PREDICTION OF TRANSCRIPTION START SITES BASED ON FEATURE SELECTION USING AMOSA

Xi Wang[1#], Sanghamitra Bandyopadhyay[3,1#], Zhenyu Xuan[2], Xiaoyue Zhao[2], Michael Q. Zhang[2,1] and Xuegong Zhang[1*]

[1]*Bioinformatics Division, TNLIST and Dep. of Automation, Tsinghua Univ., Beijing 100084, China*

[2]*Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA*

[3]*Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700 108, India*

[#]*The first two authors are joint first authors.*

[*]*Email: zhangxg@tsinghua.edu.cn*

To understand the regulation of the gene expression, the identification of transcription start sites (TSSs) is a primary and important step. With the aim to improve the computational prediction accuracy, we focus on the most challenging task, i.e., to identify the TSSs within 50 bp in non-CpG related promoter regions. Due to the diversity of non-CpG related promoters, a large number of features are extracted. Effective feature selection can minimize the noise, improve the prediction accuracy, and also to discover biologically meaningful intrinsic properties. In this paper, a newly proposed multi-objective simulated annealing based optimization method, Archive Multi-Objective Simulated Annealing (AMOSA), is integrated with Linear Discriminant Analysis (LDA) to yield a combined feature selection and classification system. This system is found to be comparable to, often better than, several existing methods in terms of different quantitative performance measures.

1. INTRODUCTION

It is known that the initiation of transcription of a gene is the important first step in gene expression. RNA polymerase II (Pol II) plays the key role during transcription and is recruited by other transcription factors (TFs) to TSS within the preinitiation complex (PIC). Determining the location of the TSSs has become crucial for mapping the cis-regulatory elements and hence for further studying the mechanism of gene regulation.

The core promoter region is centered around the TSS, within a length of ~100bp and the proximal promoter, which is also enriched by transcription factor binding sites (TFBSs), is located immediately upstream of the core promoter within several hundred base pairs. Given the relationship between promoters and TSSs, a promoter region must contain the information for PolII to recognise TSSs, this information forms the basis for identifying the TSS *in silico*. Moreover, through computational modeling, important cis-regulatory element features may be identified. Good predictive features and accurate TSSs will help forming testable hypothesis and designing targeted experiments.

Predicting the TSS *in silico* is an old but still very challenging problem. A strategy was proposed by Zhang in 1998[1]: an initial identification of a promoter approximately within a distance of 2 kb, followed by a more specific prediction method to locate the TSS within a 50 bp region. Many methods have been developed in the past decade[2], generally belonging to the two categories: namely the initial identification of a gross promoter and the more specific prediction of the TSS. It is also demonstrated by recent studies that, in vertebrates, one should treat the CpG related and the non-CpG related promoters (see the definition of non-CpG related promoters in the data section) separately for better TSS prediction, which is biologically sound and computationally feasible[3]. For the CpG related promoters, the TSS prediction is much easier and has largely been solved[2, 4]. However, predicting TSSs for non-CpG related promoters remains challenging. In this paper we focus on predicting TSSs within 50 bp for non-CpG related promoters.

Almost all the previous methods for TSS prediction have been summarized in the recent reviews[3, 5]. The main idea of those methods is to use some characteristic features, which can differentiate between a promoter region a non-promoter region, in the classification tests. The resulting classifiers (or predictive models) are applied to new input DNA sequences for TSS prediction. However, Bajic *et al*[5] describe detecting TSSs in non-

[*] Corresponding author.

CpG related promoter as a bottleneck of current technology. The reason may be due to poor understanding of transcriptional initiation mechanism and the diversity of non-CpG related promoters, especially tissue specific promoters. Hence the features which could be used to distinguish the promoter from non-promoter regions cannot be easily determined. To solve this problem, one strategy is to start with a large number of potential features and then select the most discriminative ones according to certain classification objectives.

A good feature selection not only can improve the accuracy of the prediction, but also can reveal biologically meaningful features which may provide deeper biological insights. Feature selection is the process of selecting a subset of the available features such that some internal or external criterion is optimized. The purpose of this step is the following: building simpler and more comprehensible models, improving the performance of some subsequent machine learning algorithm, and helping to prepare, clean, and understand data. Different algorithms exist for performing feature selection. One important approach is to use an underlying search and optimization technique like genetic algorithms[6, 7]. However, it may often be difficult to evolve just a single criterion that is sufficient to capture the goodness of a selected subset of features. It may thus be more appropriate and natural to treat the problem of feature selection as one of multi-objective optimization. Such an approach is adopted in this article. A newly developed multi-objective simulated annealing algorithm called Archived Multi-Objective Simulated Annealing (AMOSA)[8, 9] is utilized for this purpose.

2. MATERIALS AND METHOD

2.1. AMOSA

Archived multi-objective simulated annealing (AMOSA)[8, 9] is a generalized version of the simulated annealing (SA) algorithm based on multi-objective optimization (MOO). MOO is applied when dealing with the real-world problems where there are several objectives that should be optimized simultaneously. In general, a MOO algorithm usually admits a set of solutions that are not dominated by any solution it encountered, i.e., non-dominated solutions.[10, 11] During recent years, many multi-objective evolution algorithms, such as Multi-Objective SA (MOSA), have been suggested to solve the MOO problems.[12]

Simulated annealing (SA) is a search technique for solving difficult optimization problems, which is based on the principles of statistical mechanics[13]. Recently, SA has become very popular because not only can SA replace the exhaustive search to save time and resource, but also converge to the global optimum if annealed sufficiently slowly[14].

Although the single objective version of SA is quite popular, its utility in the multi-objective case was limited because of its search-from-a-point nature. Recently Bandyopadhyay *et al* developed an efficient multi-objective version of SA called AMOSA[8, 9] that overcomes this limitation. AMOSA is utilized in this work for selecting features for the task of TSS prediction.

The AMOSA algorithm incorporates the concept of an *archive* where the non-dominated solutions seen so far are stored. Two limits are kept on the size of the archive: a hard or strict limit denoted by *HL*, and a soft limit denoted by *SL*. The algorithm begins with the initialization of a number (γ x *SL*, $0<\gamma<1$) of solutions each of which represents a state in the search space. The multiple objective functions are computed. Each solution is refined by using simple hill-climbing and domination relation for a number of iterations. Thereafter the non-dominated solutions are stored in the archive until the size of the archive increases to *SL*. If the size of the archive exceeds *HL*, a single-linkage clustering scheme is used to reduce the size to *HL*. Then, one of the points is randomly selected from the archive. This is taken as the *current-pt*, or the initial solution, at temperature *T=Tmax*. The *current-pt* is perturbed to generate a new solution named *new-pt,* and its objective functions are computed. The domination status of the *new-pt* is checked with respect to the *current-pt* and the solutions in the archive. A quantity called amount of domination $\Delta dom_{a,b}$ between two solutions *a* and *b* is defined as follows:

$$\Delta \mathrm{dom}_{\mathrm{a,b}} = \prod_{i=1, f_i(a) \neq f_i(b)}^{M} \frac{f_i(a) - f_i(b)}{R_i} \qquad (1)$$

where $f_i(a)$ and $f_i(b)$are the i^{th} objective values of the two solutions and R_i is the corresponding range of the objective function. Based on domination status different cases may arise viz., accept the (i) *new-pt*, (ii) *current-pt*

or (iii) a solution from the archive. Again, in case of overflow of the archive, clustering is used to reduce its size to *HL*. The process is repeated *iter* times for each temperature that is annealed with a cooling rate of α(<1) till the minimum temperature *Tmin* is attained. The process thereafter stops, and the archive contains the final non-dominated solutions.

It has been demonstrated that the performance of AMOSA is better than that of MOSA[15] and NSGA-II[16], both are well-known MOO algorithm, in many applications.[9]

2.2. Data

Since TSS prediction for non-CpG related promoters is still unsolved, our newly proposed prediction system is aimed at such TSSs. All the data used come from the work of Zhao *et al*[17]. We take promoter sequences as non-CpG related if the normalized CpG content of the 3 kb centered at the TSS is less than 0.3.

All of the examples were taken from Eukaryotic Promoter Database (EPD)[18] and the Database of Transcription Start Site (DBTSS)[19], both of them have relatively high quality annotation. EPD is based on experimentally determined TSSs while DBTSS on full-length oligo-capped cDNA sequences. A total of 1,570 sequences containing non-CpG related promoters were selected, including 299 from EPD and 1,271 from DBTSS. These sequences are of 10 kb in size and are centered at the annotated TSSs.

Positive and negative samples are defined in order to train LDA classifiers (or prediction models) to accomplish the more specific prediction of TSSs. Within the known promoter regions, positive samples are the sequences around core promoters from site -250 to site 50, denoted as [-250, 50] (for a gene, site 0 is not included), and negative ones are [-850,-550), [-550,-250), (50,350] and (350,650] (Fig. 1(a)). It is obvious that negative ones can be divided into up- and down-stream negative samples.

For a test sequence with TSS unknown, we slide a 300 bp window by 1 bp step to get samples with the same length as that of the training samples (Fig. 1(b)).

Several features were extracted for all the samples, including positive and negative samples in training set

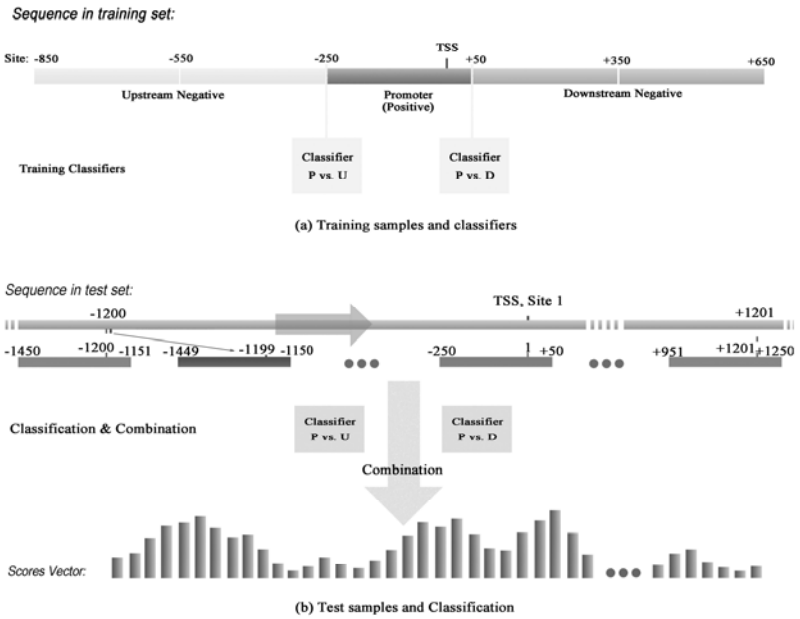

Fig. 1. Samples and Classifiers. (a) Preparation of the samples in the training set. The promoter region from 250bp upstream to 50bp downstream of the annotated TSS is taken as the positive sample, while the other upstream and downstream sequences are taken as negative samples. The upstream negative samples (U) are used with the positive samples (P) to train the classifier P vs. U, and the downstream negative samples (D) are used with the positive samples (P) to train the classifier P vs. D. (b) The classification on the samples in test set. A window of 300-bp is scanned along the DNA sequence to be tested at a 1-bp step, forming the test sets. There are 2401 samples from each sequence. The two classifiers (P vs. U and P vs. D) are applied on each of the samples and the outputs of them are combined, generating a series of prediction scores at each position of the sequence. Post-processing then is used for making the final decisions based on the scores (see text and Fig. 2 and 3).

and the samples in test set. Before feature selection, there are 210+ numeric features (see Table 1 in Zhao *et al*'s paper[17]). To make the analysis of the features easier, we categorize them as follows: (i) basic sequence features; (ii) mechanical properties; (iii) motif features. The basic sequence features include scores of core promoter elements (TATA, Inr, CCAAT-box and GC-box), the frequencies of 1-mer or 2-mer related to C or/and G, and the scores from 3rd order Markov chain modeling. The mechanical properties capture the characteristics of the energy and flexibility profiles around TSS, and the distance and correlation values are computed with different sequence lengths and smoothing window lengths. The motif features are generated by *featuretab*, part of CREAD suite of sequence analysis tools[20]. The motif weight matrices are from TRANSFAC[21] and maximal scores of the weight matrices for TFBSs are used as the motif features. There are about 66 features in this category.

If too few features are used to classify promoter and non-promoter regions and to predict TSSs, the predictive power may be very low. On the other hand, however, if the number of the features is too large, the noise may go up and the predictive power would come down. Hence, feature selection is one of the most important steps of the whole system for TSS prediction. The multi-objective optimization method AMOSA is implemented in our system for effective feature selection.

2.3. Classification Strategy

In our proposed TSS prediction system, we use Fisher's linear discriminant analysis or LDA to build the basal classifiers. LDA, originally developed in 1936 by R.A. Fisher, is a classic classification method. The main idea of Fisher's linear discriminant is to project data, usually in a high-dimensional space, onto a direction so that the distance between the means of the two classes is maximized while the variance within each class is minimized. Thereafter, the classification becomes a one dimension problem and classification can be done by a proper threshold cut-off.[22] As LDA considers all samples in the projection, it has been shown to be robust to noise and often performs well in many applications.

LDA models are built with the features selected, and their performance is used as the guide in the feature selection procedure.

2.4. Feature Selection with AMOSA

Among the 210 features to be used for predicting TSSs in non-CpG related promoters, there might be ones which contribute little to the classification but bring in more noises. In this article, a state of the AMOSA denotes the features that are selected for classification. LDA classifiers are built with only the selected features. Three objectives, namely, sensitivity (*Sn*), positive predictive value (*PPV*) and Pearson correlation coefficient (*CC*), are used to evaluate the performance of the LDA classifiers with the selected features. They are computed using 10-fold cross validation as:

$$Sn = \frac{TP}{TP + FN} \tag{2}$$

$$PPV = \frac{TP}{TP + FP} \tag{3}$$

$$CC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4}$$

where *TP*, *TN, FN*, and *FP* are the numbers of true positives, true negative, false negatives, and false positives, respectively.

We consider the three objectives are equally important, where *Sn* controls false negatives, *PPV* limits false positives, and *CC* balances classification results. However, the traditional optimization methods, which can only optimize one objective, could not deal with this problem. The multi-objective optimization method AMOSA is therefore implemented in order to solve the three-objective optimization problem. For multi-objective optimization methods usually allow multiple solutions, we get several sets of selected features in each experiment.

2.5. Prediction System

Our whole prediction system contains two phases. At first, AMOSA is combined with LDA as a feature selection and classification system. Thereafter, post processing is performed to integrate classification scores and get prediction results.

Fig. 2 shows the flow chart of the whole prediction system, which contains two phases: feature selection & classification, and post processing. Note that there are two symmetrical parts in Phase I. This is because we train two types of models (P vs. U and P vs. D) using

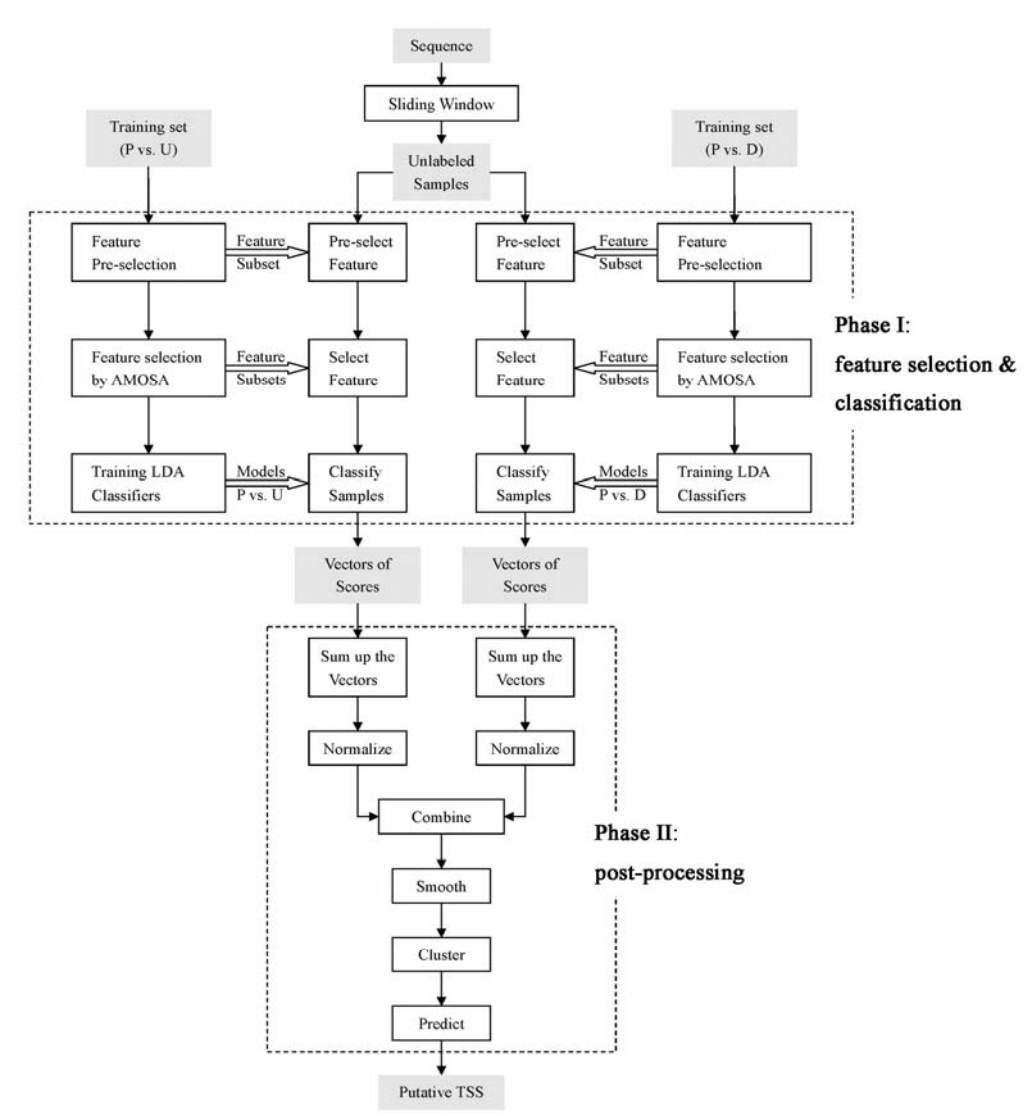

Fig. 2. Flow Chart for the Prediction System.

three categories of training samples: Positives (P), Upstream negatives (U) and Downstream negatives (D), and thus classify the test samples twice. We take the left side for example to explain the procedures in Phase I. First of all, we pre-select the features (removing those features that have almost same values for both P and U samples). Then, we come to the key step, where AMOSA is used to select the features and LDA to classify the samples. Let n denote the number of the feature subsets output by AMOSA. Correspondingly, we get n LDA classifiers trained. For an input sequence, we slide a 300-bp-width window with 1 bp step to generate the test samples. After we use the n feature subsets and the n classifiers to classify the sequential samples, n vectors of classification scores are output. Let l denote the length of the vectors. Treating the n vectors equally, we sum up the n vectors to get the summed scores S_i , i = 1,2,…,l, and then normalize them by:

$$S_i^{'} = -1.0 + \frac{S_i - S_{\min}}{S_{\max} - S_{\min}} \times 2 \qquad (5)$$

where S_i , i = 1,2,…,l, denote the initial summed scores, $S_i^{'}$, i = 1,2,…,l, the ones after normalization, and $S_{\max}$, $S_{\min}$ are the maximum and minimum of all the initial summed scores, respectively. The right-hand side of

Phase I, the classifier for P vs. D is also implemented similarly.

We consider the two vectors output by the two symmetrical classifiers to contribute equally for the prediction, and add the two scores for the prediction. We use Eq. (6) to smooth the sum:

$$S_i^{**} = \frac{\sum_{j=\max(1,i-50)}^{\min(i+50,l)} S_j^* \times \exp(-(j-i)^2/5000)}{\sum_{j=\max(1,i-50)}^{\min(i+50,l)} \exp(-(j-i)^2/5000)} \quad (6)$$

where S_j^* , j = 1,2,…,l, denote the scores before smoothed and S_i^{**} , i = 1,2,…,l, denote the smoothed scores. We choose a RBF (Radial Basis Function) window with width of ± 50 rather than a flat window because the influence decays with the distance.

We cluster the sites with the scores larger than a given threshold (the thresholds can be chosen in [-0.2, 0.25], and the corresponding results with the varying thresholds are shown in Fig.4 as a *PPV-Sn* curve.), and in a cluster the site with maximum score is the putative TSS (Fig. 3).

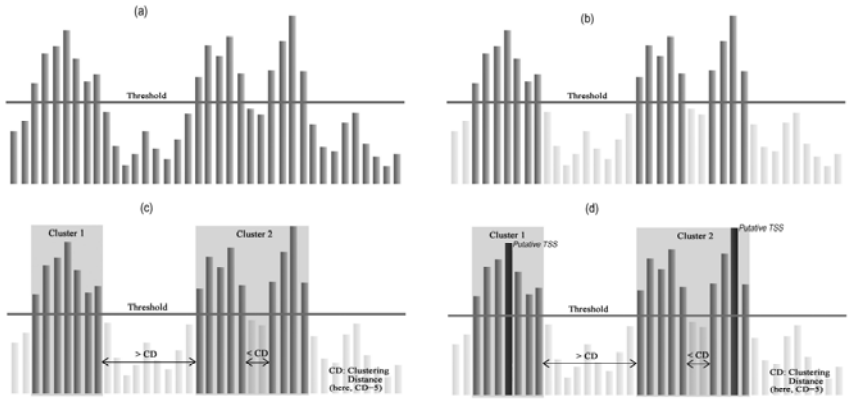

Fig. 3. Last Steps of Post Processing.
(a) The vector of bars stands for the vector of scores, high bar indicating high score. A threshold is set. (b) The bars under the threshold are removed. (c) The remaining bars within certain distance are clustered as one cluster. (d) The site with maximum score in a cluster is output as the putative TSS.

2.6. Cross Validation

We use a 5-fold cross validation to evaluate the performance of our proposed TSS prediction system. We divide the 1,271 sequences from DBTSS into 5 parts, each time 4 parts of the five and all the 299 sequences from EPD forming a training set and the remaining part forming the corresponding test set. Thus, 5 pairs of training and test sets are prepared.

For each sequence (10kbp length, with annotated TSS located at site 1) in test set, we slide a window to get 300-bp-length samples from site -1200 to 1201. Therefore, there are 2401 samples from each sequence in test set.

True positive (*TP*), false positive (*FP*) and false negative (*FN*) are defined. The putative TSS which is located within 50 bp from any of the annotated TSS is considered as a TP, otherwise an FP. If there is no predicted TSS in the $\pm$ 50bp region of an annotated TSS, this counts as one FN. Two important criteria, *Sn* and *PPV*, are as defined in Eq. (2) and (3).

Note that for one annotated TSS there is either a TP or a FN, and for one sequence in our data there is only one annotated TSS, so the sum of *TP* and *FN* equals to #sequences. The Eq. (2) can be simplified as:

$$Sn = \frac{TP}{\#\text{sequences}} \quad (7)$$

2.7. Other TSS Prediction Methods

In comparison study, we compare our newly proposed TSS prediction system with three other most effective and publicly available methods[5, 17]: McPromoter[23], Eponine[24], and CoreBoost[17]. McPromoter combines DNA sequence likelihoods and profiles of physical properties as features, and a neural network is implemented for predicting TSSs. Eponine uses a set of

weight matrices for extracting features, and applies a hybrid relevance vector machine method to build the prediction model. And CoreBoost proposes a feature selection strategy which can select useful features among basic sequence information, mechanical properties and motif features, and then a boosting technique is implemented to predict TSSs. We got McPromoter software from its authors and downloaded Eponine from its website[25]. We ran the programs of the three methods on our local computer.

3. RESULTS AND DISCUSSION

3.1. Performance of AMOSA

The performance of our system with AMOSA embedded is first compared with that without feature selection, i.e., using all available features for the prediction. Table 1 shows the *TP*, *FP, Sn* and *PPV* values comparing the two methods due to different parameters. From table 1, we can see that the system using all the features to predict TSSs does not perform as good as the one having AMOSA feature selection embedded with the same parameters. Therefore, it can be concluded that, the feature selection method with AMOSA implemented is effective, making the prediction system achieve higher *Sn* and *PPV* even using fewer features.

Table 1. Prediction results between all features and selected features using AMOSA.

Parameters[a]	Method	TP	FP	Sn(%)	PPV(%)
0.10/1000	All Features	283	887	22.3	24.2
	AMOSA	319	886	25. 1	26.5
0.10/2000	All Features	281	832	22.1	25.2
	AMOSA	316	813	24.9	28.0
0.00/1000	All Features	304	1004	23.9	23.2
	AMOSA	335	1014	26.4	24.8
0.00/2000	All Features	303	939	23.8	24.4
	AMOSA	333	932	26.2	26.3

[a]The parameters are the classification scores threshold (e.g. 0.10) and clustering distance (e.g. 1000).

3.2. Features for TSS Prediction

Among all the original features (more than 210), only about 90 features are selected each time on an average during the 5-fold cross validation. It is not surprising that not all the selected features are the same for different training sets, but there are quite a few features which are selected almost all the times. We count the number of times each feature is selected during the 5-fold cross validation experiment.

Table 2 lists the top features selected for model P vs. U and P vs. D separately while Table 3 for both models. From the tables, we can see that the known core promoter elements play great roles in the prediction, for their weight scores such as TATA90, Inr90, GCbox90 appear in the top rank. Log-likelihood ratios from 3rd order Markov chain (denoted by MC) together with some energy/flexibility characters and motif features are also among the top features. Moreover, for the two LDA classifiers (P vs. U and P vs. D), the selected features are different. For example, the weighted score of the 7 mer TATA box (denoted by TATA-7) has a lager possibility to appear in the classifier P vs. U, while the weighted score based on Bucher *et al*[26] for Inr (denoted by Inr90) is more frequently selected in the classifier P vs. D. That's why we train the two classifiers (P vs. U and P vs. D) separately. The LOGOs for the motifs mentioned in table 2 or 3 are shown in table 4.

As to the three categories of the features, namely the basic sequence features, the mechanical properties, and the motif features, the proportion of the features selected are not the same. We call the ratio of #(selected features) to #(total features) in each category as the feature selection ratio. From table 5, we can see that the selection ratio for the motif features is very low, even less than half of the other two. It indicates that the TFBS motif weight matrices seem to have less information in predicting TSSs in non-CpG related promoters. However, the reason may also be that there are many different motifs, playing different roles in different promoters (e.g. tissue-specificity). If we group the motifs according to their functions, their contribution in the prediction might be more and the performance might be further improved. Besides the motif features, there are also redundancies existed in the other two categories.

3.3. Comparison with Other Methods

We compare our system with three other effective and publicly available methods: McPromoter[23], Eponine[24], and CoreBoost[17]. Five-fold cross validation is used to evaluate the performance of our prediction system.

Table 2. Top selected features using AMOSA for models P vs. U and P vs. D. The total number of subsets of the selected features for the model P vs. U in 5-fold experiment is 43, while that for P vs. D is 46.

P vs. U	count	ratio%	P vs. D	count	ratio%
MC	43	100.00	aveTATA.flex	46	100.00
corr.flex.150.1000	43	100.00	aveTSS.flex	46	100.00
corr.eng.500.250	43	100.00	Inr90	46	100.00
corr.eng.1000.1300	43	100.00	MC	46	100.00
V$ELK1_02.pos	43	100.00	corr.flex.5.1300	46	100.00
V$HNF1_Q6.pos	43	100.00	corr.eng.5.500	46	100.00
V$MYC_Q2.pos	41	95.35	corr.eng.500.250	46	100.00
V$PAX6_01.pos	41	95.35	CCAAT90	45	97.83
aveTATA.flex	40	93.02	corr.eng.1000.1300	45	97.83
TSSdiffNew1.eng	39	90.70	V$CDPCR1_01.pos	44	95.65
eud.eng.5.250	39	90.70	TSSdiffNew2.eng	41	89.13
TATA90	38	88.37	TATACCAAT90	40	86.96
corr.flex.500.1000	38	88.37	aveTATA.eng	39	84.78
TATAdiffNew3.flex	37	86.05	TATA90	39	84.78
Density	37	86.05	TATAGCbox90.dist	39	84.78
eud.flex.1000.1000	37	86.05	mc1stmc	39	84.78

Table 3. Top selected features using AMOSA in common. The total number of subsets of the selected features for both models in 5-fold experiment is 89.

	P vs. U		P vs. D		Both	
	count	ratio%	count	ratio%	count	ratio%
corr.eng.500.250	43	100.00	46	100.00	89	100.00
MC	43	100.00	46	100.00	89	100.00
corr.eng.1000.1300	43	100.00	45	97.83	88	98.88
aveTATA.flex	40	93.02	46	100.00	86	96.63
corr.flex.5.1300	35	81.40	46	100.00	81	91.01
V$ELK1_02.pos	43	100.00	36	78.26	79	88.76
TATA90	38	88.37	39	84.78	77	86.52
V$HNF1_Q6.pos	43	100.00	34	73.91	77	86.52
eud.eng.5.250	39	90.70	35	76.09	74	83.15
V$CDPCR1_01.pos	30	69.77	44	95.65	74	83.15
Inr90	27	62.79	46	100.00	73	82.02
TSSdiffNew1.eng	39	90.70	34	73.91	73	82.02
V$PAX6_01.pos	41	95.35	32	69.57	73	82.02
Density	37	86.05	35	76.09	72	80.90

Fig. 4 depicts the plot of *PPV* vs. *Sn* to show the comparison results. Those different points for one method are due to the different parameters. The asterisks and the circles are for Eponine and McPromoter, respectively. The solid and dashed curves are for CoreBoost. And the squares and the triangles are for our system with different clustering distances, where the different points with the same symbol are for the different score thresholds from -0.20 (bottom right) to 0.22 (top left). It is clear that our prediction system with clustering distance 2000bp outperforms Eponine, McPromoter and CoreBoost. The score threshold 0.03 achieving 26.0% *Sn* and 26.5% *PPV* is chosen as the default threshold in our prediction system.

Table 4. LOGOs of motifs listed in top features.

Motif	LOGO	Note
V$ELK1_02		
V$HNF1_Q6		P vs. U
V$MYC_Q2		P vs. U
V$PAX6_01		P vs. U
V$NFY_Q6_01		P vs. U
V$PAX_Q6		P vs. U
V$CDPCR1_01		
V$HNF4_Q6_01		

Table 5. The feature selection ratios for the different feature categories.

	sequence features		mechanical properties		motif features		total	
	ave. #	ratio(%)	ave. #	ratio(%)	ave. #	ratio(%)	ave. #	ratio(%)
overall	33	100	114	100	66	100	213	100
P vs. U	15.4	46.7	55.2	48.4	15.4	*23.3*	86.0	40.4
P vs. D	16.6	50.3	57.6	50.5	13.6	*20.6*	87.8	41.2

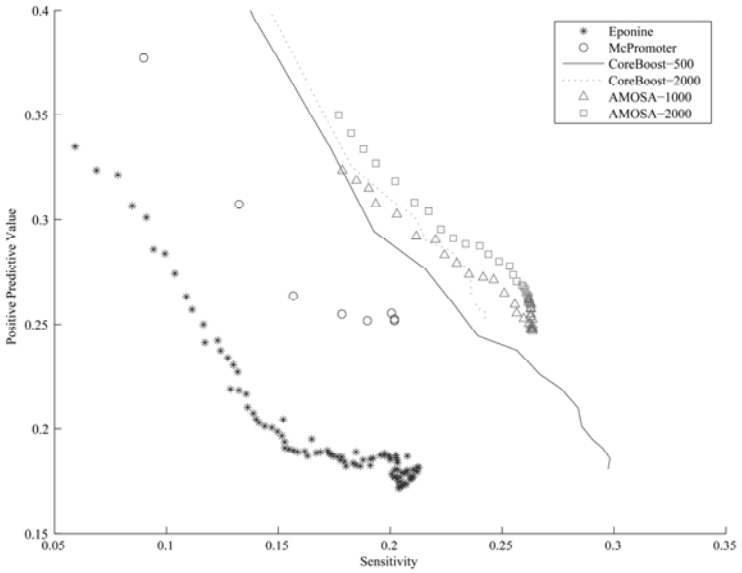

Fig. 4. Positive predictive value vs. sensitivity.
The asterisks are for Eponine, which is the default result. The circles are for McPromoter with the default clustering distance 2000bp. The solid and the dashed curves are for CoreBoost, with the solid curve for clustering scores within 500bp and the dashed one for 2000bp. The squares and the triangles are for our system with AMOSA embedded, of which the clustering distances are 2000bp and 1000bp, respectively.

3.4. Discussion

In this paper, we have proposed a new system based on AMOSA feature selection to predict TSSs in non-CpG related human promoters. Firstly, we generate features from the sequence characteristics, the mechanical properties and the TFBS motif scores. Thereafter, we implement AMOSA to select features to train LDA models. And finally, we use the LDA classification scores followed by some post-processing to predict TSSs. As a result, relatively higher prediction *Sn* and *PPV* are achieved when comparing to the other existing methods.

It can be observed that the performance of all these methods still has a lot of room for improvement. This reflects the complexity of the problem and the insufficient understanding of the underlying biology. However, considering that we are trying to predict a single TSS with 50 bp resolution *de novo* from a long genomic DNA sequence, such moderate sensitivity and specificity are still welcome. The results can be also useful, in conjunction with other gene prediction tools, for helping biologists to prioritizing their experimental targets. Further improvement will likely require more detailed information on chromatin state and tissue/stage-specificity of the promoter sequences.

Acknowledgments

This work is supported in part by NSFC grants 30625012 and 60540420569, the National Basic Research Program of China 2004CB518605, and the Changjiang Professorship Award of China. Additional support is provided by an award from the Dart Neurogenomics Alliance and by HG001696 grant from NIH.

References

1. Zhang MQ. Identification of human gene core promoters in silico. *Genome Res* 1998; **8**: 319-326.
2. Werner T. The state of the art of mammalian promoter recognition. *Brief Bioinform* 2003; **4**: 22-30.
3. Davuluri RV, Grosse I, Zhang MQ. Computational identification of promoters and first exons in the human genome. *Nat Genet* 2001; **29**: 412-417.
4. Ioshikhes IP, Zhang MQ. Large-scale human promoter mapping using CpG islands. *Nat Genet* 2000; **26**: 61-63.
5. Bajic VB, Tan SL, Suzuki Y, Sugano S. Promoter prediction analysis on the whole human genome. *Nat Biotechnol.* 2004; **22**: 1467-1473.
6. Oh IS, Lee JS, Moon BR. Hybrid Genetic Algorithms for Feature Selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2004; **26**: 1424 - 1437.
7. Raymer ML, Punch WF, Goodman ED, Kuhn LA, Jain AK. Dimensionality reduction using genetic algorithms. *IEEE Transactions on Evolutionary Computation* 2000; **4**: 164-171.
8. Bandyopadhyay S, Saha S. Simultaneous Optimization of Multiple Objectives: New Methods and Applications. *In the Proceedings of the Eighth International Conference on Humans and Computers* 2005; 159-165.
9. Bandyopadhyay S, Saha S, Maulik U, Deb K. A Simulated Annealing Based Multi-objective Optimization Algorithm: AMOSA. *IEEE Transactions on Evolutionary Computation* 2007; Submitted.
10. Coello CAC. A Comprehensive Survey of Evolutionary-Based Multiobjective Optimization Techniques. *Knowledge and Information Systems* 1999; **1**: 129-156.
11. Deb K. *Multi-Objective Optimization Using Evolutionary Algorithms* John Wiley and Sons, Ltd., England. 2001.
12. Veldhuizen DAV, Lamont GB. Multiobjective Evolutionary Algorithms: Analyzing the State-of-the-Art. *Evolutionary Computation* 2000; **8**: 125-147.
13. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics* 1953; **21**: 1087-1092.
14. German S, German D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1984; **6**: 721-741.
15. Smith KI, Everson RM, Fieldsend JE, Murphy C, Misra R. Dominance-Based Multi-Objective Simulated Annealing. *IEEE Trans. on Evolutionary Computation* 2005; Submitted.
16. Deb K, Pratap A, Agarwal S, Meyarivan T. A Fast Elitist Multi-Objective Genetic Algorithm: NSGA-II. *IEEE Trans. on Evolutionary Computation* 2002; **6**: 182-197.
17. Zhao X, Xuan Z, Zhang MQ. Boosting with stumps for predicting transcription start sites. *Genome Biol* 2007; **8**: R17.

18. Cavin PR, Junier T, Bucher P. The Eukaryotic Promoter Database EPD *Nucleic Acids Res* 1998; **26**: 353-357.
19. Suzuki Y, Yamashita R, Sugano S, Nakai K. DBTSS, DataBase of Transcriptional Start Sites: progress report 2004. *Nucleic Acids Res* 2004; **32**: D78-D81.
20. The Comprehensive Regulatory Element Analysis and Discovery (CREAD) suite. http://rulai.cshl.edu/cread.
21. Wingender E, Chen X, Hehl R, Karas H, Liebich I, Matys V, Meinhardt T, Prüß M, Reuter I, Schacherer F. TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res* 2000; **28**: 316-319.
22. Duda RO, Hart PE, Stork DG. *Pattern Classification (Second Edition)*. John Wiley & Sons, Inc., England. 2001: 117-121.
23. Ohler U, Niemann H, Liao G, Rubin GM. Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition. *Bioinformatics* 2001; **17**: S199-S206.
24. Down TA, Hubbard TJP. Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Res* 2002; **12**: 458-461.
25. Eponine. http://www.sanger.ac.uk/Users/td2/eponine/.
26. Bucher P. Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J Mol Biol* 1990; **212**: 563-578.