Published online 26 October 2006 *Nucleic Acids Research, 2006, Vol. 34, No. 20* ***5943–5950***
doi:10.1093/nar/gkl608

Identification of core promoter modules in *Drosophila* and their application in accurate transcription start site prediction

Uwe Ohler[1,2]

Institute for Genome Sciences and Policy, [1]Department of Biostatistics and Bioinformatics and [2]Department of Computer Science Duke University, Durham, NC 27708, USA

Received June 21, 2006; Revised August 2, 2006; Accepted August 3, 2006

ABSTRACT

The reliable recognition of eukaryotic RNA polymerase II core promoters, and the associated transcription start sites (TSSs) of genes, has been an ongoing challenge for computational biology. High throughput experimental methods such as tiling arrays or 5′ SAGE/EST sequencing have recently lead to much larger datasets of core promoters, and to the assessment that the well-known core promoter sequence elements such as the TATA box appear to be much less frequent than thought. Here, we address the co-occurrence of several previously identified core promoter sequence motifs in *Drosophila melanogaster* to determine frequently occurring core promoter modules. We then use this in a new strategy to model core promoters as a set of alternative submodels for different core promoter architectures reflecting these different motif modules. We show that this system improves greatly on computational promoter recognition and leads to highly accurate *in silico* TSS prediction. Our results indicate that at least for the case of the fruit fly, we are getting closer to an understanding of how the beginning of a gene is defined in a eukaryotic genome.

INTRODUCTION

The concerted and differentiated expression of genes is necessary for the existence of complex eukaryotic organisms with an intricate development that requires precise control of the expression of information. Understanding the regulation of gene expression is therefore one of the most interesting challenges in molecular biology today. We are only beginning to elucidate the impressive logic and organization of tightly interwoven players that a cell uses to determine the active state of every component in it (1,2). One of the most important control levels is the first step of gene expression, the transcription of a gene into messenger-RNAs. Protein-coding genes, as well as some classes of regulatory non-coding genes (in particular miRNAs), are transcribed by RNA polymerase II (pol-II). The transcriptional control region of pol-II regulated genes encompasses a core promoter, a proximal promoter region, and possibly distal enhancers, all of which contain transcription factor binding sites (TFBS) (3).

Core promoters are responsible for guiding the polymerase to the correct transcription start site (TSS) and span the region from [−50, +50] relative to the TSS. Accurate initiation of transcription depends on assembling a complex containing pol-II and at least six general transcription factors (4,5), the most important of which is TFIID. TFIID is a protein complex which consists of the TATA binding protein (TBP) and at least a dozen other components known as TAFs (TBP associated factors) which also interact, directly or indirectly, with sequence elements in the core promoter. Comprehensive analyses of fly core promoters (6,7) as well as plants and mammals (8,9) have suggested that the well-known TATA box, which is located ∼25 nt upstream of the TSS, occurs in at most 25–30% of the genes within a genome, despite its conservation in all eukaryotes. However, so-called TBP related factors (TRFs) may substitute for TBP in the TFIID complex and contribute to the activation of specific subsets of genes. TRF1 binds to TATA-box like sequences in *Drosophila*; TRF2 is present in vertebrates as well but does not show sequence specific interactions. Instead, a study on fly genes showed that it occurs in a complex with the DNA replication element binding factor [DRE factor (10)]. Compared to other core promoter sequence elements, its location appears to be much less restricted relative to the start site.

In *Drosophila* as well as vertebrates, the sequence at the TSS is conserved and referred to as the Initiator motif (Inr). Sequences downstream of the TSS were also found to exert influence on basal transcription activity. Experimental evidence for a specific downstream promoter element DPE (6) suggests that the DPE is as widely used as the TATA box. Its core motif is located exactly from 28 to 33 bp downstream of the TSS and is recognized by two distinct TAFs. A second

Tel: +1 919 668 5388; Fax: +1 919 668 0795; Email: uwe.ohler@duke.edu

distinct downstream element called MTE (Motif 10 Element), located at positions 17–22, has been computationally predicted and experimentally verified (7,11), and is also thought to interact with parts of TFIID. Despite evidence for downstream vertebrate elements, current knowledge suggests that DPE and MTE may play a less important role in these organisms. However, a different downstream element called DCE has been found in human promoters (12). Thus, elements show organism specific patterns: the initiator has higher information content, and the DPE is much more frequent in fly promoters compared to mammals. In addition, the majority of mammalian promoters are associated with so-called CpG islands—regions of atypically high density of CG dinucleotides—a phenomenon which does not exist in many other species, including *Drosophila*.

An important area of computational sequence analysis is concerned with the analysis and identification of regulatory DNA (13). Recent computational methods for core promoter prediction are based on solid machine learning techniques like probabilistic sequence models or support vector machines, and have shown good performance on fly predictions (7,14). As mentioned above, species-specific distinctions in promoter motifs make it necessary to adapt systems to particular species. An evaluation on the human genome (15) showed that vertebrate systems have also recently improved, but most are still hampered by a strong bias towards CpG island promoters, as well as our lesser understanding of core promoter motifs when compared to *Drosophila*. A recent publication has explicitly modeled mammalian promoters not associated with CpG islands (16), and reports promising results by combining positional preference and co-occurrence of oligomers, yet without reporting distinct new motifs which may serve as targets of the general transcription machinery. In this paper, we will further investigate the architecture of core promoters in the *Drosophila melanogaster* genome, define a set of core promoter modules, and show that making use of our increasing knowledge of core promoter modules in computational systems leads to an accurate prediction of core promoters/TSSs.

METHODS

McPromoter: TSS prediction with probabilistic sequence models

A hidden Markov model [HMM; (17)] is a state-based generative model which transitions stochastically from state to state, emitting a single feature (here, a symbol representing a nucleotide in the DNA) from each state according to that state's emission probabilities. A generalized HMM [gHMM; also referred to as semi-Markov model (18)] extends this process to emitting a sequence of consecutive features, or segments, from each state. A state now incorporates a joint emission distribution for a sequence of features, and a discrete length distribution on the size of the sequence. States can e.g. represent short sequence motifs (such as the TATA box), sequence fragments of a particular size but no discernible motif (the spacer between TATA box and initiator), or the general base composition. Efficient algorithms exist to calculate the best division of a sequence into these segments during training and evaluation.

The McPromoter system, which we have developed previously (7,19), consisted of one linear gHMM for a promoter sequence, with states representing different segments of the promoter from −250 to +50: upstream 1 and 2, TATA box, spacer, initiator, and downstream. In addition to the core promoter *per se*, our promoter model incorporates states representing the proximal promoter region up to −250 because it is characterized by a distinct profile in the GC content. Interpolated Markov chains were used as emission distributions, and histograms bounded by a minimum and maximum value as length distributions. The Viterbi algorithm was used to compute the likelihood of the most probable segmentation of a sequence and the associated state path. We also employed the Viterbi algorithm during training: After initialization on overlapping adjacent subsequences, both emission and length distributions were iteratively re-estimated according to the most probable segmentations of the training data. Non-promoter sequences were represented by two interpolated Markov chain models, trained on coding and non-coding/intronic genomic sequences. To accurately reflect the potential presence of a gene on both sense and anti-sense strand, both non-promoter models were mixtures with two components trained on sense or anti-sense sequences, respectively. To localize promoters in genomic sequences, an input window of 300 bp was shifted along both sides of the sequence, and the log likelihood ratio score of promoter versus best non-promoter model, as well as the position of the initiator state, was stored for each window. Local maxima were reported as TSS if they exceeded a pre-set threshold on the score. We had previously augmented the sequence features with features for DNA physicochemical properties such as bendability (20); however, the additional benefits of these became negligible as available training data became larger (7).

In this study, we compared the performance of this previous system with one model against a new system with parallel gHMMs representing different core promoter motif modules. To establish these gHMMs, we performed a semi-supervised clustering on the training set of experimentally inferred core promoters in the following way:

(i) Initial data partitioning. We obtained initial overlapping partitions of the data based on strong hits to known core promoter motifs/modules.
(ii) Model initialization. We specified several gHMMs representing the different core promoter modules and trained each model on the respective initial partition of sequences. The topology of each model was motivated by the number of promoter motifs in the partition (one or two) as well as the approximate location of the motifs relative to the start site and each other, and the number of states varied accordingly between five and seven.
(iii) Clustering. We performed three iterations of an expectation-maximization-style semi-supervised clustering.
  (a) Re-classify all promoters in the data set according to the model yielding the highest likelihood. If the number of sequences for one class falls below a pre-specified cutoff, eliminate its model from the set of models and assign its sequences to the class with the second highest likelihood.

(b) Re-estimate the model parameters according to the new classification. This was done in a 5-fold cross-validation setup, i.e. five promoter models were trained on four-fifths of the training data each and evaluated on the fifth of sequences which were not part of the training data.

To localize promoters in genomic sequences, the score of each genomic window was computed as the log likelihood ratio score of the best promoter versus best non-promoter model, where the non-promoter models were identical to the ones described above. In this way, we were able make predictions about the type of core promoter along with its location.

Data sets

To allow for an unbiased comparison of our new approach with previous ones, we used comprehensive training data which we have previously described (7), in particular, a set of 1941 promoters (covering the region of [−250, +50] with respect to the TSS), inferred from clusters of aligned cap-trapped ESTs. A set of 1864 non-redundant sequences was prepared by removing highly similar sequences in this set according to the standards used by the Eukaryotic Promoter Database (21). As genomic test data, we used a benchmark set of 92 TSSs located in the 2.9 Mb Adh genomic fragment (22). We removed an additional 31 sequences from the training set which were located in this region.

A promoter was considered as containing a specific core promoter motif if it had a match to one of the core promoter weight matrices given in (7), with a significance level of $P \leqslant 0.001$ as computed by patser (23). This was used to obtain the initial partition of the training data which is then clustered, and to assess the co-occurrence of pairs of motifs. Numbers referring to the frequency of particular motifs are therefore conservative estimates.

Evaluation

We evaluated the performance of the system in two different settings, as described previously. First, the accuracy of the classifier was determined by cross-validation on the promoter and non-promoter sequence data sets of 300 bp long sequences (20). We evaluated the accuracy in terms of average equal recognition rate (i.e. the threshold for which the rate of true positives equals the rate of true negatives), cross correlation, and receiver operating characteristics (ROC) integral. Second, the classifier was tested on the task of genome annotation, where we evaluated the success on locating TSSs in genomic DNA (22). TSSs were counted as true positive if they fall into a region of [−500, +50] around the annotated TSS of the test genes. Predictions outside this region and overlapping the test genes were counted as false positives.

RESULTS

In a previous study, we selected experimentally validated Drosophila TSSs from alignments of cap-trapped ESTs, by determining the 5′ end of only the most upstream entry within a cluster of overlapping ESTs if it fulfilled certain quality criteria (7). An analysis to identify over-represented motifs within this large set of ∼2000 core promoters surrounding the TSSs lead to a list of 10 significant sequence elements.

Briefly summarizing this result, the list included the known TATA box (Motif 3), initiator (Inr, Motif 4), and downstream promoter element (DPE, Motif 9). These elements showed a preference for a specific distance to the TSS: TATA boxes for the −25, Inr for the +1, and DPE for the +30 region. In addition to these three well known and positionally restricted sequence elements, another motif (named Motif 10 Element or MTE after its position in the list of enriched motifs) showed a strong preference for a downstream location immediately next to DPE. We have subsequently shown that motif 10 is a new functional core promoter element, located between the Inr and DPE elements at position +20 (11). Motif 2 corresponded to the DNA replication element DRE, which had at the same time been shown to interact with the TBP replacing factor TRF2 (which itself does not bind specifically to DNA) (10). We confirmed the role of this motif as a frequent promoter element not only in DNA replication related genes which it was named after. The other five elements are still unconfirmed as functional elements, including somewhat surprisingly the strongest motif (Motif 1; M1), and Motif 6 (M6) which loosely resembles a variation of the TATA box. In the following, we will not consider the remaining three motifs from our previous analysis: Motif 5 is a consensus E box motif, the typical binding site of transcription factors with basic helix–loop–helix DNA-binding protein domains which is enriched in core promoters (24) but likely not a direct interaction partner of the basal machinery. Motifs 7 and 8 do not show a location preference and may be artifacts of the motif finding algorithm, reflecting the sequence composition rather than particular motifs.

Establishing subclasses of promoters

Several core motifs have been known to work in cooperation, most notably the initiator motif paired with either TATA box, DPE, or MTE (6,11). When looking at motif pairs occurring preferentially in the same promoter (Table 1), the most prominent cases include combinations of the Inr with downstream elements (DPE and MTE) and also with the TATA box, as well as the combination of yet uncharacterized motifs 1 and 6. We examined the location preference of the motif 1/motif 6 (M1/6) module in detail: Despite a rather weak location preference of each individual motif (Figure 1A and B), the distinct preferred distance between the two elements is striking (Figure 1C). Given the low co-occurrence of motif 1 with the Inr and motif 6 with the TATA box (Table 1), and the similar spacing preference, this new combination may serve as an alternative to the TATA/Inr module. In an independent computational study, M1/6 was also found to be a highly significant core module (14). The observed co-occurrence pattern of these core promoter motifs strongly indicates a scenario in which at least four frequent modules of core promoter motifs exist, and where the widely studied TATA/initiator pair is simply one of several options. This idea of different core promoter architectures was initially posed for Drosophila in conjunction with the DPE element (6), and has been recently put forward for plants and vertebrates as well (25,26).

Table 1. Frequency of occurrence of individual core promoter motifs and pairs of motifs, modified after (7)

Motif X	% Seq with Motif X	% Of promoters with Motif X also containing Motif Y below						
		M1	DRE	TATA	INR	M6	DPE	MTE
M1	25.1	100.0	21.3	13.1	12.7	***28.3***	4.9	6.1
DRE	26.0	20.6	100.0	14.9	16.8	14.1	5.7	6.9
TATA	19.3	17.1	20.1	100.0	***28.9***	14.4	4.8	***9.4***
INR	26.3	12.1	16.6	***21.1***	100.0	12.1	***14.9***	***12.9***
M6	15.8	***45.1***	23.2	17.6	20.3	100.0	4.6	4.2
DPE	7.9	15.6	18.8	11.7	***49.4***	9.1	100.0	8.4
MTE	8.5	18.2	21.2	***21.2***	***40.0***	7.9	7.9	100.0

The first column gives the motif name. The second column shows the overall fraction of promoters containing a hit to the corresponding weight matrix model. The remaining columns list the frequency with which the motif in each row co-occurs with a particular second motif in the same core promoter. Cells which contain a higher fraction of promoters with a particular second motif than its overall frequency (column 2) are printed in bold and italic.

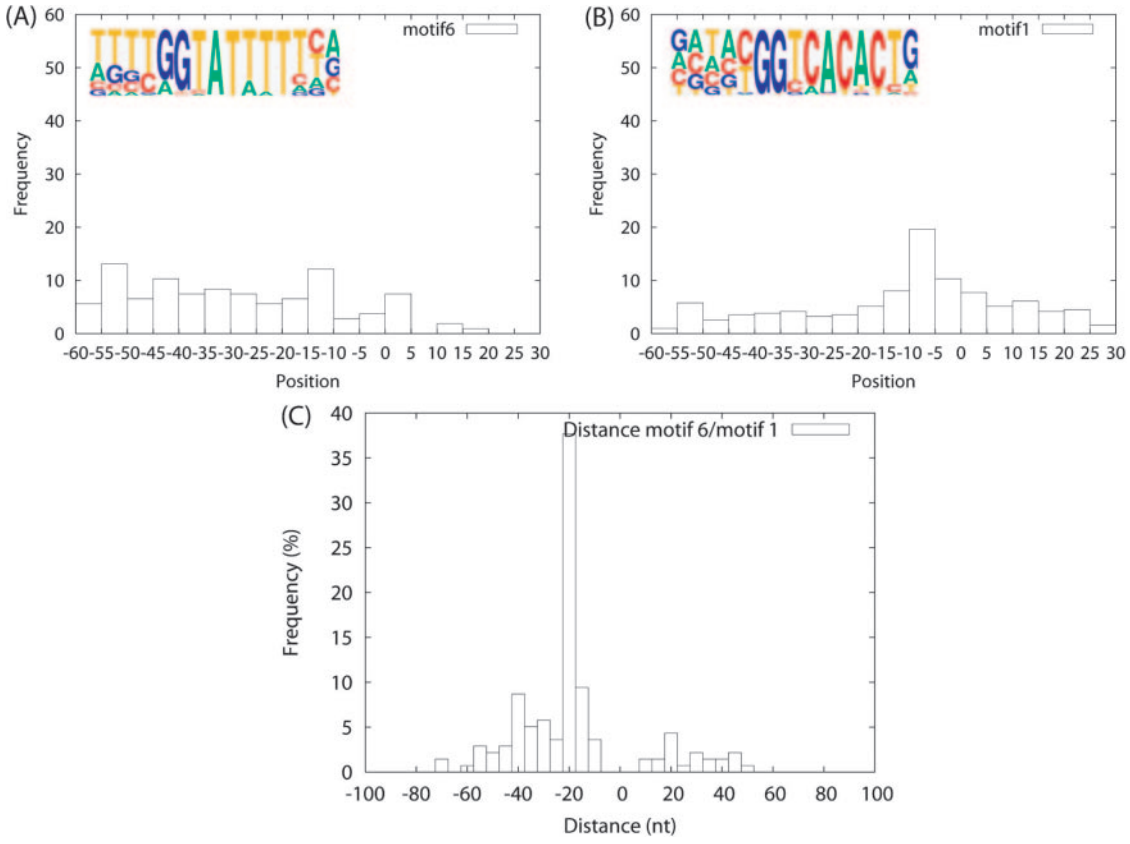

Figure 1. A new *Drosophila* core promoter module. (**A** and **B**) show the location distributions of motifs 6 and 1 relative to the TSS (pos. 0), and (**C**) shows the distance between motif occurrences in the same promoter.

In addition to the four pair modules, promoters with the experimentally proven Inr and DRE motifs alone were found to be frequent. We thus divided the set of experimentally inferred core promoters in six partitions (Figure 2), and specified generalized HMMs representing these different core promoter modules. The model topologies (i.e. the number of states per model and the minimum and maximum value of the state length distribution) were chosen as follows: Inr and DRE models: 5 states [(100,115), (105,115), (15,45), (5,15), (35,50)]; M1/6 and TATA/Inr models: 6 states [(100,115), (105,115), (5,15), (10,30), (5,15), (35, 50)]; Inr/DPE and Inr/MTE models: 7 states [(100,115), (105,115), (15,45), (5,15), (5,20), (5,15), (5,15)]. Next, we performed several iterations of semi-supervised clustering of the promoter training data. Figure 2 shows that the initial partitions generally proved to be stable and did not cluster together large fractions of sequences initially assigned to a different class. A notable exception was the Inr/MTE class, the sequences of which were gradually split up among the other classes and which was therefore removed previous to the last iteration. The DPE class was the one with the highest fraction of sequences initially assigned to a different cluster; this may be due to the DPE weight matrix we used, which extends beyond the DPE requirements as defined in Ref. (6) and may in fact be too stringent. Instead of one model averaging over all core promoters, we thus arrived at a set

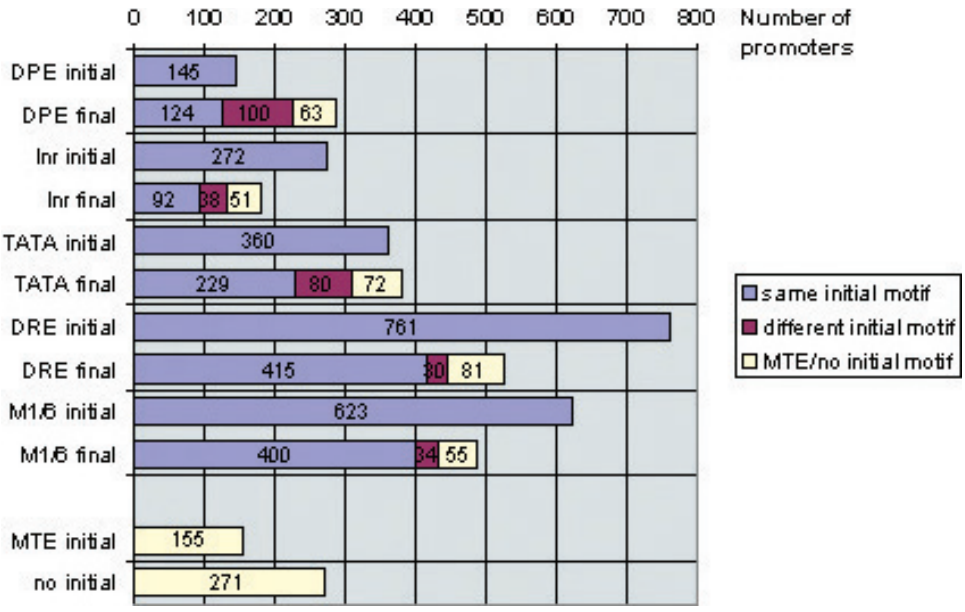

Figure 2. Comparison of motif module frequency in the initial and the final partitioning after semi-supervised clustering. Only initial frequencies are shown for the partition of sequences without a strong motif hit ('no initial'), i.e. which were initially not assigned to a particular motif class, and for the MTE motif partition, which proved to be not stable and was gradually split up among the other classes. For each of the final partitions, we show the number of promoters with the same motif/module, i.e. which are left from the initial partitions (blue); the number of promoters which were initially assigned to a different partition among the five stable subclasses (red); and the number of promoters from the initial 'no motif' and MTE partitions (yellow). Promoters were assigned to several initial partitions in case several motifs/modules had a good hit, and the combined size of the initial partitions thus adds up to more than the total dataset of 1864 promoters.

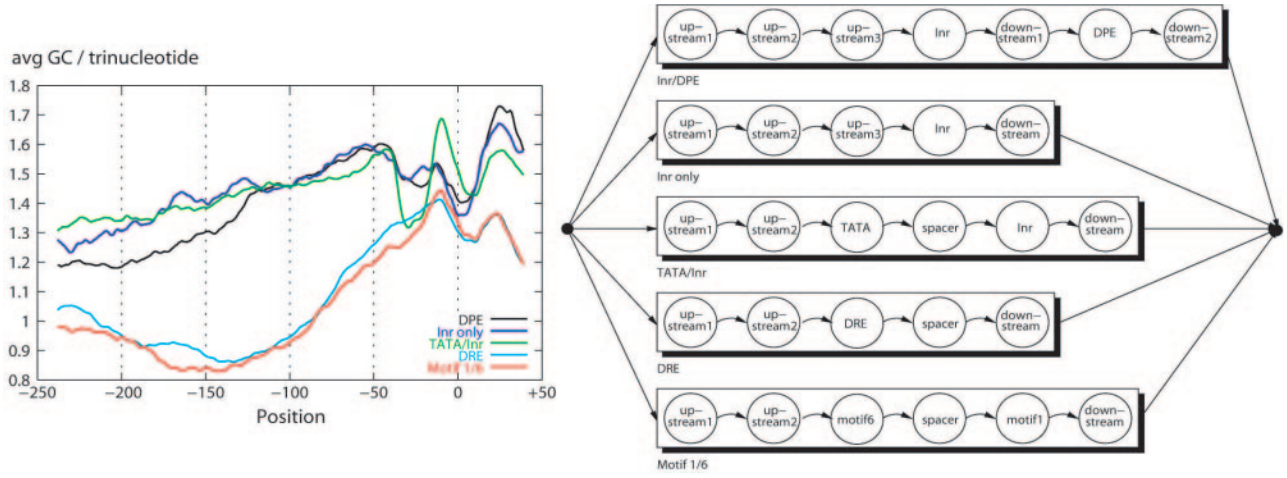

Figure 3. Specific sequence profiles (left) in five different subclasses of core promoters (right). The left shows the average GC trinucleotide content in the region [−250, +50]. The right depicts the different core modules currently modeled in the McPromoter system.

of five models representing different sub-classes and core modules. Figure 3 shows the models for the five subclasses as specified above, and their average GC content in the [−250, 50] region with respect to the TSS. There were two roughly equally frequent groups of core promoters: One with a prominent rise of GC content followed by a 'dip' in the −25 region of the TATA box (comprised of the established TATA, Inr, and DPE classes), and a second one with a distinctly different, GC-poor profile (comprised of the motif 1/6 and DRE classes) which corresponded to the promoters whose functional motifs show a less pronounced location preference. Note that the clustering indicated that there were distinct features within these two broad classes, as can be seen in the M1/6 and DRE class which showed little evidence of taking on new sequences from other initial partitions.

Using explicit models for core promoter subclasses to improve on prediction results

These results suggested adapting computational models to this situation where a core promoter is not regarded as a fixed functional element any longer. Rather, it shows one of several distinct motif arrangements using a repertoire of

motifs, some of which occur preferentially together as pairs and can be regarded as simple transcriptional modules. We applied the five new core promoter models on our previously described benchmark data of 300 bp long promoter and non-promoter sequences (coding and intronic/non-coding; http://www.fruitfly.org/sequence/drosophila-datasets.html), using the same non-promoter models. Comparing the performance of the previous single-promoter-model system with the new five-model system, the best average cross-correlation value of promoter versus non-promoter classification increased from 0.79 to 0.89, with an equal recognition rate improvement from 89.9 to 94.1% and the ROC integral increasing from 0.958 to 0.981. Applying the system to predict promoters in the genome, Table 2 shows the clear improvement we obtained on our curated benchmark test set of 92 promoters from the Adh region (22). In comparison, other state-of-the-art predictors (14) achieve a performance similar to our previous one-model system. The results indicate that we may finally have arrived at a point where we begin to understand the core promoter architecture. For instance, at a detection level of two-thirds, false predictions were made at a frequency of one per gene (the average gene density in Drosophila is ~1 in 9 kb). Using this threshold corresponding to 64% true positives, we made ~35 000 predictions in the non-repeat-masked release four of the *D.melanogaster* genome in total. Table 3 shows the breakdown how often each of the five different core modules was predicted. Compared to the final partitioning of in the training data, the TATA module was predicted more frequently across the whole genome (36% versus 20%), with fewer predictions seen for all other modules. Applied to the *Drosophila pseudoobscura* genome, McPromoter also predicted all five subclasses, but with some distinct differences in their frequency.

Table 2. Comparison of McPromoter using one respectively five promoter models, with the most recent Drosophila predictor proposed in Ref. (14)

McPromoter (one model)		Sharan and Myers		McPromoter (five models)		
Sn	Sp	Sn	Sp	Sn	Sp	FP rate
20	69	20	79	23	91	1/426 590
37	51	35	53	36	79	1/94 797
52	40	50	33	50	47	1/16 097
67	29	65	20	64	36	1/8 203

To enable a fair comparison, the evaluation is done on the same dataset and annotation as in previous publications. Sn: sensitivity, i.e. fraction of correctly identified TSSs among the set of annotated start sites. A TSS is counted as correct if one or more predictions fall within a window of [−500, +50] of the 5′ end of genes in the set. Sp: specificity, i.e. fraction of correctly identified TSSs among the set of total predictions, where predictions are counted within the regions spanned by the genes. FP rate: false positive rate, i.e. the frequency of additional predictions per nucleotide. Numbers for the one-model McPromoter and Sharan and Myers were adapted from Ref. (14).

Table 3. Relative frequency of predicted core modules in *D.melanogaster* and *D.pseudoobscura* (Figure 2).

Core module	Frequency mel/pse (%)
Inr/DPE	12/12
Inr only	6/17
TATA/Inr	36/37
DRE	23/15
Motif 1/6	23/19

Taken together, our results show (i) that a dramatic improvement in promoter prediction accuracy is achieved when taken the variable structure of core promoters into account; (ii) that additional, as yet not fully experimentally validated, promoter motifs play an important role in this; and (iii) that these different architectures are likely conserved across different fly species.

DISCUSSION

It has been argued that identifying TSSs without including some model of the gene structure means to deliberately exclude helpful information. For instance, one system which included features of the first exon reportedly improved on previous approaches (27). For the purpose of accurate annotations of protein-coding genes, this may in fact be an advantageous approach. However, one has to remember that the cell does not know about codons or splice signals at the stage of transcription initiation, and coupling with protein-coding gene models will limit predictions to this class of genes, ignoring other pol-II TSSs like those for miRNAs. In this sense, there is a distinction between the predictions of 5′-untranslated regions (5′-UTRs) (28) and core promoters. For computational investigations of the biology of gene regulation, it makes sense to use only the information available to the cell, and model the core promoter region using sequence-derived features such as arrangements of specific TFBSs. This is in analogy to recent 'splicing simulators' (29) which aim to recapture the process of splicing of a primary transcript without explicitly using coding models, but rather based on motifs which interact with components of the spliceosome. Note that this does not mean that such core promoter information cannot include sequences downstream of the TSS; in fact, any DPE and MTE motifs are part of the mRNA transcript.

Computational methods which aim at the identification of core promoters in this way have traditionally been classified as belonging to one of two groups, depending on how the model captures promoter features (30):

- *Search-by-signal* algorithms make predictions based on the detection of core promoter elements such as the TATA box or the initiator, and/or TFBSs outside the core region (31).
- *Search-by-content* algorithms identify regulatory regions by using measures based on the sequence composition of promoter and non-promoter examples (32).

Search-by-content and hybrid methods combining both ideas have turned out to be more successful than signal-based systems. In the light of our and other recent studies which address the variability of the core promoter regions, this is no longer surprising; approaches which e.g. explicitly expect a TATA box (33) cannot be very successful, given that this motif is lacking in the majority of promoters. In retrospect, the weak performance of predictors of a decade ago was to a large extent due to our lack of data and biological knowledge, rather than due to poorly designed computational approaches.

Today, there is no a priori advantage or disadvantage for either one paradigm—which method is chosen will depend on the particular organism and how much is known about

its functional promoter elements. For Drosophila, we now have an initial comprehensive set of core motifs at our disposition. It is not unlikely that future analyses may discover additional sequence motifs preferentially targeted by components of general transcription factors. In particular, ~15% of the promoters in our curated set do not show strong matches to the core promoter elements in our list. A very recent study has revisited our analysis of Drosophila core promoters in this regard (34), using a larger (albeit not curated) collection of putative TSSs. In addition to the motifs we identified and made use of in this paper, the authors identified new elements: putative variants of DPE and Inr, as well as one positionally restricted and two positionally unrestricted motifs, all of which are present in relatively few core promoters compared to the previously known motifs. Motivated by the increasing evidence of the modularity of core promoters, our approach to base promoter prediction on a collection of several models instead of only one turned out to be very successful: It surpasses the accuracy of our previous results and other Drosophila predictors, and to our best knowledge, achieves the most reliable core promoter predictions for a single eukaryotic species (Table 3). Further improvements to this model can include these recently identified additional motifs, or deviate from a strictly parallel collection of architectures to include motifs such as MTE which do not appear to be part of one particular module only. From a computational view, it will also be interesting to apply model selection techniques, i.e. to automatically derive a set of suitable core promoter architectures, instead of initializing them manually as we did in this study.

Eukaryotic genes may be controlled by several alternative promoters, leading to transcript initiation from multiple distinct start sites. Increasing evidence shows that this is a widespread phenomenon, in organisms ranging from rice to human (35–37). The availability of several promoters may provide a mechanism for more refined expression patterns such as limitation to several tissues. Current large-scale experimental data however suggests that the very concept of a transcription start 'site' is much looser than anticipated; it may not exist for a large number of genes in higher organisms, at least not in the traditional view as precise site defined by functional motifs. High throughput mammalian 5′ SAGE/CAGE data show a broad distribution of initiation events. In fact, the majority of human and mouse genes appear to have more than one distinct start site or a diffuse start 'region', and a large number of additional events are observed in close proximity, with the notable exception of TATA box containing promoters (38). For many human genes, it may simply not be functionally important where exactly the start of transcription occurs, as long as it does not lead to the inclusion of sequence in the primary transcript which has an effect on its post-transcriptional processing. The distinction between an alternative and an imprecise start may not be easy in some cases, but it may be possible to assess the importance of some alternative sites by conservation across species. In light of recent reports on wide-spread intergenic transcription (39), using a second genome does not only serve to filter out false positive predictions, but rather to determine which elements are likely to have a function because they have been conserved. With our current classifier accuracy (94% equal recognition rate), it is thus our opinion that many remaining computational 'false predictions' in genomic DNA may in fact serve as functional TSSs under the right circumstance, or that they could do so, but that e.g. the chromatin structure renders them inaccessible to the transcriptional machinery (40).

Why is the core promoter architecture in Drosophila so variable, given that it should only serve to recruit the polymerase to the start of the gene? One scenario sees this variability as serving to regulate different subsets of genes on a high level, e.g. through the replacement of TBP by TRFs, and recent studies could show associations of sequence motifs with GO categories (16) and additionally expression data (34). Different core promoter architectures also allow for communicating selectively with distal enhancer regions and specific transcription factors (41). The complex structure of TFIID and the core promoter regions may thus serve as modular machinery that allows for an integration of information from different *cis*-regulatory modules. A still possible alternative scenario is that the diversity is frequently a side-effect of TFIID consisting of many subunits, many of which allow for specific interactions with DNA sequence motifs, but which are often interchangeable with each other. The large number of available Drosophila genomes will now allow us to investigate those two scenarios.

In summary, the picture of transcriptional regulatory regions as consisting of a fixed core promoter component and a separate variable component defining specific expression patterns is undergoing a rapid change. Even the role of pol-II as the only player in protein-coding gene expression has changed with the recent discovery that an isoform of a mitochondrial polymerase participates in the transcription of nuclear genes in eukaryotes (42). The precise set of genes regulated in this way, as well as the *cis*-elements and *trans*-factors assisting in their transcription, have yet to be deciphered. Computational methods have to be adapted to keep pace with this and make the most of the increasing wealth of data at our disposition.

ACKNOWLEDGEMENTS

The author thanks Josh Kaminker and Gerry Rubin, then at the Berkeley Drosophila Genome Project, for instigating the research in this paper. We are also grateful to the lab of James Kadonaga at UCSD for continuing discussion and collaboration. UO is an Alfred P Sloan Fellow in Computational Molecular Biology. McPromoter is accessible through http://tools.genome.duke.edu/generegulation, and flat files for all predictions in the release five of the *D.melanogaster* genome assembly are provided. The linux command line version is available from the author for non-commercial purposes. Funding to pay the Open Access publication charges for this article was provided by an award from the Alfred P Sloan foundation.

Conflict of interest statement. None declared.

REFERENCES

1. Li,H. and Wang,W. (2003) Dissecting the transcription networks of a cell using computational genomics. *Curr. Opin. Genet. Dev.*, **13**, 611–616.
2. Levine,M. and Tjian,R. (2003) Transcription regulation and animal diversity. *Nature*, **424**, 147–151.

3. Arnosti,D.N. (2003) Analysis and function of transcriptional regulatory elements: insights from Drosophila. *Annu. Rev. Entomol.*, **48**, 579–602.
4. Wray,G.A., Hahn,M.W., Abouheif,E., Balhoff,J.P., Pizer,M., Rockman,M.V. and Romano,L.A. (2003) The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.*, **20**, 1377–1419.
5. Smale,S.T. and Kadonaga,J.T. (2003) The RNA polymerase II core promoter. *Annu. Rev. Biochem.*, **72**, 449–479.
6. Kutach,A.K. and Kadonaga,J.T. (2000) The downstream promoter element DPE appears to be as widely used as the TATA box in Drosophila core promoters. *Mol. Cell. Biol.*, **20**, 4754–4764.
7. Ohler,U., Liao,G.C., Niemann,H. and Rubin,G.M. (2002) Computational analysis of core promoters in the Drosophila genome. *Genome Biol.*, **3**RESEARCH0087.(Epub December 20, 2002).
8. Suzuki,Y., Tsunoda,T., Sese,J., Taira,H., Mizushima-Sugano,J., Hata,H., Ota,T., Isogai,T., Tanaka,T., Nakamura,Y. *et al.* (2001) Identification and characterization of the potential promoter regions of 1031 kinds of human genes. *Genome Res.*, **11**, 677–684.
9. Molina,C. and Grotewold,E. (2005) Genome wide analysis of Arabidopsis core promoters. *BMC Genomics*, **6**, 25.
10. Hochheimer,A., Zhou,S., Zheng,S., Holmes,M.C. and Tjian,R. (2002) TRF2 associates with DREF and directs promoter-selective gene expression in Drosophila. *Nature*, **420**, 439–445.
11. Lim,C.Y., Santoso,B., Boulay,T., Dong,E., Ohler,U. and Kadonaga,J.T. (2004) The MTE, a new core promoter element for transcription by RNA polymerase II. *Genes. Dev.*, **18**, 1606–1617.
12. Lee,D.H., Gershenzon,N., Gupta,M., Ioshikhes,I.P., Reinberg,D. and Lewis,B.A. (2005) Functional characterization of core promoter elements: the downstream core element is recognized by TAF1. *Mol. Cell. Biol.*, **25**, 9674–9686.
13. Ohler,U. and Frith,M.C. (2005) In Bajic,V. and Tan,T. W. (eds.), *Information Processing and Living Systems*. Imperial College Press, London.
14. Sharan,R. and Myers,E.W. (2005) A motif-based framework for recognizing sequence families. *Bioinformatics*, **21** (Suppl 1), i387–i393.
15. Bajic,V.B., Tan,S.L., Suzuki,Y. and Sugano,S. (2004) Promoter prediction analysis on the whole human genome. *Nat. Biotechnol.*, **22**, 1467–1473.
16. Wang,J. and Hannenhalli,S. (2006) A mammalian promoter model links *cis*-elements to genetic networks. *Biochem. Biophys. Res. Commun.*, **347**, 166–177.
17. Durbin,R., Eddy,S.R., Krogh,A. and Mitchison,G. (1998) *Biological sequence analysis*. Cambridge University Press, Cambridge.
18. Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.
19. Ohler,U., Stemmer,G., Harbeck,S. and Niemann,H. (2000) Stochastic segment models of eukaryotic promoter regions. *Pac. Symp. Biocomput.*, 380–391.
20. Ohler,U., Niemann,H., Liao,G. and Rubin,G.M. (2001) Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition. *Bioinformatics*, **17** (Suppl 1), S199–S206.
21. Schmid,C.D., Perier,R., Praz,V. and Bucher,P. (2006) EPD in its twentieth year: towards complete promoter coverage of selected model organisms. *Nucleic Acids Res.*, **34**, D82–D85.
22. Reese,M.G., Hartzell,G., Harris,N.L., Ohler,U., Abril,J.F. and Lewis,S.E. (2000) Genome annotation assessment in *Drosophila melanogaster*. *Genome Res.*, **10**, 483–501.
23. Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.
24. Hulf,T., Bellosta,P., Furrer,M., Steiger,D., Svensson,D., Barbour,A. and Gallant,P. (2005) Whole-genome analysis reveals a strong positional bias of conserved dMyc-dependent E-boxes. *Mol. Cell. Biol.*, **25**, 3401–3410.
25. Florquin,K., Saeys,Y., Degroeve,S., Rouze,P. and Van de Peer,Y. (2005) Large-scale structural analysis of the core promoter in mammalian and plant genomes. *Nucleic Acids Res.*, **33**, 4255–4264.
26. Schug,J., Schuller,W.P., Kappen,C., Salbaum,J.M., Bucan,M. and Stoeckert,C.J., Jr (2005) Promoter features related to tissue specificity as measured by Shannon entropy. *Genome Biol.*, **6**, R33.
27. Davuluri,R.V., Grosse,I. and Zhang,M.Q. (2001) Computational identification of promoters and first exons in the human genome. *Nature Genet.*, **29**, 412–417.
28. Brown,R.H., Gross,S.S. and Brent,M.R. (2005) Begin at the beginning: predicting genes with 5′-UTRs. *Genome Res.*, **15**, 742–747.
29. Wang,Z., Rolish,M.E., Yeo,G., Tung,V., Mawson,M. and Burge,C.B. (2004) Systematic identification and analysis of exonic splicing silencers. *Cell*, **119**, 831–845.
30. Fickett,J.W. and Hatzigeorgiou,A.G. (1997) Eukaryotic promoter recognition. *Genome Res.*, **7**, 861–878.
31. Prestridge,D.S. (1995) Predicting Pol II promoter sequences using transcription factor binding sites. *J. Mol. Biol.*, **249**, 923–932.
32. Audic,S. and Claverie,J.M. (1997) Detection of eukaryotic promoters using Markov transition matrices. *Comput. Chem.*, **21**, 223–227.
33. Reese,M.G. (2001) Application of a time-delay neural network to promoter annotation in the *Drosophila melanogaster* genome. *Comput. Chem.*, **26**, 51–56.
34. Fitzgerald,P.C., Sturgill,D., Shyakhtenko,A., Oliver,B. and Vinson,C. (2006) Comparative genomics of Drosophila and human core promoters. *Genome Biol.*, **7**, R53.
35. Landry,J.R., Mager,D.L. and Wilhelm,B.T. (2003) Complex controls: the role of alternative promoters in mammalian genomes. *Trends Genet.*, **19**, 640–648.
36. Kitagawa,N., Washio,T., Kosugi,S., Yamashita,T., Higashi,K., Yanagawa,H., Higo,K., Satoh,K., Ohtomo,Y., Sunako,T. *et al.* (2005) Computational analysis suggests that alternative first exons are involved in tissue-specific transcription in rice (Oryza sativa). *Bioinformatics*, **21**, 1758–1763.
37. Zavolan,M., Kondo,S., Schonbach,C., Adachi,J., Hume,D.A., Hayashizaki,Y. and Gaasterland,T. (2003) Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome. *Genome Res.*, **13**, 1290–1300.
38. Carninci,P., Sandelin,A., Lenhard,B., Katayama,S., Shimokawa,K., Ponjavic,J., Semple,C.A., Taylor,M.S., Engstrom,P.G., Frith,M.C. *et al.* (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genet.*, **38**, 626–635.
39. Cheng,J., Kapranov,P., Drenkow,J., Dike,S., Brubaker,S., Patel,S., Long,J., Stern,D., Tammana,H., Helt,G. *et al.* (2005) Transcriptional maps of 10 human chromosomes at 5 nt resolution. *Science*, **308**, 1149–1154.
40. Lieb,J.D. and Clarke,N.D. (2005) Control of transcription through intragenic patterns of nucleosome composition. *Cell*, **123**, 1187–1190.
41. Butler,J.E. and Kadonaga,J.T. (2001) Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs. *Genes. Dev.*, **15**, 2515–2519.
42. Kravchenko,J.E., Rogozin,I.B., Koonin,E.V. and Chumakov,P.M. (2005) Transcription of mammalian messenger RNAs by a nuclear RNA polymerase of mitochondrial origin. *Nature*, **436**, 735–739.