Published online 22 October 2009

Nucleic Acids Research, 2010, Vol. 38, No. 1 ***17–25***
doi:10.1093/nar/gkp866

Modeling tissue-specific structural patterns in human and mouse promoters

Alexis Vandenbon[1] and Kenta Nakai[1,2,3,*]

[1]Department of Medical Genome Sciences, Graduate School of Frontier Sciences, [2]Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan and [3]Institute for Bioinformatics Research and Development (BIRD), Japan Science and Technology Agency, 5-3 Yonbancho, Chiyoda-ku, Tokyo 102-0081, Japan

Received June 26, 2009; Revised September 9, 2009; Accepted September 28, 2009

ABSTRACT

Sets of genes expressed in the same tissue are believed to be under the regulation of a similar set of transcription factors, and can thus be assumed to contain similar structural patterns in their regulatory regions. Here we present a study of the structural patterns in promoters of genes expressed specifically in 26 human and 34 mouse tissues. For each tissue we constructed promoter structure models, taking into account presences of motifs, their positioning to the transcription start site, and pairwise positioning of motifs. We found that 35 out of 60 models (58%) were able to distinguish positive test promoter sequences from control promoter sequences with statistical significance. Models with high performance include those for liver, skeletal muscle, kidney and tongue. Many of the important structural patterns in these models involve transcription factors of known importance in the tissues in question and structural patterns tend to be conserved between human and mouse. In addition to that, promoter models for related tissues tend to have high inter-tissue performance, indicating that their promoters share common structural patterns. Together, these results illustrate the validity of our models, but also indicate that the promoter structures for some tissues are easier to model than those of others.

INTRODUCTION

The activity of genes and their products is regulated on many levels. As one of the first levels, regulation of transcription plays an important role in this regulatory system. Transcription factors (TFs) bind the regulatory regions of genes on specific sites in the genome—commonly referred to as transcription factor binding sites (TFBSs) or *cis*-regulatory motifs—and in this way help the formation of the transcription initiation complex. Therefore, the understanding of TFs and their binding sites is key to understanding the regulation of transcription. As such, they have received considerable attention, both in the field of their prediction and the modeling of their binding motifs (1,2), and their functions in different tissues (3,4).

It is becoming more and more clear that the regulation of transcription in higher eukaryotes is a complex process. A single promoter can contain multiple binding sites for up to 10 or more different TFs, and significant cooperation and competition exist between different regulatory proteins (5). Therefore, in higher organisms, it is no longer sufficient to regard TFs as individually operating molecules and it is not surprising that attention is shifting towards modeling regulatory regions on a higher level. Various techniques, such as thermodynamic models (6,7), Bayesian networks (8,9), hidden Markov models (10,11) and Markov chains (12) have been used, and recent studies have successfully focused attention on local clusters of TFBSs, so called *cis*-regulatory modules (CRMs). Smith *et al.* (13) modeled CRMs using the MARS algorithm and used their models to predict whether genes are up- or down-regulated in 28 tissues. Blanchette *et al.* (14) scanned the genome for CRMs, computing scores from conserved sites using windows of different sizes. Van Loo *et al.* (15) described an approach for modeling CRMs, using a Genetic Algorithm (GA) to find an optimal set of motifs distinguishing positive promoter sequences from others. However, in general these studies tend to focus only on the enrichment of predicted TFBSs, ignoring other structural patterns such as relative distances or orientation of sites.

Previously, we have presented an approach for the modeling of the structure of promoters driving tissue-specific expression (16). In this approach, structural patterns are generated on the presence of motifs and their positioning with regard to the transcription start site (TSS) and to each other. Subsequently a GA selects

*To whom correspondence should be addressed. Tel: +81 3 5449 5131; Fax: +81 3 5449 5133; Email: knakai@ims.u-tokyo.ac.jp

a small set of structural patterns that optimally distinguishes positive training promoters from negative training promoters. The model was applied successfully on a set of muscle-specific genes in the nematode *C. elegans*.

Here we present an updated version of this model and its application on sets of genes expressed specifically in 26 human and 34 mouse tissues. Using 10-fold cross-validation (CV) we evaluate the capability of the model to predict tissue-specific expression of genes based on structural patterns in their promoter regions. In about half of the sets the model performs significantly better than expected at random. We find that the performance of our model is highly variable between different tissues. Interestingly, tissues traditionally used in the study of tissue-specific regulation of transcription, such as skeletal muscle, liver and kidney, tend to be easier to model. We find high performance when applying human models on mouse datasets and vise versa, indicating that the structural patterns are conserved between human and mouse promoters. Additionally, high performance is observed when applying models trained on one tissue on promoters expressed specifically in functionally related tissues. Finally, the GA in this updated model assigns weights to the structural patterns, and top structural patterns turn out to contain many patterns describing the positioning of motifs proximal to the TSS. This stresses the importance of not focusing all attention solely on CRM modeling.

MATERIALS AND METHODS

Tissue-specific datasets

As expression data we used the microarray expression data of Su *et al*. (17), which includes data for 79 tissues and cell types in human and 61 tissues and cell types in mouse. For each gene we processed raw expression data to *Z*-scores, using the mean and the standard deviation of the gene's signal over all tissues. We defined genes as 'tissue-specific' when the *Z*-score exceeded the threshold value of 3. Genes with *Z*-scores exceeding the threshold in multiple tissues are considered 'tissue-specific' in each of these tissues. All tissues with at least 50 tissue-specific genes were used for further analysis.

TSS data and promoter sequences

TSSs of genes were defined using a combination of data DBTSS (for both human and mouse data) (18), CAGE data (for mouse) (19) and data from the UCSC Genome Browser (20) (for both human and mouse). Where possible TSSs were assigned using data from DBTSS. If not, CAGE data (for mouse) and UCSC TSS annotations were used. The regions from -1000 to $+200$ were extracted from the repeat-masked hg18 and mm9c versions of the human and mouse genome, respectively.

Training, validation and test samples

We used 10-fold CV for evaluating performance of the models. Therefore, positive promoter sequences were randomly divided into 10 partitions. Each of the 10 partitions was used once as test set. The remaining 90% were used for training the models. They were further divided randomly into two parts of equal size; a training set and a validation set. Motif prediction and generation of structural patterns were performed on the training set. The validation set was used in the GA runs, for selecting individuals for the next generations. Finally, performance was evaluated on the test set.

Control sequences were sampled from the genomic set of promoter sequences, excluding positive sequences. No overlap was allowed between sequences for training and testing purposes. We prefer the term 'control sequences' over the term 'negative sequences', as the promoters we used as controls are not necessarily inactive in the tissue under consideration.

Motif prediction, evaluation and selection

In training sequences we predicted *de novo* motifs using a set of four popular motif prediction programs, and converted motifs to PWMs. Subsequently, we scanned these sequences, and a set of randomly picked up control promoters using all *de novo* PWMs and all vertebrate PWMs from the TRANSFAC database version 2008.4 (21). For each motif a measure for over-representation was calculated from the predicted sites in the positive and control sequences. For this measure of over-representation, a corresponding *p*-value was calculated using random sampling from a second set of randomly selected promoters. Redundancies in the PWM set were removed using tomtom (22), stepwise removing less over-represented PWMs from pairs of similar PWMs. From the final set of non-redundant PWMs, PWMs with a $p\text{-value} \leq 0.01$ were retained. We refer to the Supporting material for a more detailed description.

Promoter structure model

Using this final set of PWMs the positive and control training and test sequences were scanned, and based on the predicted sites the promoter models were trained. For a detailed description we refer to Vandenbon and Nakai (16). Here we limit the description to a short summary, and a description of the changes made to the model.

From the predicted sites for the final set of PWMs in the training sequences structural patterns over-represented in the positive sequences were generated. Three types of patterns are considered: (i) presence of sites anywhere in the promoter sequences, (ii) presence of sites within a certain region relative to the TSS and (iii) relative positioning of pairs of sites. From the large set of generated patterns, a GA subsequently selects a set of meaningful patterns. Briefly, in GA optimization, initially a population of individuals is generated with each individual representing a possible solution to the problem of interest. In this study, the solutions are subsets of structural patterns. For each individual in the population, the solution it encodes is applied on the training samples and the resulting performance is used as a measure of the fitness of the individual. Subsequently, individuals are selected stochastically based on their fitness, and

modified to form the next generation. This process is repeated for a number of generations, until a suitable solution is obtained. For a more detailed introduction on GAs we refer to Johnson and RahmatSamii (23).

While the GA in the original study was rather simple, we have made some changes in order to get more biologically meaningful results, and to avoid over-fitting the training data. First of all, instead of just selecting a number of patterns from the total set of patterns, the new GA also gives weights to each selected pattern, and adjusts the weights in a way that improves performance on the training sequences. Let P be the set of all structural patterns; $C(s,P)$ the vector of counts of patterns present in promoter sequence s; and W, the vector of weights of the patterns as allocated by the GA. Then,

$$Score(s,P,W) = C(s,P)^t W, \quad \mathbf{1}$$

where $Score(s,P,W)$ is the score of promoter s given P and W.

The GA adjusts the vector W in such a way that the fitness of the individuals in the GA is increased. Let F be the fitness of an individual. Whereas in the original model as described by Vandenbon and Nakai (16), F was only a function of the area under the curve (AUC) of the receiver operator characteristic (ROC) curve of the training sequences, we have expanded the definition of F to also include a direct measure of the over-representation of true positive training sequences among the top scoring sequences. Let $Sens_{95}$ represent the sensitivity at 95% specificity, and let $Sens_{90}$ represent the sensitivity at 90% specificity. Then, the fitness of each individual is defined to be

$$F(AUC, Sens_{95}, Sens_{90}) = (AUC/0.5) + (Sens_{95}/0.05) + (Sens_{90}/0.1) \quad \mathbf{2}$$

A third change is in the use of the training data. While using separate sets of samples for training a model and testing gives a more correct estimation of the performance of the model than training and testing on the same samples, it does not necessarily prevent the model from over-learning the training data. Here, we used the following approach for reducing over-learning: we used a separate set of samples for estimating performance during training (referred to as the validation set), as described in Gagne *et al.* (24). In our GA, in the reproduction step, individuals were selected based on their fitness on the training sequences. Subsequently, during the selection step, individuals were selected from the expanded population based on their fitness on the validation set. This way, individuals performing well on the training set but not on the validation set (e.g. over-fitted individuals) will not be selected for the next generation. Note that the training and validation sets do not contain any of the test set sequences. Finally, we introduced a penalty function to limit the complexity of the models: individuals received a penalty as the number of patterns with non-zero weights increases. A simulated annealing approach was used, resulting in lower penalties at the start of training, and gradual increase of penalties as training progresses.

For each of the tissue datasets, for each of the 10 CV runs, the GA was run 100 times. The weights given to each pattern were averaged over the 100 runs to make the final weights of each model.

Evaluation of performance

Using the averaged weights of each pattern, and the counts for each pattern in each promoter sequence, the final score of each sequence was calculated. From the scores of the positive test sequences, and the control test sequences, a ROC curve was constructed and the AUC value was calculated for each CV set. Two other measures for performance are the sensitivity at 95% specificity and the sensitivity at 90% specificity. For these three measures of performance, p-values were calculated by random shuffling of the scores for positive and negative test sequences. From these shuffled scores the three measures of performance were calculated. This was repeated one million times. For each of the actual measures of performance, the fraction of shuffled measures of performance equal to or higher than the actual measure was used as P-value.

RESULTS AND DISCUSSION

Tissue-specific motif over-representation

In a first step we identified over-represented *de novo* and known TFBS in the training sequences of each of the 60 sets of tissue-specific genes. For many tissues we found known vertebrate TFBS motifs from the TRANSFAC database to be significantly over-represented in each of the promoter sets (see Tables 1 and 2 in the Supplementary Data for an overview). For many of the over-represented TFBSs, the role of the corresponding TF in the tissue in question is known and has been extensively reported in the literature, and in many cases similar motifs were found to be over-represented in human and mouse tissues. Some examples are ETS binding sites (including PU.1) in myeloid cells, monocytes, NK cells, and B and T cells (25–30), HNF4 sites in human and mouse liver, and mouse small and large intestine (31,32), HNF1 sites in human fetal liver, human and mouse liver (33), and MEF2 sites in human and mouse skeletal muscle tissue (34,35), SRF sites in human skeletal muscle and mouse heart (34,36). Other *in silico* studies on similar datasets have reported similar results (3,4,13).

Promoter structure model performance

For each dataset we constructed a set of non-redundant over-represented PWMs, and generated structural patterns concerning their presence and positioning in the promoter sequences. A GA was used to train a promoter structure model for each of the datasets, by assigning a weight to each structural pattern (see Materials and Methods section for a more detailed description). Finally, the performance of the trained models was evaluated by applying them on test sequences.

Table 1 shows an overview of the performance of the model on the human and mouse datasets. It is obvious

Table 1. Overview of results for (A) 26 human and (B) 34 mouse tissues and cell types

Description	Size (No. of seqs)	AUC value	
		Value	Corrected *p*-value
Human datasets			
Tongue	76	0.8066	<6.0e−5
Fetal liver	89	0.7879	<6.0e−5
Kidney	95	0.7056	<6.0e−5
Skeletal muscle	67	0.6986	<6.0e−5
Liver	276	0.6814	<6.0e−5
Testis interstitial	131	0.6680	<6.0e−5
Bronchial epithelial cells	75	0.6656	<6.0e−5
Placenta	124	0.6521	<6.0e−5
PB- CD14+ monocytes	142	0.6272	6.0e−5
Testis	159	0.6028	1.8e−4
Heart	80	0.6411	4.2e−4
Pancreas	58	0.6592	7.8e−4
Lung	74	0.6390	1.4e−3
BM- CD71+ early erythroid	187	0.5799	4.4e−3
Whole blood	110	0.5917	0.024
PB- CD56+ NK cells	146	0.5757	0.043
BM- CD33+ myeloid	160	0.5702	0.063
PB- CD8+ T cells	62	0.6064	0.11
721 B lymphoblasts	215	0.5478	0.46
Smooth muscle	81	0.5676	1.0
PB- BDCA4+ dentritic cells	95	0.5619	1.1
Testis leydig cell	66	0.5718	1.3
Adipocyte	50	0.5737	2.1
PB- CD19+ B cells	70	0.5579	2.8
BM- CD105+ endothelial	53	0.5547	5.1
BM- CD34+	85	0.5227	14.1
Mouse datasets			
Small intestine	205	0.7267	<6.0e−5
Tongue	102	0.7148	<6.0e−5
Snout epidermis	125	0.7018	<6.0e−5
Digits	105	0.6799	<6.0e−5
Liver	325	0.6736	<6.0e−5
Kidney	213	0.6677	<6.0e−5
Eye	103	0.6653	<6.0e−5
Testis	787	0.6428	<6.0e−5
Large intestine	124	0.6403	<6.0e−5
Fertilized egg	589	0.6396	<6.0e−5
Thyroid	156	0.6314	<6.0e−5
Skeletal muscle	145	0.6209	<6.0e−5
Oocyte	655	0.6147	<6.0e−5
Pancreas	381	0.5794	<6.0e−5
Umbilical cord	83	0.6626	6.0e−5
Bone	87	0.6372	3.0e−4
Placenta	109	0.6203	4.2e−4
Bone marrow	96	0.6258	6.0e−4
CD4+ T-cells	64	0.6476	1.0e−3
Heart	75	0.6365	1.1e−3
Dorsal root ganglia	63	0.6367	4.4e−3
Blastocysts	165	0.5685	0.068
Stomach	77	0.5882	0.22
Lung	88	0.5823	0.23
Spleen	83	0.5733	0.62
Salivary gland	107	0.5641	0.65
Medial olfactory epithelium	120	0.5580	0.83
Mammary gland (lact)	62	0.5681	1.9
Vomeralnasal organ	63	0.5575	3.4
B220+ B-cells	163	0.5327	4.5
Adrenal gland	58	0.5413	8.3
Prostate	57	0.5102	23.7
Thymus	55	0.4692	47.1
Embryo day 6.5	68	0.4239	59.1

A description of each dataset, the number of promoter sequences it contains, the average AUC value of the ROC curves obtained from the 10 cross-validation runs, and a corrected *P*-value for this value is shown. Tissues are ranked in order of increasing *P* values and decreasing AUC values (PB: peripheral blood; BM: bone marrow).

Table 2. Overview of the top five structural patterns in the cross-validation run with highest performance for the human fetal liver dataset

Pattern rank	Pattern content	Pattern weight
1	Motif 1 in region −100 to +200 relative to TSS	0.015
2	Motif 2 in region −250 to +200 relative to Motif 3	0.012
3	Motif 4 in region −350 to +50 relative to Motif 5	0.011
4	Motif 2 in region −200 to +150 relative to TSS	0.010
5	Motif 6 in region −150 to +200 relative to TSS	0.010

The rank, content and weight of each pattern is indicated. Motif IDs refer to the motif logos shown in Figure 1.

Table 3. Overview of the 10 tissue pairs with the highest inter-tissue performance for the human models

Model tissue	Sequence tissue	AUC	Correlation of expression
Kidney	Fetal liver	0.7305	0.21
Liver	Kidney	0.6939	0.38
Pancreas	Fetal liver	0.6876	0.16
Liver	Fetal liver	0.6870	0.29
Skeletal muscle	Tongue	0.6846	0.35
Kidney	Tongue	0.6806	0.11
PB- CD14+ monocytes	Lung	0.6542	0.03
Pancreas	Tongue	0.6497	0.08
PB- CD14+ monocytes	Tongue	0.6460	−0.20
Liver	Tongue	0.6431	0.11

The tissue on which the model was trained, and the tissue on which it was applied are shown, along with an AUC value as measure of performance. The value shown is an average of the 10 cross-validation runs. The final column shows the Pearson correlation coefficient of the expression of the genomic set of genes for the model tissue and target sequences tissue.

that performance greatly varies between the tissues: AUC values averaged over the 10 CV runs range from as high as 0.81 for the human tongue model down to values lower than expected at random for the mouse embryo day 6.5 model (a random scoring process has an expected average AUC value of 0.5). After Bonferroni correction for multiple testing (60 tissues), for 35 out of 60 sets the average AUC values were significantly higher than expected at random (Bonferroni-corrected p-value < 0.01, as determined by random shuffling of scores between positive and negative test sequences). These include 14 human and 21 mouse sets. Likewise, 28 sets have significantly higher average sensitivity at 95% specificity than expected at random (14 human and 14 mouse sets), and 32 sets have significantly higher average sensitivity at 90% specificity than expected at random (16 human and 16 mouse sets). For 24 out of 60 sets, all three measures of performance were significantly higher than expected (10 human sets and 14 mouse sets—see Tables 3 and 4 in the Supplementary Data for more data on the performances).

In the following sections we will shortly discuss the content of some models and their performances.

Table 4. Overview of the inter-species performance of some models

Human model	Mouse sequences	AUC	Sensitivity at 95% specificity	Sensitivity at 90% scpecificity
Tongue	Tongue	0.6899	18.7	28.6
Liver	Liver	0.6694	18.0	29.1
Kidney	Kidney	0.6321	14.2	22.9
Heart	Heart	0.6221	15.3	24.1
Skeletal muscle	Skeletal muscle	0.6064	15.2	22.3
Lung	Lung	0.5997	12.3	19.4
Pancreas	Pancreas	0.5803	9.5	17.5
Testis	Testis	0.5436	8.6	15.1
Placenta	Placenta	0.5135	7.2	12.9

Models are trained on human datasets and applied on mouse datasets. The tissue on which the model was trained, and the tissue it was applied on are shown, along with three measures of performance. Ten randomly selected pairs of sets gave the following values for the measures of performance (average ± SD): AUC: 0.529 ± 0.055; sensitivity at 95% specificity: 6.6 ± 3.7; Sensitivity at 90% specificity: 12.7 ± 5.2.

Fetal liver promoter structure models

Liver is a tissue that is often used in the study of tissue-specific regulation of transcription (32,37). In this study too, the liver promoter models perform relatively well. One advantage of our model is that more important structural patterns are assigned higher weights by the GA during training, allowing for easy identification of biologically important structural features. Table 2 and Figure 1 show the top five structural patterns we found for the best performing CV run of the human fetal liver dataset, and the corresponding sequence motifs. The fetal liver promoter model is dominated by a small set of known TFs: HNF1, and to a lesser extent HNF3, HNF4, C/EBP and GATA3. The importance of HNF1, HNF3, HNF4 and C/EBP in liver-specific gene expression regulation has been well reported (38,39), and other computational studies have found similar results (37,40,41). In eight of the 10 CV runs, a rule on the positioning of HNF1 in the region around −200 to +150 is present in the top five rules, with remarkably little variation in the region depicted by the rule. Similar rules were found to have high weights in the human liver promoter model and the mouse liver promoter model, illustrating their validity. In Table 2, in addition to the positioning of HNF1 to the TSS, also a relative positioning rule involving HNF1 and a motif resembling HNF1 (COMP1) in its close proximity is shown. This might indicate the presence of strong and weak HNF1 binding sites in close proximity of each other. A recent study using a thermodynamic approach for the analysis of regulatory regions reported that a combination of a strong and a weak site might provide a more sensitive regulation than a combination of two strong sites or two weak sites (42). Recently other studies too, have stressed the importance of weak binding sites (7,43).

We refer to the Supplementary Data for a description of human and mouse liver, human skeletal muscle and human tongue promoter models. The top patterns in these tissues illustrate the general nature of our model: they contain patterns on the positioning of CRMs with regard to the TSS, patterns describing the distal and proximal positioning of pairs of motifs, and patterns on the unrestricted presence of sites anywhere in the promoter sequences. Cooperating TFs or clustering of sites in CRMs can be represented by patterns describing proximal positioning of pairs of sites. Distal interactions between TFs through the formation of loops in the DNA sequence can be modeled by patterns describing distal positioning.

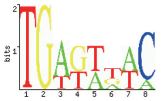
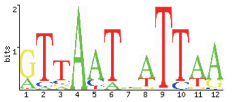
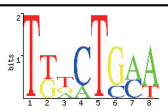
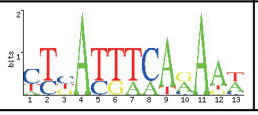
Motif ID	Motif Logo	Comment
Motif 1		unknown motif
Motif 2		HNF1 (E: 6.4E-6)
Motif 3		NF-Y (E: 0.85)
Motif 4		COMP1 (E: 0.069)
Motif 5		unknown motif
Motif 6		C/EBPgamma (TRANSFAC PWM)

Figure 1. The motif logos of the sequence motifs present in the top five structural patterns of the human fetal liver model as shown in Table 2. The motif IDs correspond to the IDs used in Table 2. For each motif, the motif logo and a comment are shown. Motif logos labeled as 'unknown motif' do not show a significant resemblance to any PMW in the Transfac database. For logos similar to known motifs, the known motif is indicated together with the tomtom *E*-value. 'TRANSFAC PWM' indicates that the motif used corresponds directly to a PWM from the Transfac database.

Inter-tissue performance

An additional illustration of the validity of the promoter models was found when applying models of one tissue on the promoter sequences of genes expressed specifically in other tissues. We used the promoter models of each tissue to score the positive and control promoter sequence sets, and evaluated the ability of the model to distinguish between these two sets of sequences. We limited this analysis to datasets for which the model showed significant performance on its original test sequences (14 datasets for human, 21 for mouse). Promoter sequences labeled as specific for both the model tissue and the target tissue were excluded before the evaluation of performance between tissues, in order

to avoid sequences used as training samples for the model to be also used for the evaluation of performance.

We found that applying promoter structure models of one tissue on the sequences of a tissue with similar expression values [as measured by the Pearson correlation coefficient (PCC) between expression values of all genes in the model and target tissue] tend to result in higher performance (as measured by AUC values). The PCC between AUC values and correlation of expression values was 0.34 for human models ($P = 3.3e{-6}$), and 0.40 for mouse models ($P < 2.2e{-16}$). Table 3 shows the human tissue combinations with the highest performance (see also Table 9 in the Supplementary Data). While the average AUC value over all combinations of tissues was close to that expected at random (average AUC: 0.54; SD: 0.07), performance of some tissue combinations was significantly higher. Examples include the performance of the kidney model on fetal liver-specific promoters, the liver model on kidney promoters and the pancreas model on fetal liver promoters.

Another example of a pair of related tissues is tongue and skeletal muscle. The expression values of the genomic set of genes in these tissues show a relatively high correlation, with a PCC of 0.35 (average PCC over all human tissue pairs: 0.04; SD: 0.15). They are thus likely to share a common regulatory mechanism, and indeed the models of these two tissues show high mutual performance (AUC of the skeletal muscle model on tongue sequences: 0.6846). In addition, there is a high correlation between the scores of the genomic set of promoter sequences for both models (PCC: 0.63; average PCC over all human tissue pairs: 0.18; SD: 0.22), which indicates that these two models contain similar patterns. Nevertheless, top patterns of the two models do not show a clear similarity (see the Supplementary Data for a more detailed description of the top patterns of both tissues), and sites that are mainly responsible for tongue-specific expression are different from sites responsible for skeletal muscle expression.

For a number of high performing pairs of tissues (for example monocytes and lung) there is no apparent correlation between their expression values. A previous study on CRMs has lead to similar observations (13). We hypothesize that expression in these tissues is regulated by different TFs binding similar DNA motifs, or by similar TFs interacting with different cofactors. Many TFs binding similar motifs are known, and previous studies have found that promoter regions of genes specifically expressed in different tissues are still enriched for similar motifs. Smith et al. (44) found that nuclear receptor binding site motifs and E-box motifs were over-represented in the tissue-specific promoters of 11 and 10 of the 14 human and mouse tissues they investigated, respectively (44).

Similar results were obtained in mouse (see Table 10 in the Supplementary Data). For the mouse tissues especially, high performing tissue combinations show a tendency to have a high correlation of expression over the genomic set of genes. The top 10 performing combinations have an average PCC of 0.28 (average PCC over all mouse tissue combinations: 0.00; SD: 0.13).

Inter-species performance

In addition to inter-tissue similarities, we also found inter-species similarities. We used models with significant performance of one species and applied them on the promoter sequences of the same tissue in the other species. Table 4 shows the performance of some human promoter structure models on mouse datasets (see Table 11 in the Supplementary Data for the performance of mouse models on human datasets). We found that in many cases applying human models on the same tissue they were trained on in mouse lead to high performance. The reverse case was also true (mouse models on human datasets). This was not the case for randomly selected pairs of sets, which showed performances close to what can be expected at random (average AUC: 0.529; SD: 0.055). Models with high performance in one species—such as tongue and liver—tend to also have high performance when applied on the other species dataset. In general, mouse models applied on human datasets seem to have somewhat higher performance than human counterparts applied on mouse datasets. We suspect that this is caused by the higher quality of the human TSS annotations compared to those of mouse.

Easy and difficult tissues: where is the difference?

The promoter structures of some tissues are clearly easier to model than those of others. It is not surprising that tissues that are often used for computational analysis of tissue-specific regulation, such as skeletal muscle, liver and kidney, have a relatively high performance, and can thus be considered as 'easy' to model. In addition to that, a tendency for similar tissues to show high performance in both human and mouse is observed. Examples are tongue, skeletal muscle, liver and kidney models. There are a number of possible causes for the variance of performance between different tissues. One obvious cause might be that the model is capable of capturing important structural features for some of the tissues, but not for others. However, while there might be some additional structural patterns that our model cannot capture, it is unlikely that the regulation of transcription in different tissues is controlled by completely different types of molecules and interactions. We did, however, observe a tendency for high-scoring tissues to have a clear regulatory signal located in the region immediately upstream of and around the TSS. For each promoter model, the weights of patterns describing positioning of TFBSs relative to the TSS were mapped to the region described in each pattern. Thus we obtained a 'total weight' for each position in the promoter sequences. Figure 2 shows a visualization of the average weights for the five human models with the highest performance, and the five human models with the lowest performance (see also Supplementary Figures 10 and 11 for figures on all models). When comparing the weight distribution in high-performing models with that of low-performing models, we observed a clear pattern of higher importance around the TSS in high-performing models. Previous studies have reported the importance of this region in the regulation of tissue-specific expression (45) and transcriptional regulation in general (46).

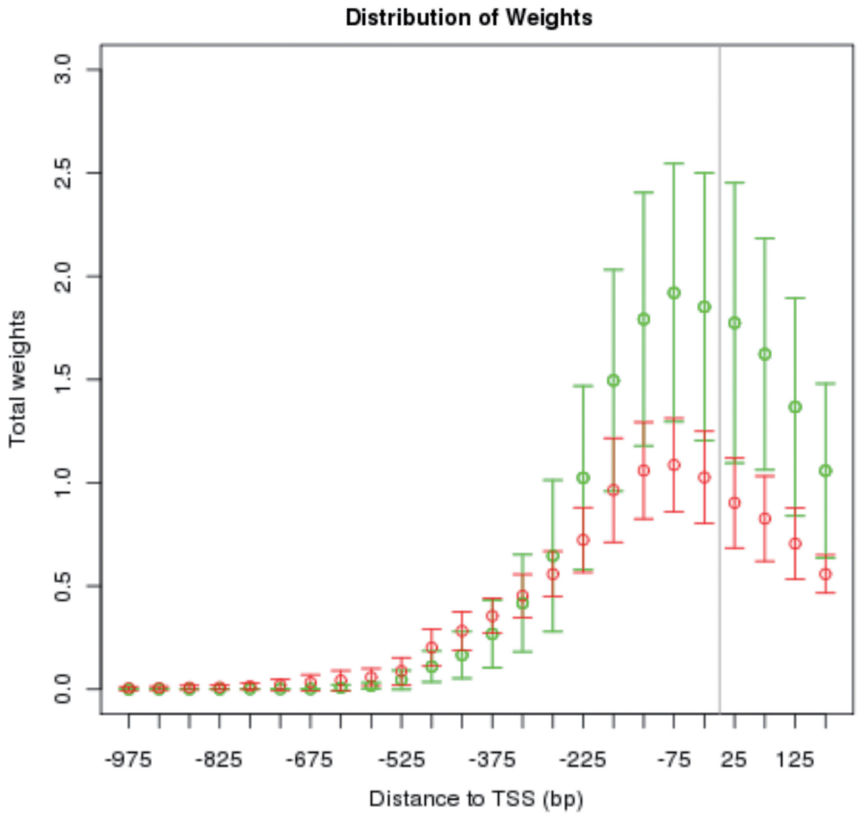


Figure 2. The distribution of the weights of patterns describing the positioning of regulatory motifs relative to the TSS. Average weights with error bars corresponding to the standard deviation are shown for the five human tissues with the highest performance (tongue, fetal liver, kidney, skeletal muscle and liver, indicated in green), and for the five tissues with the lowest performance (testis leydig cell, adipocyte, PB-CD19+ B cells, BM-CD105+ endothelial cells and BM-DC34+, indicated in red). The vertical grey line indicates the position of the TSS.

Other genome-wide studies of regulatory regions have reported potential regulatory sites with positional preference for this region (4,47). In low-performing models, this pattern was far less pronounced. This observation indicates that tissues with a clear regulatory signal proximal to the TSS show a tendency to be easier to model. An additional interpretation is that in tissues that are difficult to model, this proximal regulatory signal is absent, or situated in other regions. Studies have found that regulatory modules containing binding sites for certain TFs might target proximal regions, while others show a preference for more distal regions (14,15). Regions responsible for the regulation of expression in such tissues might be located further upstream (>1 kb upstream), causing them to be missed by our model.

Another factor likely to play a role in the differences in performance might be the heterogeneity of tissues. The same set of genes might be under the control of different regulatory factors in different cell types, resulting in the binding sites for different TFs not to be significantly over-represented in the entire datasets. We did indeed observe a tendency for high performing tissues to have one or more clearly over-represented known TFBS motifs, while tissues with low performance often had no significantly over-represented known TFBS motifs. We also found that models for tissues such as liver and skeletal muscle, which are believed to be relatively homogeneous, achieved high performance. However, it is difficult to conclude that homogeneous tissues are easier to model than others, because of on one hand the difficulty of measuring the degree of homogeneity of tissues and on the other hand the fact that we do not know the precise composition of the tissue samples used in the study of Su *et al.* (17). Another level of heterogeneity is the existence of different regulatory pathways controlling expression of subsets of genes within the same cell type or tissue, which might make it more difficult for the model to capture the relatively more diverse structural patterns. Finally, experimental contaminations might be a source for differences in performance. This might be especially the case for tissues where contamination by blood cells is likely to occur, such as lung and thymus. Some results of a recent genome-wide study on human TFs have lead to similar suggestions (48).

CONCLUSION

We have presented a promoter architecture model and its performance on sets of genes that show increased expression in a 26 human and 34 mouse tissues. Our results show that performance of our model is greatly variable between

the different tissues and cell types. In 35 out of the 60 datasets (58%), our model performs significantly better than a random scoring function. The high performance in some tissues shows that structural patterns in human and mouse promoter sequences can be used to distinguish genes expressed specifically in a certain tissue from genes that are not. Examination of the models in some tissues with high performance revealed that many patterns with high weights contain structural information on known regulators of importance. In addition, application of models between related tissues, and between human and mouse indicate that the model is capable of picking up biologically meaningful structural information. No species-specific prior knowledge was used in the training of the model, and all tissues were addressed with the exact same approach. We thus believe that the model is applicable on other species and tissues as well, even though differences in performance depending on the tissue of interest are likely to appear. Where successful, our model can help us to understand the mechanisms of transcription regulation, and can provide hypothesis for conducting wet experiments.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank all members of the Nakai–Kinoshita Laboratory for helpful discussions and advice. Computational time was provided by the Super Computer System at the Human Genome Center, Institute of Medical Science, The University of Tokyo.

FUNDING

Global COE Program (Center of Education and Research for Advanced Genome-Based Medicine), MEXT, Japan. Funding for open access charge: Japan Science and Technology Agency (JST). A.V. is supported by the Japanese Government Scholarship (Monbukagakusho, MEXT). Funding for open access charge: Institute for Bioinformatics Research and Development (BIRD), Japan Science and Technology Agency (JST), Japan.

Conflict of interest statement. None declared.

REFERENCES

1. Tompa,M., Li,N., Bailey,T.L., Church,G.M., De Moor,B., Eskin,E., Favorov,A.V., Frith,M.C., Fu,Y., Kent,W.J. *et al.* (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, **23**, 137–144.
2. Stormo,G.D., Schneider,T.D. and Gold,L.M. (1982) Characterization of translational initiation sites in *E. coli*. *Nucleic Acids Res.*, **10**, 2971–2996.
3. Nelander,S., Larsson,E., Kristiansson,E., Mansson,R., Nerman,O., Sigvardsson,M., Mostad,P. and Lindahl,P. (2005) Predictive screening for regulators of conserved functional gene modules (gene batteries) in mammals. *BMC Genomics*, **6**, 68.
4. Xie,X., Lu,J., Kulbokas,E.J., Golub,T.R., Mootha,V., Lindblad-Toh,K., Lander,E.S. and Kellis,M. (2005) Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, **434**, 338–345.
5. Arnone,M.I. and Davidson,E.H. (1997) The hardwiring of development: organization and function of genomic regulatory systems. *Development*, **124**, 1851–1864.
6. Bussemaker,H.J., Li,H. and Siggia,E.D. (2001) Regulatory element detection using correlation with expression. *Nat. Genet.*, **27**, 167–171.
7. Segal,E., Raveh-Sadka,T., Schroeder,M., Unnerstall,U. and Gaul,U. (2008) Predicting expression patterns from regulatory sequence in Drosophila segmentation. *Nature*, **451**, 535–540.
8. Beer,M.A. and Tavazoie,S. (2004) Predicting gene expression from sequence. *Cell*, **117**, 185–198.
9. Yuan,Y., Guo,L., Shen,L. and Liu,J.S. (2007) Predicting gene expression from sequence: a reexamination. *PLoS Comput. Biol.*, **3**, e243.
10. Noto,K. and Craven,M. (2007) Learning probabilistic models of cis-regulatory modules that represent logical and spatial aspects. *Bioinformatics*, **23**, e156–e162.
11. Won,K.J., Sandelin,A., Marstrand,T.T. and Krogh,A. (2008) Modeling promoter grammars with evolving hidden Markov models. *Bioinformatics*, **24**, 1669–1675.
12. Vandenbon,A., Miyamoto,Y., Takimoto,N., Kusakabe,T. and Nakai,K. (2008) Markov chain-based promoter structure modeling for tissue-specific expression pattern prediction. *DNA Res.*, **15**, 3–11.
13. Smith,A.D., Sumazin,P., Xuan,Z. and Zhang,M.Q. (2006) DNA motifs in human and mouse proximal promoters predict tissue-specific expression. *Proc. Natl Acad. Sci. USA*, **103**, 6275–6280.
14. Blanchette,M., Bataille,A.R., Chen,X., Poitras,C., Laganiere,J., Lefebvre,C., Deblois,G., Giguere,V., Ferretti,V., Bergeron,D. *et al.* (2006) Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res.*, **16**, 656–668.
15. Van Loo,P., Aerts,S., Thienpont,B., De Moor,B., Moreau,Y. and Marynen,P. (2008) ModuleMiner – improved computational detection of cis-regulatory modules: are there different modes of gene regulation in embryonic development and adult tissues? *Genome Biol.*, **9**, R66.
16. Vandenbon,A. and Nakai,K. (2008) Using simple rules on presence and positioning of motifs for promoter structure modeling and tissue-specific expression prediction. *Genome Informatics*, **21**, 188–199.
17. Su,A.I., Wiltshire,T., Batalov,S., Lapp,H., Ching,K.A., Block,D., Zhang,J., Soden,R., Hayakawa,M., Kreiman,G. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA*, **101**, 6062–6067.
18. Wakaguri,H., Yamashita,R., Suzuki,Y., Sugano,S. and Nakai,K. (2008) DBTSS: database of transcription start sites. *Progress Report 2008. Nucleic Acids Res.*, **36**, D97–D101.
19. Carninci,P., Kasukawa,T., Katayama,S., Gough,J., Frith,M.C., Maeda,N., Oyama,R., Ravasi,T., Lenhard,B., Wells,C. *et al.* (2005) The transcriptional landscape of the mammalian genome. *Science*, **309**, 1559–1563.
20. Karolchik,D., Kuhn,R.M., Baertsch,R., Barber,G.P., Clawson,H., Diekhans,M., Giardine,B., Harte,R.A., Hinrichs,A.S., Hsu,F. *et al.* (2008) The UCSC genome browser database: 2008 update. *Nucleic Acids Res.*, **36**, D773–D779.
21. Matys,V., Kel-Margoulis,O.V., Fricke,E., Liebich,I., Land,S., Barre-Dirrie,A., Reuter,I., Chekmenev,D., Krull,M., Hornischer,K. *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108–D110.
22. Gupta,S., Stamatoyannopoulos,J.A., Bailey,T.L. and Noble,W.S. (2007) Quantifying similarity between motifs. *Genome Biol.*, **8**, R24.
23. Johnson,J.M. and RahmatSamii,Y. (1997) Genetic algorithms in engineering electromagnetics. *IEEE Anten. & Prop. Mag.*, **39**, 7–25.
24. Gagne,C., Schoenauer,M., Parizeau,M. and Tomassini,M. (2006) Genetic programming, validation sets, and parsimony pressure. *Genetic Program. Proc.*, **3905**, 109–120.
25. Chen,H., Ray-Gallet,D., Zhang,P., Hetherington,C.J., Gonzalez,D.A., Zhang,D.E., Moreau-Gachelin,F. and Tenen,D.G.

(1995) PU.1 (Spi-1) autoregulates its expression in myeloid cells. *Oncogene*, **11**, 1549–1560.
26. Anderson,K.L., Smith,K.A., Conners,K., McKercher,S.R., Maki,R.A. and Torbett,B.E. (1998) Myeloid development is selectively disrupted in PU.1 null mice. *Blood*, **91**, 3702–3710.
27. Barton,K., Muthusamy,N., Fischer,C., Ting,C.N., Walunas,T.L., Lanier,L.L. and Leiden,J.M. (1998) The Ets-1 transcription factor is required for the development of natural killer cells in mice. *Immunity*, **9**, 555–563.
28. Eyquem,S., Chemin,K., Fasseu,M., Chopin,M., Sigaux,F., Cumano,A. and Bories,J.C. (2004) The development of early and mature B cells is impaired in mice deficient for the Ets-1 transcription factor. *Eur. J. Immunol.*, **34**, 3187–3196.
29. Wang,D., John,S.A., Clements,J.L., Percy,D.H., Barton,K.P. and Garrett-Sinha,L.A. (2005) Ets-1 deficiency leads to altered B cell differentiation, hyperresponsiveness to TLR9 and autoimmune disease. *Int. Immunol.*, **17**, 1179–1191.
30. Anderson,M.K., Hernandez-Hoyos,G., Diamond,R.A. and Rothenberg,E.V. (1999) Precise developmental regulation of Ets family transcription factors during specification and commitment to the T cell lineage. *Development*, **126**, 3131–3148.
31. Hayhurst,G.P., Lee,Y.H., Lambert,G., Ward,J.M. and Gonzalez,F.J. (2001) Hepatocyte nuclear factor 4alpha (nuclear receptor 2A1) is essential for maintenance of hepatic gene expression and lipid homeostasis. *Mol. Cell Biol.*, **21**, 1393–1403.
32. Odom,D.T., Zizlsperger,N., Gordon,D.B., Bell,G.W., Rinaldi,N.J., Murray,H.L., Volkert,T.L., Schreiber,J., Rolfe,P.A., Gifford,D.K. *et al.* (2004) Control of pancreas and liver gene expression by HNF transcription factors. *Science*, **303**, 1378–1381.
33. Cereghini,S. (1996) Liver-enriched transcription factors and hepatocyte differentiation. *FASEB J.*, **10**, 267–282.
34. Black,B.L. and Olson,E.N. (1998) Transcriptional control of muscle development by myocyte enhancer factor-2 (MEF2) proteins. *Annu. Rev. Cell Dev. Biol.*, **14**, 167–196.
35. Yu,Y.T., Breitbart,R.E., Smoot,L.B., Lee,Y., Mahdavi,V. and Nadal-Ginard,B. (1992) Human myocyte-specific enhancer factor 2 comprises a group of tissue-restricted MADS box transcription factors. *Genes Dev.*, **6**, 1783–1798.
36. Shore,P. and Sharrocks,A.D. (1995) The MADS-box family of transcription factors. *Eur. J. Biochem.*, **229**, 1–13.
37. Krivan,W. and Wasserman,W.W. (2001) A predictive model for regulatory sequences directing liver-specific transcription. *Genome Res.*, **11**, 1559–1566.
38. Ktistaki,E. and Talianidis,I. (1997) Modulation of hepatic gene expression by hepatocyte nuclear factor 1. *Science*, **277**, 109–112.
39. Tronche,F., Ringeisen,F., Blumenfeld,M., Yaniv,M. and Pontoglio,M. (1997) Analysis of the distribution of binding sites for a tissue-specific transcription factor in the vertebrate genome. *J. Mol. Biol.*, **266**, 231–245.
40. Chen,X.Y. and Blanchette,M. (2007) Prediction of tissue-specific cis-regulatory modules using Bayesian networks and regression trees. *BMC Bioinformatics*, **8(Suppl. 10)**, S2.
41. Pennacchio,L.A., Loots,G.G., Nobrega,M.A. and Ovcharenko,I. (2007) Predicting tissue-specific enhancers in the human genome. *Genome Res.*, **17**, 201–211.
42. Gertz,J., Siggia,E.D. and Cohen,B.A. (2009) Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature*, **457**, 215–218.
43. Tanay,A. (2006) Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Res.*, **16**, 962–972.
44. Smith,A.D., Sumazin,P. and Zhang,M.Q. (2007) Tissue-specific regulatory elements in mammalian promoters. *Mol. Syst. Biol.*, **3**, 73.
45. Schug,J., Schuller,W.P., Kappen,C., Salbaum,J.M., Bucan,M. and Stoeckert,C.J. (2005) Promoter features related to tissue specificity as measured by Shannon entropy. *Genome Biol.*, **6**, R33.
46. Trinklein,N.D., Aldred,S.J.F., Saldanha,A.J. and Myers,R.M. (2003) Identification and functional analysis of human transcriptional promoters. *Genome Res.*, **13**, 308–312.
47. Vardhanabhuti,S., Wang,J. and Hannenhalli,S. (2007) Position and distance specificity are important determinants of cis-regulatory motifs in addition to evolutionary conservation. *Nucleic Acids Res.*, **35**, 3203–3213.
48. Vaquerizas,J.M., Kummerfeld,S.K., Teichmann,S.A. and Luscombe,N.M. (2009) A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.*, **10**, 252–263.