Research

Computational analysis of core promoters in the *Drosophila* genome

Uwe Ohler*†, Guo-chun Liao*, Heinrich Niemann‡ and Gerald M Rubin*§

Addresses: *Department of Molecular and Cell Biology and §Howard Hughes Medical Institute, University of California at Berkeley, Berkeley, CA 94720-3200, USA. ‡Computer Science 5, University of Erlangen-Nuremberg, Martensstrasse 3, D-91058 Erlangen, Germany. †Current address: Department of Biology, Massachusetts Institute of Technology, 77 Massachusetts Ave 68-223, Cambridge, MA 02139, USA.

Correspondence: Uwe Ohler. E-mail: ohler@mit.edu

Abstract

Background: The core promoter, a region of about 100 base-pairs flanking the transcription start site (TSS), serves as the recognition site for the basal transcription apparatus. *Drosophila* TSSs have generally been mapped by individual experiments; the low number of accurately mapped TSSs has limited analysis of promoter sequence motifs and the training of computational prediction tools.

Results: We identified TSS candidates for about 2,000 *Drosophila* genes by aligning 5′ expressed sequence tags (ESTs) from cap-trapped cDNA libraries to the genome, while applying stringent criteria concerning coverage and 5′-end distribution. Examination of the sequences flanking these TSSs revealed the presence of well-known core promoter motifs such as the TATA box, the initiator and the downstream promoter element (DPE). We also define, and assess the distribution of, several new motifs prevalent in core promoters, including what appears to be a variant DPE motif. Among the prevalent motifs is the DNA-replication-related element DRE, recently shown to be part of the recognition site for the TBP-related factor TRF2. Our TSS set was then used to retrain the computational promoter predictor McPromoter, allowing us to improve the recognition performance to over 50% sensitivity and 40% specificity. We compare these computational results to promoter prediction in vertebrates.

Conclusions: There are relatively few recognizable binding sites for previously known general transcription factors in *Drosophila* core promoters. However, we identified several new motifs enriched in promoter regions. We were also able to significantly improve the performance of computational TSS prediction in *Drosophila*.

Background

Transcription initiation is one of the most important control points in regulating gene expression [1,2]. Recent observations have emphasized the importance of the core promoter, a region of about 100 base-pairs (bp) flanking the transcription start site (TSS), in regulating transcription [3,4]. The core promoter serves as the recognition site for the basal transcription apparatus, which comprises the multisubunit RNA polymerase II and several auxiliary factors. Core promoters show specificity both in their interactions with enhancers and with sets of general transcription factors that control distinct subsets of genes. Although there are no

known DNA sequence motifs that are shared by all core promoters, a number of motifs have been identified that are present in a substantial fraction. The most familiar of these motifs is the TATA box, which has been reported to be part of 30-40% of core promoters [5].

Prediction and analysis of core promoters have been active areas of research in computational biology [6], with several recent publications on prediction of human promoters [7-10]. In contrast, prediction of invertebrate promoters has received much less attention and has focused almost exclusively on *Drosophila*. Reese [11] described the application of time-delay neural networks, and in our previous work [12] we used a combination of a generalized hidden Markov model for sequence features and Gaussian distributions for the predicted structural features of DNA. Structural features were also examined by Levitsky and Katokhin [13], but they did not present results for promoter prediction in genomic sequences.

As with computational methods for predicting the intron-exon structure of genes [14], the computational prediction of promoters has been greatly aided by cDNA sequence information. However, promoter prediction is complicated by the fact that most cDNA clones do not extend to the TSS. Recent advances in cDNA library construction methods that utilize the 5′-cap structure of mRNAs have allowed the generation of so-called 'cap-trapped' libraries with an increased percentage of full-length cDNAs [15,16]. Such libraries have been used to map TSSs in vertebrates by aligning the 5′-end sequences of individual cDNAs to genomic DNA [17,18]. However, it is estimated that even in the best libraries only 50-80% of cDNAs extend to the TSSs [16,19], making it unreliable to base conclusions on individual cDNA alignments.

We describe here a more cautious approach for identifying TSSs that requires the 5′ ends of the alignments of multiple, independent cap-selected cDNAs to lie in close proximity. We then examine the regions flanking these putative TSSs, the putative core promoter regions, for conserved DNA sequence motifs. We also use this new set of putative TSSs to retrain and significantly improve our previously described probabilistic promoter prediction method. Finally, we report the results of promoter prediction on whole *Drosophila melanogaster* chromosomes, and discuss the different challenges of computational promoter recognition in invertebrate and vertebrate genomes.

Results and discussion
Selection of expressed sequence tag (EST) clusters to determine transcription start sites
Stapleton *et al.* [20] report the results of aligning 237,471 5′ EST sequences, including 115,169 obtained from cap-trapped libraries, on the annotated Release 2 sequence of the *D. melanogaster* genome. They examined these alignments for alternative splice forms and grouped them into 16,744

clusters with consistent splice sites, overlapping 9,644 known protein-encoding genes. We applied the following set of criteria to select those 5′ EST clusters most likely to identify TSSs. Clusters were required to either overlap a known protein-coding gene or have evidence of splicing. One of the three most 5′ ESTs in the cluster had to be derived from a cap-trapped library. In some cases, disjoint clusters overlap the annotation of a single gene; here, we only considered the most 5′ cluster. We required the distance to the next upstream cluster to be greater than 1 kb. This requirement, together with the selection of only the most 5′ cluster, leads to the selection of only one start site per gene. By doing so, we minimize the erroneous inclusion of ESTs which are not full-length, but also exclude alternative start sites or start sites of genes with overlapping transcript. Because the 5′ ends of ESTs derived from full-length cDNAs are expected to lie in a narrow window at the TSS, we required that the 5′ ends of at least three ESTs fall within an 11-bp window of genomic sequence, and that the number of ESTs whose 5′ ends fall within this window comprise at least 30% of the ESTs in the cluster. With a single EST we cannot be sure to have reached the true start site, even if it was generated by a method selecting for the cap site of the mRNA [17,19]; with a cluster of ESTs within a small range, we can be more confident that we have defined the actual TSS. By requiring selected clusters to have at least three ESTs we are, however, introducing a bias against genes with low expression levels. The requirement that 30% or more of the 5′ ESTs in a cluster terminate within the 11-bp window was introduced because, for large EST clusters, a simple numerical requirement is insufficiently stringent.

We identified a total of 1,941 clusters, representing about 14% of annotated genes, which met all of the above criteria. Table 1 shows how the numbers of selected clusters varies when we change a single parameter specified in the requirement for distance to next upstream cluster and the requirement that the 5′ ends of at least three ESTs fall in a specified window of sequence to a higher or lower value, leaving the other selection requirements constant. Not surprisingly, the most sensitive criterion by far is the window size. A large number of clusters show slightly different 5′ ends, which was also observed by other large-scale full-length cDNA projects [17,18]. At the moment, it is an open question how much of this variation is a result of incomplete extension to the 5′ end during library construction or an indication of a larger than expected variation in the transcription initiation process. The most 5′ EST of each selected cluster, along with its corresponding genomic location, is presented in Supplementary Table 1 in the additional data files available with this paper online (see Additional data files).

We defined the start of the most 5′ EST in each of the 1,941 clusters as the predicted TSS and refer to this as position +1 in the analyses reported below. We extracted the genomic sequences from 250 bp upstream to 50 bp downstream of

Table I

Influence of the parameter values on the number of selected clusters

Parameter	Value	Number of clusters
Minimum distance to next upstream cluster (bp)	0	1,997
	2,000	1,852
Window size (bp)	21	2,691
	16	2,321
	6	1,597
	1	865
Percentage of 5′ ends in window	20	2,008
	40	1,770

The default values are distance, 1,000 bp; window, 11 bp; percentage, 30. The table shows how the number of 1,941 selected clusters varies when one of the parameters is set to a lower or higher value, leaving the others at the default.

each of these sites as a set of putative proximal promoter regions to compare with previous collections of promoters, to identify possible core promoter motifs, and to use as training set for computational promoter prediction. To study the motifs in core promoters with more sensitivity, we also carried out analyses on subsequences from -60 to +40.

Comparison with previous collections of core promoters

Two small collections of curated *Drosophila* TSSs have been assembled previously on the basis of information carefully extracted from the literature. The *Drosophila* Promoter Database (DPD) was the set of 247 TSSs used to train earlier computational promoter finding systems such as NNPP [11] and McPromoter [12]. This DPD was assembled by combining *Drosophila* promoters in the Eukaryotic Promoter Database release 63 [21], and a set of promoters extracted according to similar criteria [22]. The second set was the *Drosophila* Core Promoter Database (CPD [5]) with 205 start sites.

To assess the quality of our inferred TSSs, we aligned the 1,941 300 bp sequences against sequences flanking the TSSs in the DPD and CPD using BLAST [23]. The derivation of our TSS set, which corresponds to just over 14% of all *Drosophila* genes, did not depend on the scientific literature and thus we expect it to be largely non-overlapping with the DPD and CPD sets. It was therefore not surprising that of the 247 core promoter regions in the DPD, only 44 (18%) could be aligned to those in our set. The positions of the TSSs in 28 of these alignments differed by less than 10 bp and are considered identical for our purposes; in five cases, the DPD entries lie more than 10 bp upstream, and in 11 cases, a newly derived putative TSS was more than 10 bp 5′ of the corresponding TSS in the DPD. Of the 205 core promoter regions in the CPD, 32 sites (16%) belonging to

30 genes could be aligned successfully; in 21 out of 30 cases, the difference was again smaller than 10 bp, in six cases, a CPD entry was more 5′, in three cases a newly derived TSS. This simple assessment suggests that our new set of putative TSSs is of similar accuracy to the DPD and CPD. However, our set is eight times larger, containing the predicted TSS for one in seven *Drosophila* genes.

Identification of over-represented sequence motifs in core promoters

Core promoters are known to contain binding sites for proteins important for transcription initiation, and our first analysis of the sequence content of our set of 1,941 core promoters was to assess the representation of two well-established core promoter sequence motifs, the initiator (Inr) and the TATA box. We used the CPD consensus strings for the *Drosophila* Inr and TATA box, TCA(G/T)T(C/T) and TATAAA, respectively [5], permitting up to one mismatch. A search with these consensus strings in random sequences of equal length would result in a frequency of 29.3% for the initiator and 11.6% for the TATA box. Of the CPD promoters, 67.3% have a match to the Inr consensus in the region from -10 to +10 and 42.4% have a TATA box in the region from -45 to -15. We observed that 62.8% of our core promoters had a match to the Inr consensus in the -10 to +10 interval, an almost identical fraction as observed for the CPD. However, we observed a frequency for the TATA box consensus of 28.3%, only about two-thirds of the frequency observed in the CPD; extending the region over which we allowed matches to -60 to -15 only increased the frequency to 33.9%.

We next looked for over-represented motifs using the MEME system to analyze the core promoter regions from -60 to +40 on the leading strand ([24,25] and see Materials and methods). MEME uses the iterative expectation-maximization algorithm to identify conserved ungapped blocks in a set of query sequences, and delivers weight matrix models of the found non-overlapping motifs. The 10 most statistically significant motifs found by this method are listed in Table 2, and their location distributions within the sequences that MEME used in its alignments are shown in Figure 1. Well-known motifs such as the TATA box and Inr are readily found (the third and fourth most significant motifs in Table 2). These motifs are known to have largely fixed locations relative to the TSS, and the tight distribution in the locations of the motifs we observe (Figure 1) implies that the TSSs in our core promoter set have been accurately mapped. Motif 9 matches the previously derived downstream promoter element (DPE) consensus (A/G)G(A/T)(C/T)GT, but Figure 1 suggests a second, distinct, DPE (motif 10).

From the location distribution it is apparent that motif 1 is preferentially found close to the TSS, although not as tightly localized as the Inr motif. Motif 2, which shows a broad spatial distribution within the core promoter regions, corresponds

Table 2

The ten most significant motifs in the core promoter sequences from -60 to +40, as identified by the MEME algorithm

Motif	Pictogram	Bits	Consensus	Number	E value
1		15.2	YGGTCACACTR	311	5.1e-415
2 DRE		13.3	WATCGATW	277	1.7e-183
3 TATA		13.2	STATAWAAR	251	2.1e-138
4 INR		11.6	TCAGTYKNNNTYNR	369	3.4e-117
5		15.2	AWCAGCTGWT	125	2.9e-93
6		15.1	KTYRGTATWTTT	107	1.9e-62
7		12.7	KNNCAKCNCTRNY	197	1.9e-63
8		14.7	MKSYGGCARCGSYSS	82	5.1e-29
9 DPE		15.4	CRWMGCGWKCGGTTS	56	1.9e-12
10		15.3	CSARCSSAACGS	40	8.3e-9

We show the identified motifs in pictogram representation, where the height of letters corresponds to their frequencies relative to the single-nucleotide background used when running MEME. The information content in bits is also calculated with respect to this background. The consensus sequence represents only the highly conserved part of each motif, using the IUPAC code for ambiguous nucleotides. The number of occurrences refers to the sequences that MEME decided to use to build each motif model. The E-value refers to the probability that a motif of the same width is found with equally or higher likelihood in the same number of random sequences having the same single-nucleotide frequencies as our promoter set.

to the target of the DNA replication-related-element binding factor (DREF). At the same time our study was being carried out, DREF was interestingly found to be part of a complex with *Drosophila* TBP-related factor (TRF) 2 [26]. TRF2 replaces the TATA-box-binding element TBP in a distinct subset of promoters, and our data suggest that it is used in a larger fraction of promoters than previously thought.

Because different algorithms for detecting overrepresented motifs have different properties (see Materials and methods),
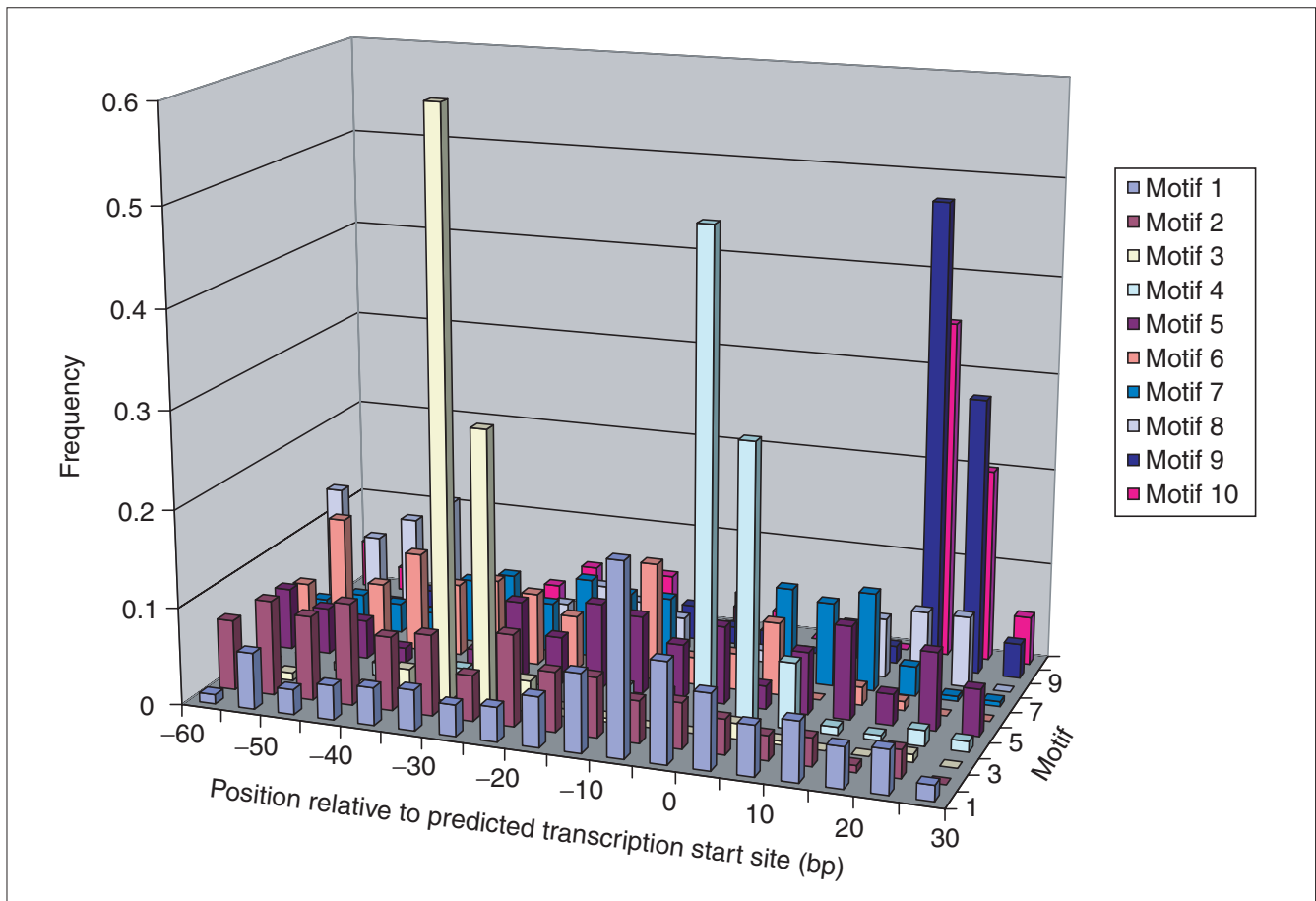
Figure 1
Positional distributions of the occurrence of the 10 most significant motifs relative to the putative transcription start site, as determined by MEME. The positions of base 1 of the motifs as given in the pictograms of Table 2 were binned in 5-bp intervals. (The numerical values plotted here are given in Supplementary Table 4 of the additional data files.)

we compared the motifs identified using a Gibbs sampling algorithm [27,28] with those identified by MEME. Gibbs sampling is non-deterministic and generally delivers different results each time it is run; therefore, the motifs that are most significant are those that are reported multiple times with high likelihood scores. We performed 100 iterations of the algorithm; each iteration was stopped after a maximum of ten motifs were reported. We sorted the resulting 986 motifs by their log-likelihood score (see [28]), which ranged from negative infinity (the logarithm of zero) to 1,353. Among the 354 motifs with scores higher than 700, we observed no motifs that were unrelated to the ten most significant motifs found by MEME (Table 2). Instead, these 354 motifs were comprised of several variants each of motifs 1-6 and 8 of Table 2; the best variant of motif 1 scored 1,353 and the best variant of the TATA box scored 742. We still consider it likely that the three motifs in Table 2 that were not also identified by Gibbs sampling are biologically meaningful. For example, motif 9, is similar in both sequence and positional restriction to the well-established DPE motif.

We were interested in determining which of the 10 motifs shown in Table 2 tend to occur together in individual promoters. We searched the core promoters with each of the 10 weight matrix models, using the program Patser ([29] and see Materials and methods). We restricted the sequence range in which the first base of the model must lie to count as a match as follows: -60 to -15 for the TATA box, -20 to +10 for the Inr, +10 to +25 for the DPE, and -60 to +25 for the other six models. Table 3 gives the percentage of hits for each separate motif, as well as the percentage of promoters containing a specific motif that also contain one of the other motifs. Some previously known dependencies are apparent; for example, DPE containing promoters very often contain an Inr motif, but rarely any of the other motifs. Other obvious correlations are a tendency for motif 6-containing promoters to also contain motif 1, and a tendency for motif 7-containing promoters to contain motif 2 (DRE). Conversely, motif 7 is rarely observed in promoters with a TATA box. There is also a large difference in the likelihood of the DPE and the DPE-like motif 10 to occur in the same promoter

Table 3

Frequency of occurrence of pairs of the ten most significant motifs in the same core promoter

Motif	Percent of promoters with motif	Percent of promoters with each motif that also containing the indicated second motif									
		1	2 DRE	3 TATA	4 Inr	5	6	7	8	9 DPE	10
1	25.1	100.0	21.3	13.1	12.7	20.5	28.3	27.0	27.0	4.9	6.1
2	26.0	20.6	100.0	14.9	16.8	20.0	14.1	33.1	19.4	5.7	6.9
3	19.3	17.1	20.1	100.0	28.9	13.9	14.4	12.6	24.9	4.8	9.4
4	26.3	12.1	16.6	21.1	100.0	14.1	12.1	12.9	25.2	14.9	12.9
5	18.5	27.9	28.1	14.5	20.1	100.0	14.8	29.2	30.6	6.7	8.4
6	15.8	45.1	23.2	17.6	20.3	17.3	100.0	18.6	19.6	4.6	4.2
7	23.3	29.2	36.9	10.4	14.6	23.2	12.6	100.0	30.3	4.9	6.0
8	23.2	29.3	21.8	20.7	28.7	24.4	13.3	30.4	100.0	7.6	10.0
9	7.9	15.6	18.8	11.7	49.4	15.6	9.1	14.3	22.1	100.0	8.4
10	8.5	18.2	21.2	21.2	40.0	18.2	7.9	16.4	27.3	7.9	100.0

The first column lists the motifs given in Table 2. The second column shows the frequency of promoters with a hit to the corresponding weight matrix model (*p* value 1.0e-3). Each of the other columns is labeled with a motif number and the intersection of a row and column shows the frequency with which the two motifs occur in the same core promoter. We did not normalize for the different sizes of the subsets, but entries in the same column can be compared. As we set all thresholds to deliver the same false-positive rate of one in 1,000 nucleotides, we would expect 8.5% of random sequences to contain a match to motifs 1, 5-8 and 10, as the length of the sequence searched allows for 85 different alignment positions of a 15 base motif. Because the sequence windows searched for the other motifs were smaller, the expected false-positive rate was reduced to 4.5% for the TATA box, 3.0% for the Inr, and 1.5% for DPE. Note that the percentage of promoters with TATA boxes or Inr motifs is much lower when estimated using the weight matrix models and Patser than when using matches to the more degenerate consensus strings.

as the TATA box. Motif 1 is the only other motif in addition to the TATA, Inr and DPE motifs to show a marked spatial preference within the core promoter region and tends to occur near the TSS. It is therefore worth noting that there is a bias against co-occurrence of motif 1 and the Inr, suggesting that they may have similar roles in distinct subsets of promoters. Weight matrices for all 10 motifs are provided in Supplementary Table 2 of the additional data files.

To determine whether any of the motifs were enriched in promoters for genes associated with a specific function, process or cellular component, we retrieved the gene ontology (GO) terms [30] associated with the group of genes whose core promoter contained a particular motif. Even though some differences can be seen, it is too early to say whether any are biologically significant. DREF/TRF2 promoters were reported to control genes involved in DNA replication and cell proliferation, but we see no such clear restriction of promoters containing DREs. The most frequent GO terms associated with each motif are given in Supplementary Table 3 of the additional data files.

In a final experiment, we ran MEME on both strands of the entire 300-bp core promoter regions. Table 4 shows the consensus sequences of the 10 most statistically significant

motifs. Note that the initiator and TATA box are no longer identified, owing to the background model and the extended sequences, but that some of the other core motifs of Table 2 are still highly statistically significant. Motif 2 of Table 4 may be related to the reverse complement of the GAGA box, which has been reported to occur in clusters of adjacent copies [31].

Using the core promoter set to retrain the McPromoter TSS prediction tool

McPromoter is a probabilistic promoter prediction system that identifies likely TSSs in large genomic sequences [12,32]. To a certain extent, the performance of such probabilistic systems can be improved by increasing the size of the training set used to estimate the system parameters. The dataset we originally used to train McPromoter consisted of only 247 promoter, 240 non-coding and 711 coding sequences. We took advantage of our new large TSS set to retrain McPromoter (see Materials and methods). We used a slightly smaller set of 1,841 promoter sequences that eliminated instances of related promoters in the 1,941 promoter set, along with a newly extracted representative set of non-promoter sequences taken from *Drosophila* genes (2,635 coding and 1,755 non-coding sequences). The tests we describe below document a markedly improved performance. The

Table 4

The most significant motifs in the extended promoter sequences spanning from -250 to +50, as identified by the MEME algorithm

Motif	Consensus	Number of sequences	IC (bit)	E value
I	YGGTCACACTR	391	16.4	5.0e-369
2	CKCTCTCTCKCTCTC	166	19.1	1.7e-203
3	KCGRCGNCGRCNGCR	153	18.5	1.1e-151
4	TTTKTTTWTWTWTWT	514	13.8	1.5e-155
5	TATCGATAR	246	15.0	4.4e-78
6	CAGCCTGWTTY	187	15.8	1.5e-80
7	STGGCAACGCYR	104	17.6	1.4e-55
8	GTGYGTGTGTGYGTG	106	19.1	6.4e-96
9	YTGCTKYTGCYKYTG	58	20.0	1.2e-39
10	GCGCYTWACAGCAC	34	21.9	9.5e-24

See Table 2 for explanation. Motif 1 corresponds to motif 1 in Table 2; motif 5 to motif 2 (DRE); motif 7 appears to be a variant of motif 8.

largest part of the improvement was due to the increase in the size of the promoter set. Cross-validation experiments in which only a subset of promoters was used suggest that additional sequences similar to those in the current set of 1,841 would not further improve the results significantly (data not shown). However, it is possible that assembling a more representative set of promoters that includes promoters from genes expressed at low levels would improve performance. It might also be useful to make training sets consisting of subsets of promoters that share combinations of motifs and use a collection of McPromoter variants each trained on a different class of core promoter.

Evaluating the performance of the retrained McPromoter

Analysis of the 2.9 Mb Adh region

Table 5 shows the results of the retrained system on the test set of promoters used in the Genome Annotation Assessment Project, which consists of 92 genes annotated with the help of full-length cDNAs in the *Adh* region [33]. A prediction is counted as correct if it falls in the region between -500 and +50 of the annotated 5′ end.

By comparison with another predictor for *Drosophila* promoters, NNPP [11], the retrained McPromoter system has 10-15 times fewer false positives, but some of this improvement simply results from the larger training set and not from differences in the underlying algorithm. When McPromoter was trained on the same, smaller, dataset as NNPP, the reduction in false positives was only about three- to sixfold (data not shown; for details see [34]). The McPromoter system output ranges from 0 to 1, and Table 5 gives the results for a number of thresholds between 0.98 and 0.8; additional predictions below 0.7 are no longer distinguishable

from randomly spaced predictions (see [34] for discussion). Consider, for example, the predictions obtained at a threshold of 0.9; 21 out of 48 true positives were located within +/- 50 bases of the annotated 5′ end, and the average distance of all 48 true positives was 109 bases. As the real TSS is likely to be further upstream than the annotated 5′ end, we believe this result indicates that McPromoter performs well in predicting the precise location of a TSS.

The complete genome: a case study of chromosome arm 2R

The annotation of genomes is a process in flux as new data and analysis tools are continuously refined. We used the reannotated chromosome arm 2R [35] to evaluate our ability to predict promoter regions in a complete eukaryotic chromosome. Of the set of 1,941 TSSs used for retraining McPromoter, 423 correspond to genes located on arm 2R. Of the 2,231 annotated genes on this chromosome arm, 2,130 genes have at least one transcript with an annotated 5′ untranslated region (UTR) (that is, a 5′ UTR with length > 0) and these genes produce a total of 2,742 transcripts with annotated 5′ UTRs. While the average size of a 5′ UTR is about 265 bp [35], the average genomic distance from 5′ UTR start to the beginning of the annotated open reading frame (ORF) in chromosome arm 2R is 1,444 bp; a more detailed distribution is depicted in Figure 2. As evidence for a 5′ UTR, the annotators use the full set of 5′ ESTs, but with less stringent criteria than we used to select for TSSs. Many *Drosophila* genes have large introns in their 5′ UTRs and these introns may not be detected if the cDNA clone from which the EST is derived is not full length; this will lead to frequent placement of the annotated TSSs too close to the ORF.

We therefore counted a hit as positive if it fell within -1,000/+100 of the annotated 5′ end, a region twice as large as for the *Adh* set. Because of the large number of genes with more than one annotated 5′ end, which often lie closely

Table 5

McPromoter results on the *Adh* test set

	Threshold	Sensitivity	Specificity	False-positive rate
McPromoter	0.98	19.5	69.2	1/106,647
	0.95	36.9	50.7	1/25,853
	0.9	52.1	40.3	1/12,016
	0.8	65.2	29.3	1/5,884
NNPP	0.99	21.7	13.5	1/6,227
	0.97	38.0	9.5	1/2,416
	0.92	53.2	6.3	1/1,096

Shown are the sensitivity, which is defined as the percentage of actual TSSs that were correctly predicted, the specificity, which is defined as the percentage of predicted TSSs that correspond to actual TSSs, and the false-positive rate per base. The last three rows show the results obtained with the system NNPP, as reported in [11].
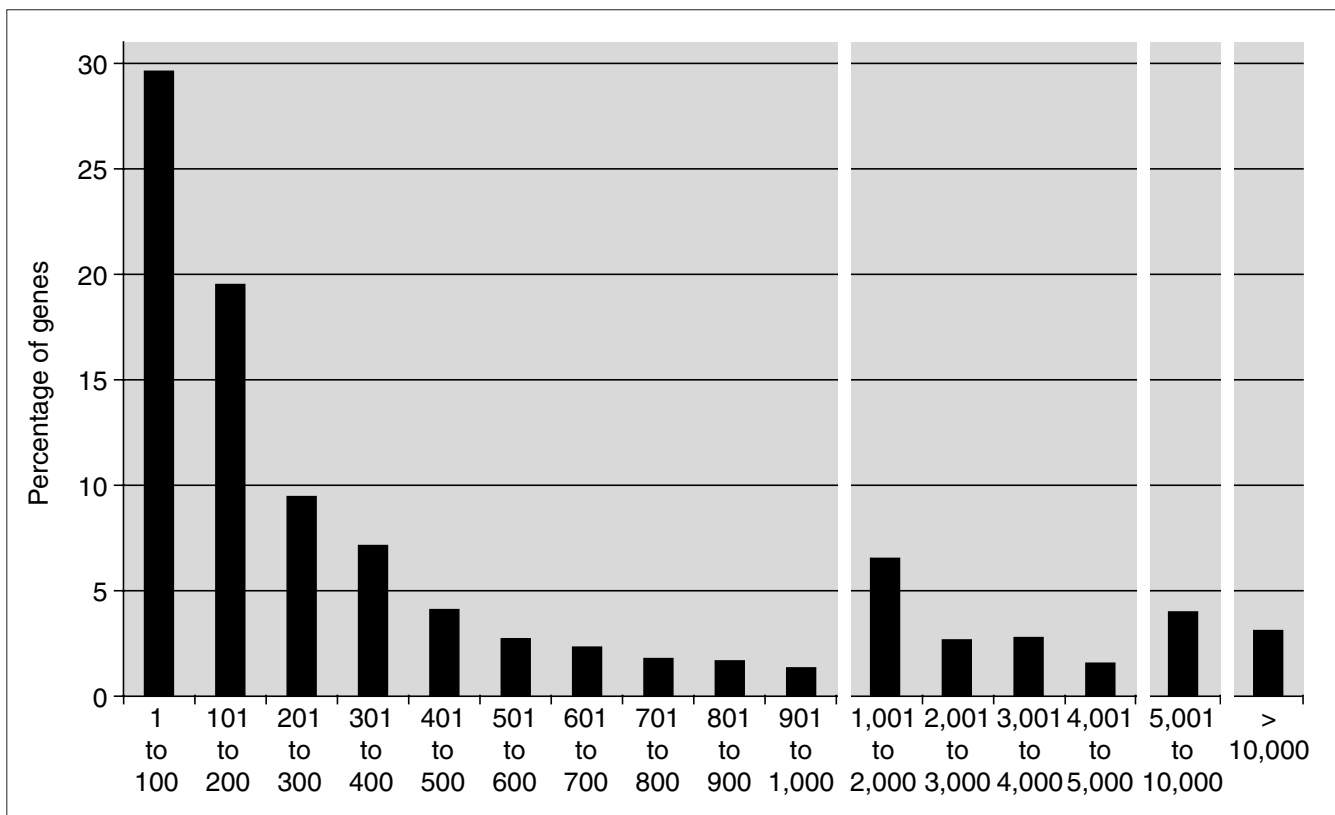
Figure 2
Genomic distance between the predicted TSS and the beginning of the ORF for protein-coding genes with annotated 5′ UTRs on chromosome arm 2R of *D. melanogaster*.

together, we did not evaluate success on the level of individual TSSs, but rather on the level of genes; that is, if at least one prediction falls within the positive region of any of its annotated start sites, it is considered as positive hit. At threshold 0.8, 1,232 TSSs referring to 1,176 (55.2%) genes are correctly identified, with one additional prediction every 5,663 bases across the whole chromosome. This is comparable to the performance on the *Adh* region (see Table 5). We did not pre-filter for low-complexity regions, and 101 genes were left out of the test set because of missing UTR information, so the real false-positive rate is most likely to be lower.

Although it is desirable for a promoter recognition system to be as good as possible when used with no additional information on naked genomic sequence, tools are not used in isolation in a genome annotation project. For example, we used information about the position of genes and considered only the first hit upstream of the translation start codon of each gene, but less than 5 kb upstream of the annotated 5′ UTR end. We therefore make at most one prediction per gene - zero if there is no hit above the predefined threshold within the scanned region. Leaving the threshold at 0.8, McPromoter predicts TSSs for 1,017 genes (47.8%) within the -1,000/+100 region, makes no prediction

for 392 genes (18.4%), and delivers predictions outside the -1,000/+100 region with respect to the annotated TSSs in 721 cases (33.8%).

As more experimental data from mapping TSSs are obtained, it will be interesting to see if the percentage of successful prediction increases. One particularly promising approach will be to use the large amount of additional information we have from 5′ ESTs. As described above, we applied very strict criteria for identifying the 1,941 TSSs used in our training set. We have, however, at least one 5′ EST for an additional 8,000 genes [20]. Looking for coincidences between the TSSs predicted by McPromoter and the genomic positions of the 5′ ends of these ESTs is likely to be a powerful approach. As an initial test of this idea, we ran McPromoter with a very low threshold of 0.75 on the whole genome, and retained all the hits as long as they were more than 100 bp apart. As above, no filtering for repeats and low-complexity regions was carried out. Because of the very low threshold, we generated an average of one predicted TSS every 3,000 nucleotides in the genome. We found that 11,160 - approximately one in seven - of these predictions is on the appropriate strand and within 500 bp of an EST 5′ end. If these predictions

are valid, we would expect them to be closer than 500 bp to an EST end. Indeed, we found that 56% of the 11,160 predictions in fact lie within 50 bp of an EST, much higher than would be expected by chance, and in accordance with the results in the much smaller *Adh* region set. Even when we require co-localization within 10 bp, we still retain 25% of the predictions originally selected to be within 500 bp. The next step is to optimize both the threshold we use for predicting the TSSs and the window size we use for assessing their correlation with 5′ EST ends. In the end, we will still need to conduct experiments such as primer extension to distinguish real from false-positive predictions.

A comparison with vertebrate promoter finding

It is instructive to examine our results on the *Drosophila* genome in the context of recent work on promoter finding in vertebrates. As a benchmark set, the 'known' genes of human chromosome 22 are widely used, as it was the first completely sequenced human chromosome [36]. The annotation of release 2.3 of May 2001 contained 339 known genes. We applied the first version of McPromoter trained on human data [37], which uses only sequence features and a simple Markov chain model for the whole promoter sequence. Along the guidelines of earlier evaluations, we counted a prediction as correct if it fell within -2,000 and +500 bp relative to the 5′ end of the annotation [9,10], and retained only the best hit within a window of 2,000 bases. As a negative set, we used the sequences downstream of +500 until the end of each gene annotation. Table 6 shows the results for different thresholds. Despite the large region where predictions count as correct, the average distance of true positive predictions is about 250 bp, and about 40% of these are located within ±100 bp, making our results comparable to those reported by other groups. For example, Scherf *et al.* [10] report a sensitivity of 45% on the set of known genes and specificity of 40% for the whole of chromosome 22. They do not attempt to predict the strand of the promoter (and also do not predict TSS locations but rather promoter regions of an average of 555 bp); as our evaluation considers only the sense strand of the genes, the numbers are comparable.

The good results obtained on human data with even a simple sequence model are apparently largely due to the strong correlation of vertebrate promoters with CpG islands. 60% of the 5′ ends of known genes on chromosome 22 are located within CpG islands, regions in the genome that are not depleted of CG dinucleotides and are associated with an open chromatin structure [38]. At 64% sensitivity, 82% of the true predictions by McPromoter are located within CpG islands, and the correlation gets stronger as the specificity increases. This was also reported for other, more recently developed promoter-finding systems [9,10]. Therefore, promoter-finding methods based on sequence information successfully identify almost the exact same subset of promoters, owing to the high correlation

Table 6

Results of McPromoter on human data

Sensitivity	Specificity	False-positive rate
39.5	72.0	1/237,475
52.8	62.6	1/115,408
64.3	36.4	1/32,411

Sensitivity, specificity and false-positive rate are defined as in Table 5.

with the subset of promoters located within CpG islands [8]. Vertebrate promoter recognition thus appears to be reaching its limit when the models use the proximal promoter sequence as the only information source.

In the only system so far with a significantly better performance and smaller correlation with CpG islands, promoter recognition is guided by a simultaneous recognition of first exons [7]. When we used the version of McPromoter that includes analysis of the physical properties of the genomic DNA, as we do with *Drosophila*, the true positives are much less correlated with CpG islands (61%), and therefore constitute a broader subset of vertebrate promoters. Unfortunately, the false-positive rate increased roughly fivefold, and the predictions tend to be located farther away from the 5′ end of genes [34]. The promoters of vertebrate and invertebrate organisms differ in that invertebrate genomes do not contain CpG islands, a feature of more than half of vertebrate genes [39]. This makes computational recognition of invertebrate promoters, and those vertebrate promoters not found in CpG islands, more difficult.

Conclusions

In this paper, we present a strategy for annotating core promoters in the complete *Drosophila* genome by a two-step process of 5′ EST cluster selection and computational prediction. With the help of a larger training set, we were able to significantly improve the performance of McPromoter, our computational TSS prediction tool. Probably for the first time in invertebrate promoter prediction, the results are sensitive and specific enough to guide verification by subsequent wet lab experiments such as primer extension.

A first analysis of motifs prevalent in core promoters revealed that less than one-third of *Drosophila* promoters have a consensus TATA box. In contrast, the DRE motif that is part of the recognition site for an alternative transcription initiation complex that utilizes TRF2 [26] is more frequent than previously thought. One surprising result of our work is that there are relatively few recognizable binding sites for known general transcription factors in *Drosophila* core promoters. Our analysis did, however, reveal previously undescribed or underappreciated motifs,

an encouraging sign that there are distinct features in *Drosophila* core promoter regions.

McPromoter is accessible at [32]. The training sets are available from the *Drosophila* Genome Project website [40]. McPromoter predictions are part of the analysis results in the Genome Annotation Database Gadfly [41].

Materials and methods
EST clustering
Alignment and clustering of ESTs on the complete *Drosophila* sequence are described in detail in [20].

Motif finding
We used the MEME system, version 3.0 [24,25] to look for core promoter motifs using a motif width range of 5-15 bp and zero or one occurrence per sequence. That is, MEME decides during the motif identification whether or not to include a sequence; the numbers of sequences that MEME used for the alignments are given in Tables 2 and 4. We observed that MEME is rather restrictive in deciding whether to use a particular sequence or not; thus, the number of sequences in any given motif alignment is rather small, and searching for binding sites as described below resulted in larger numbers of sequences with a particular motif. We used the standard background model of MEME, which corresponds to the single-nucleotide frequencies in the set. For the analysis of the extended sequences from -250 to +50, we used a third-order Markov chain as background model, which should prevent the algorithm from reporting ubiquitous short repeat motifs. The background parameters were estimated on the -250 to +50 sequences.

We also applied the Gibbs sampling algorithm for motif identification [27]. Our data set was too large to be submitted to the available web-based sites. We therefore used an implementation adapted to DNA sequence analysis that we could install locally [28] to analyze sequences from -60 to +40 in the same set of promoters used with MEME. As with other implementations of Gibbs sampling, this implementation [28] does not automatically estimate the width of motifs, which we fixed at 10 bases. Also, there is no equivalent of the 'zero or one occurrence per sequence' statistics of MEME; instead, we used a prior probability of 0.075 for motif occurrence. This property makes the identification of rarely occurring motifs intrinsically harder.

Searching for binding sites
To look for binding sites of the motifs, we used the weight matrices as computed by MEME with the program Patser v3d [29]. As the searched regions are rather small, we set a rather low threshold corresponding to a *p*-value of 1e-3; that is, once every 1,000 nucleotides in a random sequence we can expect to see a score equal or better than the threshold, assuming that the sequence has the same base composition as our core promoter set.

Datasets
To retrain the McPromoter system, we used the newly determined TSSs, and extracted 250 bp upstream and 50 bp downstream of each TSS as training sequences. We constructed a representative set starting from these 1,941 sequences similar to the guidelines of the Eukaryotic Promoter Database [21]: Sequences are only included if they share less than 50% identity in the core promoter region (in our case, -60 to +40) to each of the promoters in the growing set. This resulted in a data set of 1,864 sequences. We aligned this set to the *Adh* region that we use as test sequence, and 23 promoters within the *Adh* region were removed from that set.

For non-promoter sequences, we started with a representative set of genes compiled in 1999 that comprised 416 single- and multi-exon genes [33]. In addition, we used the genes contained in the *Drosophila* Gene Collection as of 31 December, 2001. In a first step, we eliminated genes that showed evidence of alternative splicing. The remaining set was aligned to the SWISS-PROT database [42] using TBLASTX [43], and we retained all cDNA entries where the protein completely aligned with the cDNA, and the similarity was above 90% for *Drosophila* proteins and 40% for non-*Drosophila* entries. 5′ and 3′ UTRs were truncated, and the remaining sequences were aligned to the genomic sequence using Sim4 [44], to map the exon-intron structure. This resulted in a set of 264 entries with hits to *Drosophila* SWISS-PROT entries, and 29 against proteins of other organisms. These sequences were then added one by one to the existing initial set of 416 genes, provided that the similarity of the coding sequence was less than 80% to genes already in the set. A total of 229 sequences were added in this way, resulting in a final set of 645 single- and multi-exon genes. Exons and introns of these genes were then divided into 300-bp long non-overlapping stretches to generate 2,635 coding and 1,755 intronic sequences of 300 bp each.

Computational promoter prediction
The McPromoter system consists of a generalized hidden Markov model of a promoter sequence, with six interpolated Markov chain submodels representing different segments of the promoter sequence from -250 to +50: upstream 1 and 2, TATA box, spacer, initiator and downstream. It is augmented by a set of Gaussian densities describing profiles of DNA physicochemical properties [45] in the six segments. The likelihoods of sequence and profile segments are finally combined in a neural network classifier, which is trained to output values of 1 for promoter and 0 for non-promoter sequences. Cross-validation tests of 14 different properties such as DNA bendability and nucleosome positioning (compiled in [46]) showed that stacking energy delivered the best classification results for both human and *Drosophila*

sequences when combined with the sequence probabilities. On our previous small *Drosophila* dataset, a combination of all features by principal-component analysis (PCA, see for example [47]) led to an improvement in classification; however, on the current larger dataset, no additional improvement is observed when using PCA.

We carried out five cross-validation experiments on the new dataset, setting aside one-fifth as a test set and taking the remaining four-fifths for training. The best average cross-correlation value (CC) of promoter versus non-promoter classification was 0.79, with an equal recognition rate of 89.9% and a receiver operating characteristics integral of 9,575.

To localize promoters in genomic sequences, an input window of 300 bp is shifted along both sides of the sequence, and the neural network score as well as the position of the initiator segment is stored for each window. After smoothing the score graph along the sequence, the best local maximum within a certain range is reported if it exceeds a preset threshold on the score. A detailed description can be found in [12,34].

Evaluation measures

For classification into promoter/non-promoter, we used the equal recognition rate (ERR) and the integral over the receiver operating characteristics (ROC) as measures. A ROC curve gives the recognition rate (true positives) for preselected values of false positives, which was in our case 0 to 100% in 1% steps. Then, the trapezoid rule is used to compute the integral numerically. The highest achievable value is 10,000 (100 x 100, the perfect recognition for all rates of false positives); a random classification results in a value of 5,000. ERR gives the recognition performance at the point where the rate of true positives equals the rate of true negatives; the ROC integral judges the performance in a more global manner. We also give the correlation coefficient (*CC*), defined as:

$$CC = \frac{(TP \cdot TN) - (FN \cdot FP)}{\sqrt{(TP + FN) \cdot (TN + FP) \cdot (TP + FP) \cdot (TN + FN)}}$$

where *TP* stands for true positives, *TN* for true negatives, *FP* for false positives, and *FN* for false negatives; these numbers denote the absolute numbers of correctly and incorrectly classified sequences. CC values should be considered with caution, as they depend on the relative size of the dataset for each class and can therefore generally not be compared across different datasets.

Additional data files

Additional data files containing the following supplementary tables are available with the online version of this paper: Supplementary table 1, showing alignment positions of the most 5′ ESTs of the 1,941 selected EST clusters, for Release 2 and the realignment to Release 3; Supplementary table 2, listing the weight matrixes for the 10 motifs shown in Table 2; Supplementary table 3, listing the most frequent GO terms associated with each of the 10 motifs shown in Table 2; and Supplementary table 4 with the raw data for the positional distribution of motif hits (Figure 1).

References
1. Latchman DS: *Gene Regulation - A Eukaryotic Perspective*, 4th edn. Cheltenham, UK: Nelson Thornes, 2002.
2. Lemon B, Tjian R: **Orchestrated response: a symphony of transcription factors for gene control.** *Genes Dev* 2000, **14:**2551-2569.
3. Smale ST: **Core promoters: active contributors to combinatorial gene expression.** *Genes Dev* 2001, **15:**2503-2508.
4. Butler JEF, Kadonaga JT: **The RNA polymerase II core promoter: a key component in the regulation of gene expression.** *Genes Dev* 2002, **16:**2583-2592.
5. Kutach AK, Kadonaga JT: **The downstream promoter element DPE appears to be as widely used as the TATA box in *Drosophila* core promoters.** *Mol Cell Biol* 2000, **20:**4754-4764.
6. Fickett JW, Hatzigeorgiou AG: **Eukaryotic promoter recognition.** *Genome Res* 1997, **7:**861-878.
7. Davuluri RV, Grosse I, Zhang MQ: **Computational identification of promoters and first exons in the human genome.** *Nat Genet* 2001, **29:**412-417.
8. Down TA, Hubbard TJ: **Computational detection and location of transcription start sites in mammalian genomic DNA.** *Genome Res* 2002, **12:**458-461.
9. Hannenhalli S, Levy S: **Promoter prediction in the human genome.** *Bioinformatics* 2001, **17:**S90-S96.
10. Scherf M, Klingenhoff A, Frech K, Quandt K, Schneider R, Grote K, Frisch M, Gailus-Durner V, Seidel A, Brack-Werner R, *et al*: **First pass annotation of promoters on human chromosome 22.** *Genome Res* 2001, **11:**333-340.
11. Reese MG: **Application of a time-delay neural network to the annotation of the *Drosophila melanogaster* genome.** *Comput Chem* 2001, **26:**51-56.
12. Ohler U, Niemann H, Liao GC, Rubin GM: **Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition.** *Bioinformatics* 2001, **17:**S199-S206.
13. Levitsky VG, Katokhin AV: **Computational analysis and recognition of *Drosophila melanogaster* gene promoters.** *Mol Biol* 2001, **35:**826-832.
14. Haas BJ, Volfovsky N, Town CD, Troukhan M, Alexandrov N, Feldmann KA, Flavell RB, White O, Salzberg SL: **Full-length messenger RNA sequences greatly improve genome annotation.** *Genome Biol* 2002, **3:**research0029.1-0029.12.
15. Carninci P, Shibata Y, Hayatsu N, Sugahara Y, Shibata K, Itoh M, Konno H, Okazaki Y, Muramatsu M, Hayashizaki, Y: **Normalization and subtraction of cap-trapper-selected cDNAs to prepare full-length cDNA libraries for rapid discovery of new genes.** *Genome Res* 2000, **10:**1617-1630.

16. Suzuki Y, Yoshitomo-Nakagawa K, Maruyama K, Suyama A, Sugano S: **Construction and characterization of a full length-enriched and a 5′-end-enriched cDNA library.** *Gene* 1997, **200:**149-156.
17. Suzuki Y, Taira H, Tsunoda T, Mizushima-Sugano J, Sese J, Hata H, Ota T, Isogai T, Tanaka T, Morishita S, *et al.*: **Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites.** *EMBO Rep* 2001, **2:**388-393.
18. Suzuki Y, Tsunoda T, Sese J, Taira H, Mizushima-Sugano J, Hata H, Ota T, Isogai T, Tanaka T, Nakamura Y, *et al.*: **Identification and characterization of the potential promoter regions of 1031 kinds of human genes.** *Genome Res* 2001, **11:**677-684.
19. Sugahara Y, Carninci P, Itoh M, Shibata K, Konno H, Endo T, Muramatsu M, Hayashizaki Y: **Comparative evaluation of 5′-end-sequence quality of clones in CAP trapper and other full-length-cDNA libraries.** *Gene* 2001, **263:**93-102.
20. Stapleton M, Liao GC, Brokstein P, Hong L, Carninci P, Shiraki T, Hayashizaki Y, Champe M, Pacleb J, Wan K, *et al.*: **The *Drosophila* Gene Collection: identification of putative full-length cDNAs for 70% of *D. melanogaster* genes.** *Genome Res* 2002, **12:**1294-1300.
21. Cavin Perier R, Praz V, Junier T, Bonnard C, Bucher P: **The Eukaryotic Promoter Database (EPD).** *Nucleic Acids Res* 2000, **28:**302-303.
22. Arkhipova I: **Promoter elements in *Drosophila melanogaster* revealed by sequence analysis.** *Genetics* 1995, **139:**1359-1369.
23. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.
24. Bailey TL, Elkan C: **The value of prior knowledge in discovering motifs with MEME.** *Proc Int Conf Intell Syst Mol Biol* 1995, **3:**21-29.
25. **The MEME/MAST system** [http://meme.sdsc.edu]
26. Hochheimer A, Zhou S, Zheng S, Holmes MC, Tjian R: **TRF2 associates with DREF and directs promoter-selective gene expression in *Drosophila*.** *Nature* 2002, **420:**439-444.
27. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC: **Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment.** *Science* 1993, **262:**208-214.
28. Thijs G, Marchal K, Lescot M, Rombauts S, De Moor B, Rouze P, Moreau Y: **A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes.** *J Comput Biol* 2002, **9:**447-464.
29. Hertz GZ, Stormo GD: **Identifying DNA and protein patterns with statistically significant alignments of multiple sequences.** *Bioinformatics* 1999, **15:**563-577.
30. The Gene Ontology Consortium: **Creating the gene ontology resource: design and implementation.** *Genome Res* 2001, **11:**1425-1433.
31. Katsani KR, Hajibagheri MA, Verrijzer CP: **Co-operative DNA binding by GAGA transcription factor requires the conserved BTB/POZ domain and reorganizes promoter topology.** *EMBO J* 1999, **18:**698-708.
32. **McPromoter MM:II** [http://genes.mit.edu/McPromoter.html]
33. Reese MG, Hartzell G, Harris NL, Ohler U, Abril JF, Lewis SE: **Genome annotation assessment in *Drosophila melanogaster*.** *Genome Res* 2000, **10:**483-501.
34. Ohler U: **Computational Promoter Recognition in Eukaryotic Genomic DNA.** PhD thesis, University of Erlangen-Nuremberg, 2001.
35. Misra S, Crosby MA, Mungall CJ, Matthews BB, Campbell KS, Hradecky P, Huang Y, Kaminker JS, Millburn GH, Prochnik SE, *et al*: **Annotation of the *Drosophila melanogaster* euchromatic genome: a systematic review.** *Genome Biol* 2002, **3:**research0083.1-0083.22.
36. Dunham I, Shimizu N, Roe BA, Chissoe S, Hunt AR, Collins JE, Bruskiewich R, Beare DM, Clamp M, Smink LJ*, et al.*: **The DNA sequence of human chromosome 22.** *Nature* 1999, **402:**489-495.
37. Ohler U, Harbeck S, Niemann H, Noth E, Reese MG: **Interpolated Markov chains for eukaryotic promoter recognition.** *Bioinformatics* 1999, **15:**362-369.
38. Bird A: **DNA methylation patterns and epigenetic memory.** *Genes Dev* 2002, **16:**6-21.
39. Ponger L, Duret L, Mouchiroud D: **Determinants of CpG islands: expression in early embryo and isochore structure.** *Genome Res* 2001, **11:**1854-1860.
40. **Representative benchmark data sets of *D. melanogaster* DNA sequences** [http://www.fruitfly.org/sequence/drosophila-datasets.html]
41. **GadFly: genome annotation database of *Drosophila*** [http://www.fruitfly.org/annot]
42. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000.** *Nucleic Acids Res* 2000, **28:**45-48.
43. Gish W, States D. **Identification of protein encoding regions by database similarity search.** *Nature Genet* 1993, **3:**266-272.
44. Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W: **A computer program for aligning a cDNA sequence with a genomic sequence.** *Genome Res* 1998, **8:**967-974.
45. Pedersen AG, Baldi P, Chauvin Y, Brunak S: **DNA structure in human RNA polymerase II promoters.** *J Mol Biol* 1998, **281:**663-673.
46. Liao GC, Rehm EJ, Rubin GM: **Insertion site preferences of the P transposable element in *Drosophila melanogaster*.** *Proc Natl Acad Sci USA* 2000, **97:**3347-3351.
47. Duda RO, Hart PE, Stork DG: *Pattern Classification*, 2nd edn. New York: John Wiley; 2000.
48. **Pictogram** [http://genes.mit.edu/pictogram.html]