph presents the sensitivity (black line) and specificity (gray line) over all promoter regions analyzed (see text). According to this graph, Footer performs best on a WAP threshold of 0.05%.

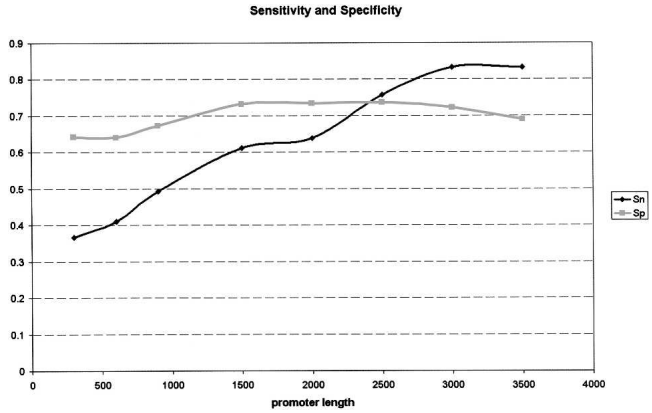

Figure 3. Sensitivity and specificity performance of Footer in relation to the analyzed promoter length. This graph presents the sensitivity (black line) and specificity (gray line) over all promoter regions analyzed (see text). According to this graph, Footer performance increases with examined promoter length up to 3 kb, while specificity remains essentially constant.

Comparison with other methods

We compared Footer with two algorithms (online versions) that use comparative genomics and PSSM models to identify phylogenetically conserved signals in mammals: ConSite (Lenhard et al. 2003) and rVista (Loots et al. 2002). The results are summarized in Table 1, while a detailed description of the comparison procedure is provided in the Supplemental material section. Overall, Footer outperformed these methods by predicting more true sites without increasing the number of false-positive predictions. ConSite had models for 49 of the 72 sites and with the default parameters showed S_N = 46.9% and S_P = 60.5%. By lowering the score threshold to 70% (default value: 80%) and running on the promoters that did not perform well with the default parameters, its S_N increased to 69.4% with S_P = 54.8%. Note that searching for the same 49 sites, Footer was able to find 44 (S_N = 89.8%) making 15 additional predictions (S_P = 74.6%) (see Supplemental material, Supplemental Table 2). rVista has PSSM models for 69 sites and it performed well in finding 54 of them, but it produced 189 additional predictions (false positives) (Supplemental Table 2). So, its overall performance was S_N = 78% and S_P = 22%. Note that searching for the same 69 sites, Footer was able to find 59 (S_N = 85.5%), making 21 additional predictions (S_P = 73.8%) (see Supplemental material, Supplemental Table 2).

Successful prediction of two novel NF-κB binding sites in the promoter of autotaxin gene

We used Footer to predict NF-κB binding sites in the promoters of the genes *iNOS* and autotaxin. NF-κB is a key factor in the onset and progression of type 1 diabetes mellitus (T1DM) (Weaver Jr. et al. 2001; Poligone et al. 2002; Sen et al. 2003). There are two known NF-κB sites in the promoter of *iNOS* gene at positions −114 and −1044, identified in inflammatory stimulated cells (Wei et al. 2004). The promoter of autotaxin gene has not been previously shown to contain NF-κB sites, although some reports indicate that NF-κB could be involved in its regulation during maturation of DCs (Le Naour et al. 2001). We analyzed the pro-

moters of these two genes with Footer, and we found three new putative NF-κB binding sites: one in the promoter of *iNOS* (position −2760) and two at the promoter of autotaxin (positions −375 and −596).

ChIP assays were performed as we describe in the Methods section on D2SC-1 cells induced with LPS for 30 min, 1 h, and 2 h. ChIP on non-induced cells was also performed (control experiment). Figure 4 plots the proportion of the ChIP precipitated DNA targets in the LPS-induced cells relative to the non-induced cells. For each site, we plot the time point where the NF-κB association was highest based on the real-time PCR values. These data confirm the strong NF-κB association to site iNOS114 (known site; five times more iNOS114 targets precipitated post-induction compared to the non-induced cells). Furthermore, these data show that NF-κB is associated with *both* predicted autotaxin sites (almost fourfold for autotaxin375 and 6.5-fold for autotaxin596), and there seems to be no association to iNOS2760 site under these conditions (Fig. 4). Note that although the two autotaxin sites are located within 220 bp, the real-time PCR on ChIP DNA yield appears to distinguish the NF-κB association patterns in the two. This is more obvious when patterns on all three time-points are compared (data not shown). In all cases, the real-time PCR values of the mock experiment were at the same level with those of the non-induced cells (data not shown).

Footer Web server

In order to make Footer available to the scientific community, we developed a Web server. The server receives a pair of DNA sequences in its input (typically, human and mouse homologous promoter sequences) and then it performs the sequence alignment (program DBA) (Jareborg et al. 1999), pattern matching, and pattern comparison in order to find the most probable site pairs. The user-defined parameters are the list of TFs, the WAP, the w_D and w_S weights (equation 4), and the number of "seed" sites considered initially per promoter. The patterns that score below the user specified WAP threshold are reported together with the color-coded alignment that depicts the conservation throughout the DNA sequence length. The server also accepts a protein sequence as input of human or mouse/rat origin. In this case, BLAST searches (Altschul et al. 1990) are used to identify the closest mouse or human protein homologs, respectively, and back searches performed to establish orthology. If orthology is not established the user is informed. In any case, the two proteins are used to search the corresponding UniGene collections (Wheeler et al. 2003), so that the longest mRNA sequences are identified and used against the genome sequences to determine the corresponding TSSs. Subsequently, a user-defined number of bases are extracted around the TSSs, and the two DNA sequences are compared with Footer. The results are presented in tabular format. A detailed description of the Web server functionality will be presented soon (Corcoran et al. 2005). Footer is available to the public through our server (http://biodev.hgen.pitt.edu/cgi-bin/Footer/Footer.cgi).

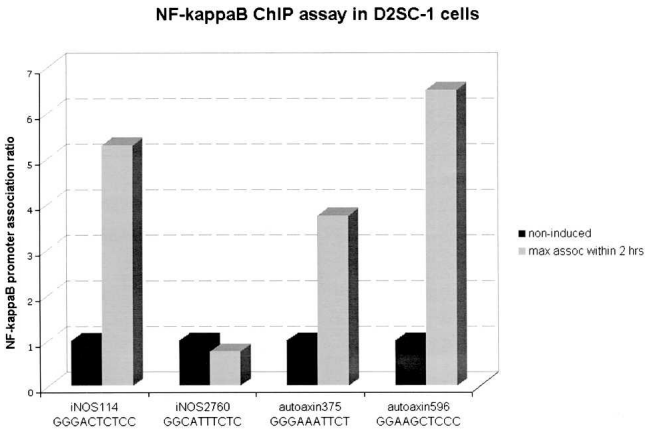

Figure 4. Real-time PCR data on NF-κB association to one known and three predicted binding sites. The real-time PCR data correspond to the in vivo NF-κB association to known and predicted sites in the promoters of genes *iNOS* and autotaxin (two sites in each promoter). This graph confirms the direct association of NF-κB with iNOS114 (known site at position −114 from the TSS) and reveals that NF-κB binds on both predicted sites in autotaxin gene (at positions −375 and −596, respectively). The peaks show the maximum fold association compared to non-induced cells (control sample) in the time interval between 30-min and 2-h post-induction. This analysis does not confirm a second predicted iNOS site (at position −2760). All predictions were made by Footer, using the default parameters described in the text.

Discussion

Transcription is a complex biological process, and its regulation depends on many factors like the presence of TFBSs in the promoters of the genes, the state of the chromatin, the localization of the TF molecules, the interactions between TFs and other protein or RNA molecules, etc. Computationally, the identification of TFBSs in large mammalian genomic sequences is a very difficult task. This is due to the low signal-to-noise ratio inherent with all DNA regulatory "signals," but also due to the more complex regulatory mechanisms, which might require or allow some functional sites to be of low affinity. The course of evolution provides a rich source of information for unraveling this biological complexity. Our method, Footer, uses prior information organized in species- or mammalian-specific PSSM models and applies a comparative genomics strategy in order to detect the most likely DNA regulatory signals given a promoter region.

With this article we are introducing two new concepts. One is that the conservation in the position of the TFBS in a promoter sequence and its agreement with known binding preferences of the corresponding TF can be quantified and used jointly to predict efficiently TFBSs. Other methods had used the position conservation to different extents, but we combine the two criteria in a single metric. Interestingly, the real-time PCR data in the four sites we analyzed (Fig. 4) exhibit a strong anti-correlation with Footer calculated WAP scores ($R = -0.96$; the lower the WAP score the higher the measured NF-κB association). Of course, the small sample size does not allow for any statistically significant conclusions to be drawn, but we think it is an interesting finding that shows the potential of Footer.

The second concept we introduce with this work is that the variability in multiorganism TFBSs should be treated with caution. Our recent data analysis (Mahony et al. 2005) showed that there are notable exceptions where the evolution of the familial PSSM models does not follow the evolution of the protein families. Although, detailed analysis has not been performed with species-specific models, these are expected to play an important role in Footer performance, due to its quantitative use of the PSSM score. We suggest that organism-specific or class-specific (in our case, mammalian) PSSM models should be used when enough sequences are available. Our PSSM models contain between six (GATA-3) and 108 (Sp1) sequences, and they performed

reasonably well in our test set. By comparison, TRANSFAC matrices (Wingender 2004) are constructed from as little as four sequences (e.g., Elk-1). We are currently looking on ways to increase the size of our data set, especially for the models with a small number of sites, by including data from other sources.

In general, we found that our scoring system is successful in finding most of the known binding sites on a test set comprised of a total of 72 sites in 24 human–mouse promoters. In addition, it successfully predicted two novel NF-κB sites in the promoter of the autotaxin gene, which we confirmed by ChIP assay coupled with quantitative real-time PCR. Overall, Footer exhibited sensitivity of 83.3% and specificity of 72.3%, which surpasses existing methods (Loots et al. 2002; Lenhard et al. 2003). The sensitivity of Footer increased with the promoter length examined (up to 3 kb). Three out of 72 sites (4%) were located in nonconserved regions, as these were identified by the alignment program DBA (Jareborg et al. 1999) and Footer correctly predicted two of them. Predicting sites in nonconserved regions is currently difficult, since most available algorithms tend to exclude nonconserved regions from the analysis.

Ten sites of the 72 in the test set were included in the PSSM models Footer used for its predictions. These are one NFAT site, one HNF-3β site, one C/EBPβ site, one CREB site, one GR-α site, one T3R-α site, one NF-1 site, one MEF-2 site, and two NF-κB sites. Exclusion of each of these sites from the corresponding PSSM models resulted in correct identification of the seven (NFAT, HNF-3β, C/EBPβ, T3R-α, NF-1, and both NF-κB sites), whereas Footer was not able to predict the MEF-2, CREB, or the GR-α sites (the GR-α site was not found in the original analysis as well). Note that if we exclude from the analysis these 10 sites, the sensitivity and specificity values are still fairly high (82% and 60%, respectively). We do not know how many of the sites in our test set were included in the models that ConSite and rVista use.

Similarly to the ConSite method (Lenhard et al. 2003), Footer scans both promoter sequences for high-scoring pairs of TFBS. But Footer and ConSite have a number of differences. ConSite compares only sites that are located within conserved regions (default is 80% identity over a window of 50 bases). Furthermore, it uses the PSSM scores for filtering out those sequences that do not score above a threshold. The top 10% of the conserved windows are analyzed further for concurrent instances of a TF "hit," which then is reported as "true TFBS." Footer, on the other hand, scans both sequences for the top 10 high-scoring TFBS per 3 kb of promoter sequence (default value; no score threshold applied). This allows Footer to include even some weaker signals in this initial pattern matching. Then, the patterns are analyzed pairwise with respect to their matching score (based on the species-specific PSSM models, when available) and their relative location in the promoter. Putative binding sites on both the conserved and nonconserved regions are considered.

PSSM models

Currently, the most widely used way to represent TF binding preferences is via PSSM models (Stormo 2000). Previous reports have pointed out that in some cases the PSSM models might constitute an oversimplification of the real binding preferences (Barash et al. 2003; Zhou and Liu 2004) and hence predictions using such models might not be accurate. Although currently Footer uses PSSM models for its initial pattern matching and the calculation of the PF_s value (equation 4), it can easily utilize any statistical representation (see also equation 3). The data for constructing the PSSM models are currently obtained from the TRANSFAC database (Wingender 2004), but well-curated data sets have started to become available (Sandelin et al. 2004).

Limitations and generalization of comparative genomics methods

One limitation of all comparative genomics methods, including Footer, is that they cannot identify regulatory elements that have been acquired/become extinct after the speciation of the compared species. For this reason, some researchers have proposed the comparison of multiple species to determine the true regulatory elements (Duret and Bucher 1997; McCue et al. 2002; Cliften et al. 2003). Footer provides a probability-based scoring function, which makes it potentially expandable to comparisons of three or more species. Another limitation of the comparative genomics methods is the presence of insertions/deletions in the promoters of the genes. ConSite (Lenhard et al. 2003) addresses this problem by focusing on the conserved regions only. Footer, on the other hand, uses promoter alignments to recalibrate the point of reference for calculation of the distance between two predicted sites. This diminishes the number of false negatives due to local insertions/deletions. Finally, the lack of availability of species-specific PSSM models presents another limitation to our program. We believe that the technologies such as ChIP (Orlando 2000) and SELEX (Choo and Klug 1994), for the high-throughput identification of TFBS in vivo will soon remove this obstacle. We expect that in future years, algorithms like Footer will become essential tools for the analysis of the wealth of the accumulated data in complex eukaryotic genomes.

Acknowledgments

We thank Dr. Gary Stormo and Dr. Wyeth Wasserman for critically reviewing earlier versions of this work. We are also grateful to three anonymous reviewers for their comments that helped us improve this manuscript significantly. This work was supported by NSF grant MCB0316255 and by intramural funds of the Department of Computational Biology, the University of Pittsburgh Cancer Institute (UPCI), School of Medicine and the Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Barash, Y., Elidan, G., Friedman, N., and Kaplan, T. 2003. Modeling dependencies in protein–DNA binding sites. In *Seventh Annual International Conference on Computational Molecular Biology (RECOMB)*.

Benos, P.V., Lapedes, A.S., Fields, D.S., and Stormo, G.D. 2001. SAMIE: Statistical algorithm for modeling interaction energies. *Pac. Symp. Biocomput.* 115–126.

Benos, P.V., Bulyk, M.L., and Stormo, G.D. 2002a. Additivity in protein–DNA interactions: How good an approximation is it? *Nucleic Acids Res.* **30:** 4442–4451.

Benos, P.V., Lapedes, A.S., and Stormo, G.D. 2002b. Probabilistic code for DNA recognition by proteins of the EGR family. *J. Mol. Biol.* **323:** 701–727.

Choo, Y. and Klug, A. 1994. Selection of DNA binding sites for zinc fingers using rationally randomized DNA reveals coded interactions. *Proc. Natl. Acad. Sci.* **91:** 11168–11172.

Cliften, P., Sudarsanam, A., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B.A., and Johnston, M. 2003. Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* **301:** 71–76.

Corcoran, D.L., Feingold, E., and Benos, P.V. 2005. FOOTER: A web tool

for finding mammalian DNA regulatory regions using phylogenetic footprinting. *Nucleic Acids Res*. (in press).
Duret, L. and Bucher, P. 1997. Searching for regulatory elements in human noncoding sequences. *Curr. Opin. Struct. Biol.* **7:** 399–406.
Fernandez, P.C., Frank, S.R., Wang, L., Schroeder, M., Liu, S., Greene, J., Cocito, A., and Amati, B. 2003. Genomic targets of the human c-Myc protein. *Genes & Dev.* **17:** 1115–1129.
Hertz, G.Z. and Stormo, G.D. 1999. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **15:** 563–577.
Hoffmann, A., Leung, T.H., and Baltimore, D. 2003. Genetic analysis of NF-κB/Rel transcription factors defines functional specificities. *EMBO J.* **22:** 5530–5539.
Hollenbeck, J.J. and Oakley, M.G. 2000. GCN4 binds with high affinity to DNA sequences containing a single consensus half-site. *Biochemistry* **39:** 6380–6389.
Jareborg, N., Birney, E., and Durbin, R. 1999. Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Res*. **9:** 815–824.
Jegga, A.G., Sherwood, S.P., Carman, J.W., Pinski, A.T., Phillips, J.L., Pestian, J.P., and Aronow, B.J. 2002. Detection and visualization of compositionally similar *cis*-regulatory element clusters in orthologous and coordinately controlled genes. *Genome Res.* **12:** 1408–1417.
Le Naour, F., Hohenkirk, L., Grolleau, A., Misek, D.E., Lescure, P., Geiger, J.D., Hanash, S., and Beretta, L. 2001. Profiling changes in gene expression during differentiation and maturation of monocyte-derived dendritic cells using both oligonucleotide microarrays and proteomics. *J. Biol. Chem.* **276:** 17920–17931.
Lenhard, B., Sandelin, A., Mendoza, L., Engstrom, P., Jareborg, N., and Wasserman W.W. 2003. Identification of conserved regulatory elements by comparative genome analysis. *J. Biol.* **2:** 13.
Loots, G.G., Ovcharenko, I., Pachter, L., Dubchak, I., and Rubin, E.M. 2002. rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res.* **12:** 832–839.
Mahony, S., Golden, A., Smith, T.J., and Benos, P.V. 2005. Improved detection of DNA motifs using a self-organized clustering of familial binding profiles. *Proc. Intell. Syst. Mol. Biol*. (in press).
McCue, L.A., Thompson, W., Carmack, C.S., and Lawrence, C.E. 2002. Factors influencing the identification of transcription factor binding sites by cross-species comparison. *Genome Res.* **12:** 1523–1532.
Orlando, V. 2000. Mapping chromosomal proteins in vivo by formaldehyde-crosslinked-chromatin immunoprecipitation. *Trends Biochem. Sci.* **25:** 99–104.
Orlando, V., Strutt, H., and Paro, R. 1997. Analysis of chromatin structure by in vivo formaldehyde cross-linking. *Methods* **11:** 205–214.
Peters, D.G., Zhang, X.C., Benos, P.V., Heidrich-O'Hare, E., and Ferrell, R.E. 2002. Genomic analysis of immediate/early response to shear stress in human coronary artery endothelial cells. *Physiol. Genomics* **12:** 25–33.
Poligone, B., Weaver Jr., D.J., Sen, P., Baldwin Jr., A.S., and Tisch, R. 2002. Elevated NF-κB activation in nonobese diabetic mouse dendritic cells results in enhanced APC function. *J. Immunol.* **168:** 188–196.
Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W.W., and Lenhard, B. 2004. JASPAR: An open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* **32:** D91–D94.
Schwartz, S., Elnitski, L., Li, M., Weirauch, M., Riemer, C., Smit, A., Green, E.D., Hardison, R.C., and Miller, W. 2003. MultiPipMaker and supporting tools: Alignments and analysis of multiple genomic DNA sequences. *Nucleic Acids Res.* **31:** 3518–3524.
Sen, P., Bhattacharyya, S., Wallet, M., Wong, C.P., Poligone, B., Sen, M., Baldwin Jr., A.S., and Tisch, R. 2003. NF-κB hyperactivation has differential effects on the APC function of nonobese diabetic mouse macrophages. *J. Immunol.* **170:** 1770–1780.
Stormo, G.D. 2000. DNA binding sites: Representation and discovery. *Bioinformatics* **16:** 16–23.
Tagle, D.A., Koop, B.F., Goodman, M., Slightom, J.L., Hess, D.L., and Jones, R.T. 1988. Embryonic ε and γ globin genes of a prosimian primate (*Galago crassicaudatus*): Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J. Mol. Biol.* **203:** 439–455.
Weaver Jr., D.J., Poligone, B., Bui, T., Abdel-Motal, U.M., Baldwin Jr., A.S., and Tisch, R. 2001. Dendritic cells from nonobese diabetic mice exhibit a defect in NF-κB regulation due to a hyperactive I κB kinase. *J. Immunol.* **167:** 1461–1468.
Wei, J., Guo, H., Gao, C., and Kuo, P.C. 2004. Peroxide-mediated chromatin remodelling of a nuclear factor κB site in the mouse inducible nitric oxide synthase promoter. *Biochem. J.* **377:** 809–818.
Wheeler, D.L., Church, D.M., Federhen, S., Lash, A.E., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M., Sequeira, E., Tatusova, T.A., et al. 2003. Database resources of the National Center for Biotechnology. *Nucleic Acids Res.* **31:** 28–33.
Wingender, E. 2004. TRANSFAC, TRANSPATH and CYTOMER as starting points for an ontology of regulatory networks. *In Silico Biol.* **4:** 55–61.
Workman, C.T., Yin, T., Corcoran, D.L., Ideker, T., Stormo, G.D., and Benos, P.V. 2005. EnoLOGOS: A versatile web tool for energy normalized sequence logos. *Nucleic Acids Res.* (in press).
Wu, X., Spiro, C., Owen, W.G., and McMurray, C.T. 1998. cAMP response element-binding protein monomers cooperatively assemble to form dimers on DNA. *J. Biol. Chem.* **273:** 20820–20827.
Zhou, Q. and Liu, J.S. 2004. Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics* **20:** 909–916.

Web site references

http://www.idtdna.com/; IDTDNA software for designing PCR primers.
http://biodev.hgen.pitt.edu/cgi-bin/Footer/Footer.cgi; Footer Web server for analysis of mammalian promoters.
http://biodev.hgen.pitt.edu/cgi-bin/enologos/enologos.cgi; enoLOGOS Web server for sequence LOGOS.

Received July 7, 2004; accepted in revised form April 5, 2005.