Novel gene discovery in the human malaria parasite using nucleosome positioning data

N. Pokhriyal, Nadia Ponts, E. Y. Harris, K. G. Le Roch, S. Lonardi

Published in final edited form as:
Comput Syst Bioinformatics Conf. 2010 August ; 9: 124–135.

Novel Gene Discovery in the Human Malaria Parasite using Nucleosome Positioning Data

N. Pokhriyal[1], **N. Ponts**[2], **E. Y. Harris**[1], **K. G. Le Roch**[2], and **S. Lonardi**[1,*]

[1]Department of Computer Science and Engineering, University of California, Riverside, CA 92521, USA

[2]Department of Cell Biology and Neuroscience, University of California, Riverside, CA 92521, USA

Abstract

Recent genome-wide studies on nucleosome positioning in model organisms have shown strong evidence that nucleosome landscapes in the proximity of protein-coding genes exhibit regular characteristic patterns. Here, we propose a computational framework to discover novel genes in the human malaria parasite genome *P. falciparum* using nucleosome positioning inferred from MAINE-seq data. We rely on a classifier trained on the nucleosome landscape profiles of experimentally verified genes, and then used to discover new genes (without considering the primary DNA sequence). Cross-validation experiments show that our classifier is very accurate. About two thirds of the locations reported by the classifier match experimentally determined expressed sequence tags in GenBank, for which no gene has been annotated in the human malaria parasite.

1. INTRODUCTION

Gene Discovery

The first step after sequencing and assembling a new genome involves annotating the genomic DNA with the predicted location of protein-coding genes. For this reason, gene discovery has been intensively studied in computational biology (see, *e.g.*, http://www.nslij-genetics.org/gene/ for an extensive bibliography). Existing techniques are based on probabilistic models (also called *ab initio*), homology with other species, or take advantage of expressed sequence tags (ESTs).

Methods using probabilistic models allow one to distinguish between the genic and non-genic regions using the structural features of genes based on the primary DNA sequence. Hidden Markov models (HMMs)[27] are perhaps the most widely used: several HMM-based gene finding tools have been developed, including GENSCAN[9], GlimmerHMM[22], and Genie[17]. Other probabilistic models such as Integrated Markov models[10], Conditional Random Fields[7], and Dynamic Bayesian networks[21], have also been employed. The key strength of probabilistic model-based approaches is that they can be efficiently trained on

*Corresponding author. stelo@cs.ucr.edu.

relatively small labeled training sets to extract the relevant features that characterize genes. Such techniques, however, often suffer from a high false positive rate because they have to model the generative probability distribution for the genic regions. In addition, these techniques perform poorly in identifying genes which are not well represented in the training data. Gene predictors based on inter-species homology make use of aligned DNA sequence from other genomes: alignments can increase predictive accuracy since protein-coding genes exhibit distinctive patterns of conservation. Rosetta[6] and Cem[4] are among the earliest methods for predicting human genes using alignments. The third category of techniques exploits collections of previously sequenced ESTs, i.e., databases of portions of transcribed sequences[25]. Several gene discovery techniques[2, 31, 11] and tools such as ProCrustes[12], GeneWise and GenomWise[8], align ESTs (allowing splicing) to the original genome to locate expressed genes. The major drawback of EST-based techniques is that they can only discover genes for which representative ESTs exist in the database, thus they miss genes which are rarely expressed.

All previously described techniques use the primary DNA sequence to annotate the location of the genes. In this paper we investigate whether nucleosome positioning data alone can be used in genome annotation processes. To our knowledge, our study is the first one that takes advantage of nucleosome positioning data for gene discovery.

Nucleosomes

In eukaryotic cells, genomic DNA organizes with various proteins in a complex structure called *chromatin*. One of the main function of chromatin is to package DNA into a smaller volume but it also regulates all DNA related processes *e.g.*, DNA replication, DNA repair, and transcription.

The fundamental unit of chromatin is the *nucleosome*, composed of 146±1 base pairs of DNA wrapped 1.65 turns around a protein complex of eight histones (Figure 1). Chromatin exists with different degrees of condensation, from *euchromatin* (relaxed) to *heterochromatin* (packed). This degree of packaging directly depends on nucleosome distribution and density. As a consequence, nucleosome occupancy directly affects a variety of cellular and metabolic processes including transcription (gene expression). The more the chromatin is condensed, the harder it is for transcription factors and other proteins to access the DNA and carry out their tasks.

Most of eukaryotic genomic DNA is organized into nucleosomes. For example, it is estimated that 75–90% of the human DNA is wrapped into nucleosomes [5, 28]. Nucleosomes tend to be regularly spaced along the genome, at a distance from each other that is organism-specific: about 18 bp in yeast[18, 23, 29], about 28 bp in drosophila[24] and *C. elegans*[32], and about 38 bp in humans[5, 28]. Significantly longer spacers or *nucleosome-free regions* (NFRs) tend to occur at the beginning and the end of protein-coding genes.

A handful of experimental techniques have been developed for genome-wide mapping of nucleosomes. One can isolate genomic regions that are bound to histones typically via chromatin immunoprecipitation (chIP) or MNase-mediated purification of mononucleosomes (MAINE) or can enrich for genomic regions that are free of nucleosomes

typically via formaldehyde-assisted isolation of regulatory elements to extract protein-free DNA (FAIRE). Then, tiling microarrays (chip) or high-throughput sequencing (seq) are used to detect and identify the isolated DNA. Finally, software tools are used to process information from the tiling array or sequencing data and generate a map of nucleosome positioning. Previous chIP-chip experiments generated genome-wide maps of nucleosome occupancy in *S. cerevisiae*[19, 18, 30, 16, 35], *C. elegans*[15], *P. falciparum*[34] and human[28] in various cell types and under a variety of physiological perturbations. More recently high throughput sequencing was used to produce nucleosome maps for *S. cerevisiae*[3, 29, 23], *C. elegans*[32], *P. falciparum*[26] and drosophila[24] at single-base resolution.

Proposed approach

Recent nucleosome positioning data generated from the studies mentioned above strongly suggest that nucleosome occupancy landscapes in the proximity of genes exhibit regular patterns. Figure 2 illustrates the nucleosome occupancy landscape for typical genes in the human malaria parasite. Two nucleosome-free regions (NFRs) are located immediately upstream of the transcription start site (TSS) or the beginning of the gene and downstream of the end of each gene. A definite pattern consisting of periodic and regularly spaced nucleosome positioning is observed at the beginning of protein-coding genes in *S. cerevisiae*[23, 35, 14], *C. elegans*[29], *D. melanogaster*[24] and *P. falciparum*[26].

Such a pattern of nucleosome distribution raises an interesting question: can we exploit the correlation between the nucleosome profiles in the vicinity of protein-coding genes to detect novel genes? More specifically, do nucleosome profiles contain enough information to allow one to predict the boundaries of gene models? To study these questions, we propose to test a binary classifier trained on the nucleosome profiles of experimentally verified genes. While the focus of this paper is on the human malaria parasite *P. falciparum* our method can be applied to any other organisms for which nucleosome positioning data with sufficient resolution are available. First, we will show that our classifier is very accurate in cross-validation experiments. Then, we will report that our classifier can discover putative novel genes in *P. falciparum* without considering the primary DNA sequence. About two third of the locations reported by the classifier as potential novel genes match ESTs in GenBank. This suggests that we might have successfully identified new genes in the human malaria parasite. A validation of these putative novel genes is currently under way.

2. DATA SOURCE, TRAINING SETS, AND LEARNING

Data source

Nucleosome positioning data at single-nucleotide resolution were previously generated by micrococcal nuclease digestion of *P. falciparum* genomic DNA followed by high-throughput sequencing (MAINE-seq)[26]. *P. falciparum*'s genome consists of 14 chromosomes for a total of about 24 Mb. We generated seven distinct time series by counting the number of sequenced reads that were mapped at each position along each chromosome throughout the *P. falciparum* intraerythrocytic infection cycle sampled at six intervals (see Ref. 26 for more details). These time series (or *occupancy profiles*) vary over time according to the nucleosomal rearrangements that occur in the malaria parasite. Since

our purpose is gene discovery, regardless of any biologically relevant regulation process, we created a general profile of nucleosome occupancy by averaging the seven profiles at every position in the genome.

Training Sets

The average general profile of nucleosome occupancy was used as input data to our classifier. The classifier was trained only on the high-confidence genes in the genome. In *P. falciparum*, only about a third of the genes are confidently annotated (hereafter called "confirmed", whereas the non-confirmed will be called "others"). We also removed the telomere regions from the training set because the occupancy profiles in these regions were not reliable due to low sequencing depth. Out of the total of 5460 genes in *P. falciparum*, 1995 were classified as "confirmed", and 3465 as "others".

Averaged occupancy profiles representing nucleosome distributions are real-valued time series, denoted in the rest of the paper by $\mathbf{S}^{(i)}$, where i is the chromosome number ($i = 1, 2, \ldots, 14$ for *P. falciparum*). We were also given a set of g annotated genes ("confirmed" and "others"), which is represented by a set $\mathscr{G}^{(i)} = \{(s_1, e_1), (s_2, e_2), \ldots, (s_g, e_g)\}$ containing start and end positions of each gene on chromosome i. To simplify the notation, hereafter we will drop the superscript i from $\mathbf{S}$ and $\mathscr{G}$.

To apply traditional classification schemes on these time series, we extracted fixed length subsequences of length w by sliding a window along $\mathbf{S}$ with a sliding step of $h \geq 1$ base. The resulting number of windows is $n = \lceil(|\mathbf{S}| - w + 1)/h\rceil$. Each time series $\mathbf{S}$ was thus converted into a set of windows $\mathscr{D} = D_1, D_2, \ldots, D_n$, where each window D_i is a vector of length w. We used $a_i = (i - 1)h + 1$ and $b_i = (i - 1)h + w$ to denote the start and the end of window D_i. Inside each window $[a_i, b_i]$ we identified the central region of length m called *margin*, which has coordinates $\left[a_i + \frac{w}{2} - \frac{m}{2}, a_i + \frac{w}{2} + \frac{m}{2}\right]$. Figure 3 illustrates a window of length 1000 bp, centered at the transcription start site, with a margin of 50 bp. Specific choice for w and m will be discussed later in the paper. The parameter h was set to one base.

After extracting windows from $\mathbf{S}$, we assigned a label to each window depending on the presence or absence of a gene in it. The labeling scheme described below is designed to train a classifier to recognize the start of the genes. While the rest of the discussion focuses on the start of the genes, a similar labeling scheme and training was used to detect the end of the genes (see end of Section 3.2). The labeling differentiates between windows that correspond to "confirmed" genes as opposed to "other" type of genes. Each window D_i was assigned one of the following labels:

- $l_i = -1$, if D_i did not overlap any gene in $\mathscr{G}$ (either "confirmed" or "other")
- $l_i = 1$, if there was a "confirmed" gene $G_j \in \mathscr{G}$ such that the start of G_j was within the margin of window D_i
- $l_i = 2$, if there was a "confirmed" gene $G_j \in \mathscr{G}$ such that the start of G_j was within the window but not inside the margin of D_i

- $l_i = 3$, if there was a "confirmed" gene $G_j \in \mathcal{G}$ such that the window overlapped G_j (completely or partially) but it did not overlap the start of any gene
- $l_i = 4$, if there was an "other" gene $G_j \in \mathcal{G}$ such that the window overlapped G_j (completely or partially).

An example of windows labeling is illustrated in Figure 4.

Learning

We used labeled windows to train a binary classifier that discriminates between windows with label 1 (*positive* windows) and windows with label −1 (*negative* windows). Windows with labels 2, 3 and 4 were ignored for the training. A random sample of positive and negative windows was used for training. The choice of the sample size is discussed later.

While in principle any binary classifier could be used to discriminate between the two types of windows, the performance of different classifiers may vary due to the nature of the data. We tested a variety of classifiers and focused on two, namely logistic regression and Radial Basis Function networks. A logistic regression classifier models the posterior probabilities of the two classes using linear functions in the input[1]. A Radial Basis Function (RBF) is an artificial neural network that uses radial basis functions as activation functions[13].

3. EXPERIMENTAL RESULTS

In order to discover novel genes in *P. falciparum*, we extracted a set $\mathcal{T}$ of test windows from the time series **S** and assigned labels as discussed earlier. Each test window T_i was then tested using the classifier. The predicted label $\hat{l_i}$ and the associated confidence score λ_i was used to assign a secondary label β_i as follows:

$$\beta_i = \begin{cases} 1 & \text{if} l_i = -1 \text{and} \hat{l_i} = 1 \text{and} \lambda_i \geq \delta \\ 0 & \text{otherwise} \end{cases}$$

where δ is a user defined threshold. If $\beta_i = 1$, test window T_i was not known to contain a gene but the classifier predicted that it did, with confidence δ or higher. Windows for which $\beta_i = 1$ are called *candidate* windows. The entire test set of windows $\mathcal{T}$ was thus processed and the set of candidate windows was identified. Contiguous candidate windows were merged, and any set that contained at least *m* windows (the same value used for the margin width) became a *candidate segment*. Observe that all of these segments belong to the non-genic regions of the chromosome and do not overlap with any genic region. However, according to the classifier these segments are highly likely to contain the start of a "missing" gene. The parameter δ controls the overall "quality" of the segments in terms of the confidence of the classifier.

In order to evaluate the performance of our method, we employed first traditional metrics, such as accuracy, precision and recall in a cross-validation framework. This initial evaluation was essential to choose the classification algorithm and the critical parameters,

such as window size, margin width, sample size, etc. Then, we validited the candidate segments by searching them in EST databases stored in GenBank.

3.1. Cross-validation experiments

As said in Section 2, two classifiers were trained, one for each strand of a chromosome. We observed that the nucleosome occupancy profiles for genes located in the forward strand is characteristically different from the profiles for the reverse strand. However, the performances of the classifier for forward and negative strands are identical. A classifier trained using the forward genes is more likely to discover genes that are located in the forward strand and a classifier trained using the reverse genes is more likely to discover genes located in the reverse strand. However, a classifier trained on forward genes can discover some genes that are located in the reverse strand and vice versa. In general, both classifiers should be used to discover the complete set of genes.

We constructed two datasets of windows that were used interchangeably as training set or test set. One was obtained by extracting and labeling windows from the seven *odd* chromosomes (i.e., chromosome number 1, 3, 5, …), while the other was obtained from the seven *even* chromosomes of *P. falciparum*. For each choice of the parameters, we ran ten experiments. In five experiments we trained on random samples from the windows on the odd chromosomes, and tested on random samples of the even chromosomes. In the other five experiments, we switched training set and test set. For each choice of the parameters, we computed mean and standard deviation for the metrics discussed later in this section.

Table 1 shows how the logistic regression classifier performed in terms of predicting labels for all windows extracted from all fourteen chromosomes of the malaria parasite (window size $w = 1000$, margin width $m = 50$ and sample size 6000). We will discuss the impact of the parameters at the end of this section. Each subtable in Table 1 shows the results for each chromosome, where each row corresponds to the labels predicted by the classifier and each column corresponds to the actual labels. Cells in boldface correspond to predicted value of 1 and actual value of −1, which represent the percentage candidate windows, *i.e.*, potential new gene locations. Observe that this value is fairly constant across all chromosomes and conforms to our intuition that only a small percentage of the chromosome should contain new genes. The classifier was able to identify the negative windows (true label −1) correctly over 90% of times for almost all of the chromosomes. The classifier was also accurate in identifying the positive windows (true label 1) correctly (> 80% for majority of the chromosomes). The performance is worse in identifying windows with true label 2 as positive, but is still higher than 65% for most of the chromosomes. This is because these windows contain the start of a “confirmed” gene, but not within the margin, and the classifier is trained to identify only those windows that contain the start of the gene within the margin. The classifier was not able to identify windows that belonged to a genic region but did not contain the start of a gene (true label 3) or windows that belonged to the genic region of “other” genes (true label 4). This indicates that windows that contained the start of a “confirmed” gene were significantly different from the windows that corresponded to the non-start portions of “confirmed” genes, as well “other” genes.

Once the number of true positive (*tp*), true negative (*tn*), false positive (*fp*) and false negative (*fn*) is computed, the following metrics can be defined:

- **Recall** $\left(\frac{tp}{tp+fn}\right)$ indicates how many of the actually positive windows were predicted as positive by the classifier. Alternatively, one can also measure the recall for the negative class $\left(\frac{tn}{tn+fp}\right)$.
- **Precision** $\left(\frac{tp}{tp+fp}\right)$ indicates how many of the positive labeled windows are actually positive. Alternatively, one can also measure the precision for the negative class $\left(\frac{tn}{tn+fn}\right)$.
- **Accuracy** $\left(\frac{tp+tn}{tp+fp+tn+fn}\right)$ indicates the total number of windows that were correctly labeled by the classifier.

Often, there is an inverse relationship between these metrics, where it is possible to increase one at the cost of reducing the other. In the context of the problem of finding missing genes, we wanted to have the highest possible recall for **both** positive and negative class. For instance, when the classifier is trained on a subset of the chromosomes and then tested on remaining chromosomes, we expect all windows overlapping known genes to be detected as positive. Similarly, we wanted the recall on the negative class to be as high as possible. We wanted the smallest proportion of test windows which are negative to contain the start of an undiscovered gene. Due to these considerations we focused on the recall on both positive and the negative classes as the primary metric to evaluate the performance of the classifiers.

In the rest of this subsection, we report recall values on positive and negative classes for a variety of parameter choices (type of classifier, window size, margin width, and training set sample size).

Logistic Regression vs. RBF Networks—Figure 5 shows the performance of Logistic Regression and Radial Basis Function Network for different window sizes (when the margin was $m = 50$ and the sample size was 6000 windows). The relative performance of the two classifiers is similar for other settings of the parameters. We also experimented with other classifiers such as Support Vector Machines and k Nearest Neighbor classifier but their recall was significantly worse. Logistic regression has an additional parameter for the underlying ridge regression step: we experimentally determined that the recall peaked when the value of this parameter was about 10^{-8}.

The results show that logistic regression is superior to RBF in terms of recall for the positive as well as the negative windows. The difference in performance is more significant for negative windows, i.e., RBF identifies a higher proportion of candidate windows. This becomes an issue when testing on *all* windows extracted from a genome, because a large proportion of these windows will be declared candidate windows, and as a consequence a large number of windows will have to be verified for the presence of new genes, possibly

resulting in waste of resources. Since logistic regression maintains a low recall on negative class, only a few but high quality candidate windows will be generated which can be validated more efficiently. Hence we conclude that logistic regression is a better classifier than RBF in the context of the proposed framework for *P. falciparum*.

Training Sample Size—As mentioned earlier, we chose a random sample from the set of labeled windows as the training data. Figure 6 summarizes the relationship between the size of the sample and the performance of the logistic regression classifier. The results indicate that sample size does not have significant impact on the performance of the classifier. While the variance is higher for smaller sample sizes, the computational cost of training the classifier increases with the sample size. We determined that a good trade-off was a sample size of 6000 (3000 samples of each positive and negative class).

Observe that we used a balanced training set, but the test set is very imbalanced because the vast majority of the windows in a chromosome are not genes. The imbalance in the test set should be reflected in the training set if the objective was to maximize the convex combination of precision and recall with the same weight. However, here we are only interested in maximizing the recall equally well for both positive and negative classes, and we completely disregard precision. In fact, when we used an imbalanced training set, we observed an increase in the recall for class with the majority of samples, but a decrease in the minority class. In order to have high recall for both positive and negative classes we had to employ a balanced training set.

Window Size—Observe that while short windows might not be able to fully capture the context around the start of a gene in the nucleosome position data, long windows might result in increased computational complexity in dealing with high dimensional spaces.

Figure 7 summarize the performance of logistic regression for different window sizes w, when $m = 50$ and the sample size was 6000. The results indicate that the recall on the positive class improves as the window size increases, but the recall on the negative class is not as sensitive to the window size. For a window of 2000bp, the recall on both classes is good but the space and time requirements for training are also high. For $w = 1000$ the performance is reasonably good, and the space/time requirements are also manageable. Hence, for further experiments we fixed the window size to 1000 base pairs.

Margin Width—Figure 8 summarizes the performance of logistic regression for different choices of margin widths m, when the window size was 1000 base pairs and the sample size was 6000. The results in Figure 8 indicate that the performance of the logistic regression classifier does not change significantly when the margin width is changed.

3.2. Evaluating Candidate Segments

The traditional evaluation metrics are useful for quantitatively assessing the performance of different classifiers as well as studying the effect of different parameter settings on the overall performance. But such an evaluation does not fully answer the question: how effective is the proposed framework in discovering “missing” genes in the genome of *P. falciparum*? Recall that the final output is a set of candidate segments which are likely to

contain the start of a novel gene. The true performance of the proposed framework could be determined only if we knew how many of the candidate segments actually contain a gene. To validate our findings, we used BLAST to compare these candidate segments with known *expressed sequence tags* (ESTs) stored in GenBank. If the candidate segment matches known ESTs, this gives an independent evidence that the segment might indeed contain a new gene. At the time of writing, a subset of these segments are being validated in our wet lab.

Results obtained from querying GenBank EST collection with the candidate segments are summarized in Table 2. We considered only those candidate segments for which the corresponding candidate windows were predicted as positive with at least δ=95% confidence. The table shows that out of 223 segments identified by our method, 65% of the segments matched one or more ESTs. Interestingly, the majority of such segments matched ESTs for *P. falciparum*, but we also found several segments that matched ESTs for other malaria parasite species, like *P. berghei*, *P. yoelii*, *P. berghei*, and *P. vivax*. We also found several segments that matched ESTs from different organisms from unrelated genus such as *Neospora caninum* (coccidian parasite), *Vitis vinifera* (common grape vine), *Arachis hypogaea* (peanut), *Aplysia californica* (California sea slug), *Citrullus lanatus* (watermelon). The fact that there are so many plants matching these segments is remarkable since like most Apicomplexa, *P. falciparum* harbor a plastid similar to plant chloroplasts.

Identification of End of Genes—All the discussion so far has concentrated on the start of genes in *P. falciparum*. The same methodology can be applied to identify other characteristics associated with genes, such as end of genes and perhaps exon/intron boundary, as long as the location of the target characteristic exhibits a distinct nucleosome landscape. The proposed methodology can be easily extended to identify the end of genes. The experiments for identifying the end of genes were conducted on averaged MAINE-seq data sets, using logistic regression classifier with window size 1000, margin width 50, and sample size 6000.

The results obtained from querying the candidate segments using BLAST are summarized in Table 3. The table shows that out of 161 segments identified by our method, 70% of the segments matched with one or more existing ESTs. We observed that for many chromosomes if a segment identified as a potential start of a gene matched a known EST, there was almost always a corresponding segment obtained for the end of genes.

4. CONCLUSIONS

We have proposed a machine learning framework that exploits the relationship between nucleosome occupancy profiles and gene locations to discover missing genes, or to confirm gene annotations of the existing ones. We have shown the applicability of the framework in the context of the malaria parasite, and our method can be extended to any other eukaryotic organism, including humans, for which high-resolution nucleosome positioning data is available.

We have shown that nucleosome positioning data alone can indeed be used to predict the location of genes. Experimental results show that our technique is able to identify known genes in *P. falciparum* more than 80% of the times. Furthermore, two-thirds of the high quality segments reported by the classifier matched one or more existing ESTs. At the time of writing, some of these segments are being validated in our wet lab.

The key challenge for our approach is how to set the various parameters. Through experimental analysis we empirically estimated the best choices for parameters such as the window size, margin width, training sample size and the classification algorithm. It is very likely that these choices are specific to *P. falciparum* data.

We also attempted to extract informative features from the nucleosome occupancy profiles instead of training on the original data. We used Principal Component Analysis and Haar wavelet transform for feature reduction but the performance of the classifier did not improve. We still believe that other feature extraction techniques might be beneficial. In particular, more experimental evaluation needs to be done to correctly ascertain the effect of Haar decomposition. Finally, since there is evidence that nucleosome occupancy profiles exhibit periodicity in the proximity of the start of genes[33], we also extracted *auto-correlation* coefficients and used them as features to our classifiers. Again, preliminary results were not promising.

The methodology proposed in this paper is relatively straightforward yet it has proved to be successful in identifying potential new genes. Although only wet lab experiments will be able fully validate our predictions, the fact that our segments match to sequenced ESTs provides strong evidence. The next step would entail incorporating our classifier with a sequence-based gene discovery tool that can accommodate additional evidence sources (e.g., Evigan[20]).

References

1. Agresti A. Categorical data analysis. Wiley-Interscience. 2002
2. Ajioka JW, Boothroyd JC, Brunk BP, Hehl A, Hillier L, Manger ID, Marra M, Overton GC, Roos DS, Wan KL, Waterston R, Sibley LD. Gene discovery by EST sequencing in Toxoplasma gondii reveals sequences restricted to the Apicomplexa. Genome Research. 1998; 8(1):18–28. [PubMed: 9445484]
3. Albert I, Mavrich TN, Tomsho LP, Qi J, Zanton SJ, Schuster SC, Pugh BF. Translational and rotational settings of H2A.Z nucleosomes across the *Saccharomyces cerevisiae* genome. Nature. 2007 Mar; 446(7135):572–576. [PubMed: 17392789]
4. Bafna, V.; Huson, DH. Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology. AAAI Press; 2000. The conserved exon method for gene finding; p. 3-12.
5. Barski A, Cuddapah S, Cui K, Roh T, Schones D, Wang Z, Wei G, Chepelev I, Zhao K. High-resolution profiling of histone methylations in the human genome. Cell. 2007 May; 129(4):823–837. [PubMed: 17512414]
6. Batzoglou S, Pachter L, Mesirov JP, Berger B, Lander ES. Human and Mouse Gene Structure: Comparative Analysis and Application to Exon Prediction. Genome Research. 2000; 10(7):950–958. [PubMed: 10899144]
7. Bernal A, Crammer K, Hatzigeorgiou A, Pereira F. Global discriminative learning for higher-accuracy computational gene prediction. PLoS Comput Biol. 2007 Mar.3(3):e54. [PubMed: 17367206]

8. Birney E, Clamp M, Durbin R. GeneWise and GenomeWise. Genome Research. 2004; 14:988–995. [PubMed: 15123596]
9. Burge, C. PhD thesis. Stanford University; 1997. Identification of Genes in Human Genomic DNA.
10. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. Improved microbial gene identification with GLIMMER. Nucleic acids research. 1999 Dec; 27(23):4636–4641. [PubMed: 10556321]
11. Gai X, Lal S, Xing L, Brendel V, Walbot V. Gene discovery using the maize genome database ZMDB. Nucleic Acids Research. 2000; 28(1):94–96. [PubMed: 10592191]
12. Gelfand MS, Mironov AA, Pevzner PA. Gene recognition via spliced sequence alignment. Proceedings of the National Academy of Sciences of the United States of America. 1996; 93(17): 9061–9066. [PubMed: 8799154]
13. Haykin, S. Neural Networks: A Comprehensive Foundation. Upper Saddle River, NJ, USA: Prentice Hall PTR; 1998.
14. Jiang C, Pugh BF. Nucleosome positioning and gene regulation: advances through genomics. Nat Rev Genet. 2009 Mar; 10(3):161–172. [PubMed: 19204718]
15. Johnson SM, Tan FJ, Mccullough HL, Riordan DP, Fire AZ. Flexibility and constraint in the nucleosome core landscape of *Caenorhabditis elegans* chromatin. Genome Research. 2006 Oct; 16(12):1505–1516. [PubMed: 17038564]
16. Kaplan T, Liu CL, Erkmann JA, Holik J, Grunstein M, Kaufman PD, Friedman N, Rando OJ, Steensel BV. Cell cycle– and chaperone-mediated regulation of H3K56ac incorporation in yeast. PLoS Genetics. 2008 Nov.4(11):e1000270. [PubMed: 19023413]
17. Kulp, D.; Haussler, D.; Reese, MG.; Eeckman, FH. Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology. AAAI Press; 1996. A generalized hidden Markov model for the recognition of human genes in DNA; p. 134-142.
18. Lee W, Tillo D, Bray N, Morse RH, Davis RW, Hughes TR, Nislow C. A high-resolution atlas of nucleosome occupancy in yeast. Nat Genet. 2007 Oct; 39(10):1235–1244. [PubMed: 17873876]
19. Liu CL, Kaplan T, Kim M, Buratowski S, Schreiber SL, Friedman N, Rando OJ. Single-nucleosome mapping of histone modifications in *S. cerevisiae*. Plos Biol. 2005 Jan.3(10):e328. [PubMed: 16122352]
20. Liu Q, Mackey AJ, Roos DS, Pereira FCN. Evigan: a hidden variable model for integrating gene evidence for eukaryotic gene prediction. Bioinformatics. 2008; 24(5):597–605. [PubMed: 18187439]
21. Liu Q, Mackey AJJ, Roos DSS, Pereira FCNC. Evigan: A hidden variable model for integrating gene evidence for eukaryotic gene prediction. Bioinformatics. 2008 Jan.
22. Majoros WH, Pertea M, Salzberg SL. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics. 2004 Nov; 20(16):2878–2879. [PubMed: 15145805]
23. Mavrich TN, Ioshikhes IP, Venters BJ, Jiang C, Tomsho LP, Qi J, Schuster SC, Albert I, Pugh BF. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. Genome Research. 2008 Jul; 18(7):1073–1083. [PubMed: 18550805]
24. Mavrich TN, Jiang C, Ioshikhes IP, Li X, Venters BJ, Zanton SJ, Tomsho LP, Qi J, Glaser RL, Schuster SC, Gilmour DS, Albert I, Pugh BF. Nucleosome organization in the drosophila genome. Nature. 2008 May; 453(7193):358–362. [PubMed: 18408708]
25. Parkinson J, Blaxter M. Expressed sequence tags: Analysis and annotation. Parasite Genomics Protocols, vol. 270 of Methods in Molecular Biology. 2004:93–126.
26. Ponts N, Harris EY, Prudhomme J, Wick I, Eckhardt-Ludka C, Hicks GR, Hardiman G, Lonardi S, Le Roch KG. Nucleosome landscape and control of transcription in the human malaria parasite. Genome Research. 2010; 20(2):228–238. [PubMed: 20054063]
27. Rabiner LR, Juang BH. An introduction to hidden Markov models. IEEE ASSP Magazine. 1986; 3(1):4–16.
28. Schones D, Cui K, Cuddapah S, Roh T, Barski A, Wang Z, Wei G, Zhao K. Dynamic regulation of nucleosome positioning in the human genome. Cell. 2008 Mar; 132(5):887–898. [PubMed: 18329373]
29. Shivaswamy S, Bhinge A, Zhao Y, Jones S, Hirst M, Iyer VR. Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. Plos Biol. 2008 Jan.6(3):e65. [PubMed: 18351804]

30. Shivaswamy S, Iyer VR. Stress-dependent dynamics of global chromatin remodeling in yeast: Dual role for SWI/SNF in the heat shock stress response. Molecular and Cellular Biology. 2008 Apr; 28(7):2221–2234. [PubMed: 18212068]

31. Tuggle CK, Green JA, Fitzsimmons C, Woods R, Prather RS, Malchenko S, Soares BM, Kucaba T, Crouch K, Smith C, Tack D, Robinson N, O'Leary B, Scheetz T, Casavant T, Pomp D, Edeal BJ, Zhang Y, Rothschild MF, Garwood K, Beavis W. EST-based gene discovery in pig: virtual expression patterns and comparative mapping to human. Mammalian Genome. 2003; 14(8):565–579. [PubMed: 12925889]

32. Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, Peckham H, Zeng K, Malek JA, Costa G, Mckernan K, Sidow A, Fire A, Johnson SM. A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. Genome Research. 2008 Jul; 18(7):1051–1063. [PubMed: 18477713]

33. Wan J, Lin J, Zack DJ, Qian J. Relating periodicity of nucleosome organization and gene regulation. Bioinformatics. 2009 May; 6(2):1–7.

34. Westenberger S, Cui L, Dharia N, Winzeler E, Cui L. Genome-wide nucleosome mapping of Plasmodium falciparum reveals histone-rich coding and histone-poor intergenic regions and chromatin remodeling of core and subtelomeric genes. BMC Genomics. 2009; 10(1):610. [PubMed: 20015349]

35. Yuan G-C, Liu Y-J, Dion MF, Slack MD, Wu LF, Altschuler SJ, Rando OJ. Genome-scale identification of nucleosome positions in *S. cerevisiae*. Science. 2005 Jul; 309(5734):626–630. [PubMed: 15961632]

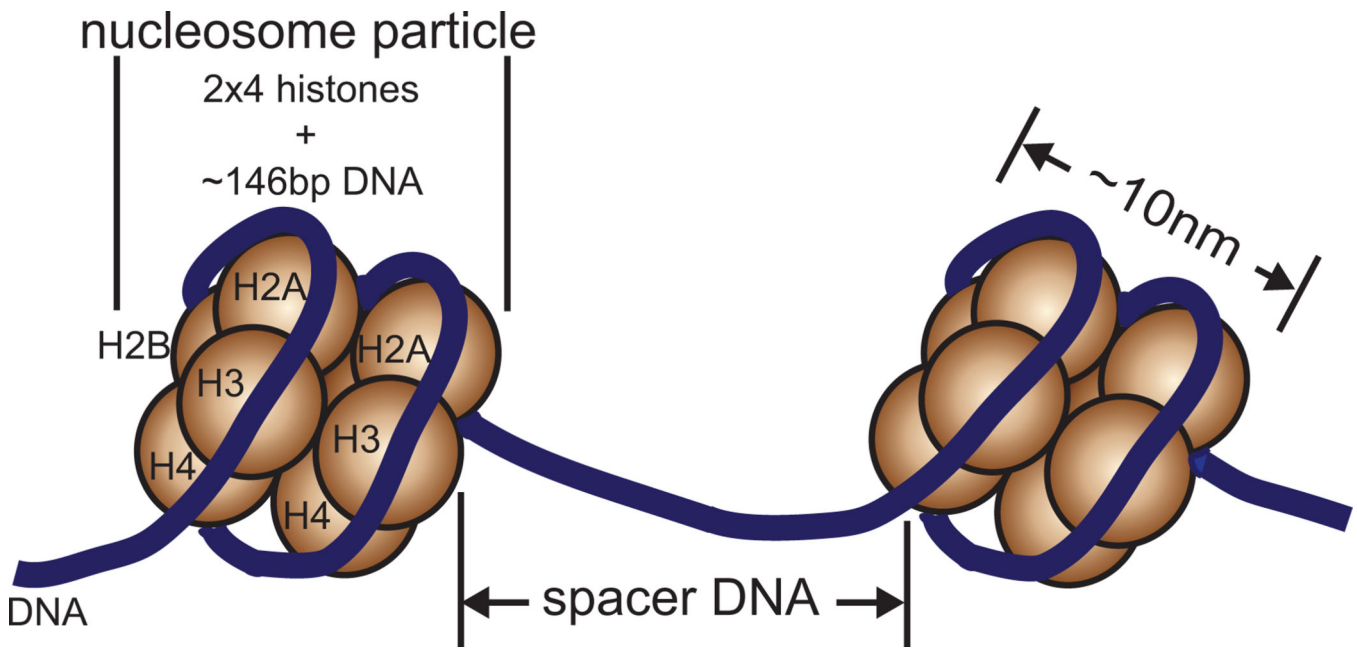

Fig. 1.

Eight histones and 146 base pairs of DNA form the nucleosome. The length of the spacer is organism specific.

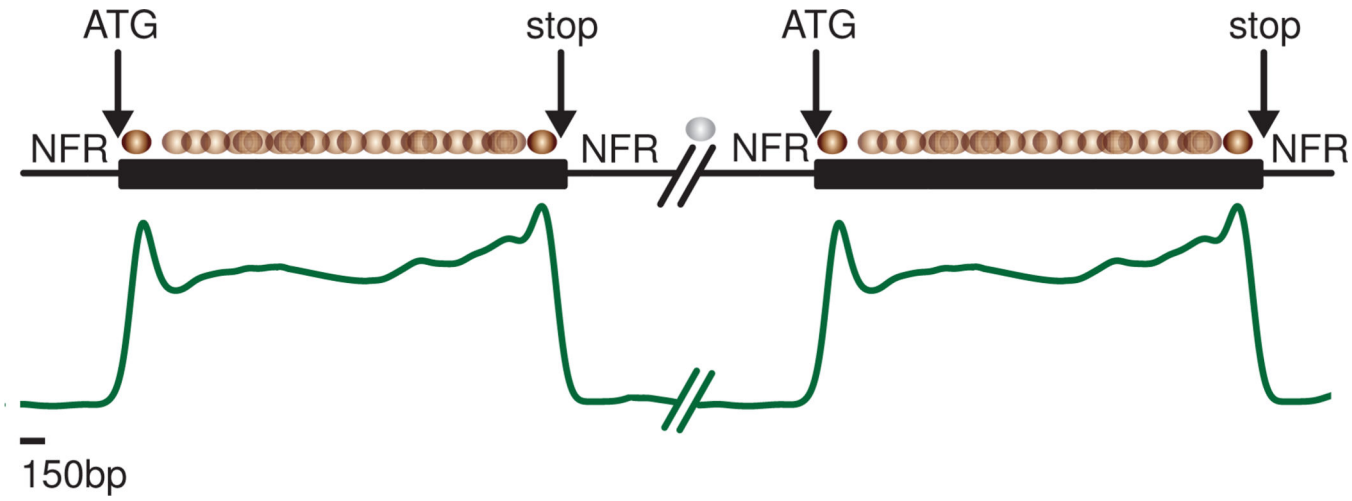

Fig. 2.

Nucleosome landscape of a typical *P. falciparum* gene. The ovals above the chromosome represent nucleosomes. The time series represents the likelihood of observing a nucleosome at that position.

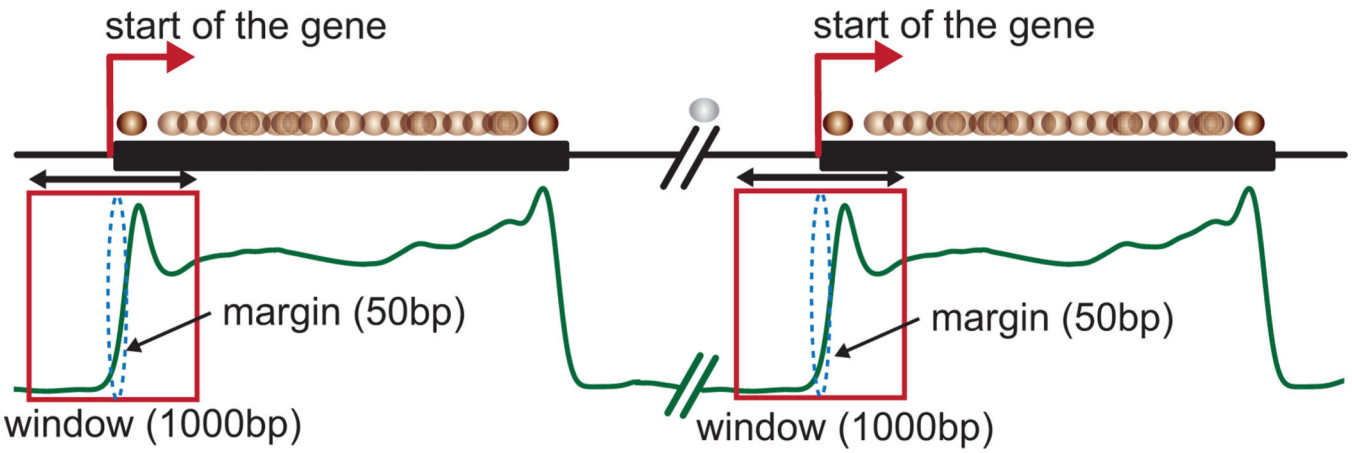

Fig. 3.
Illustrating a window, margin and the start of a gene in a nucleosomal landscape.

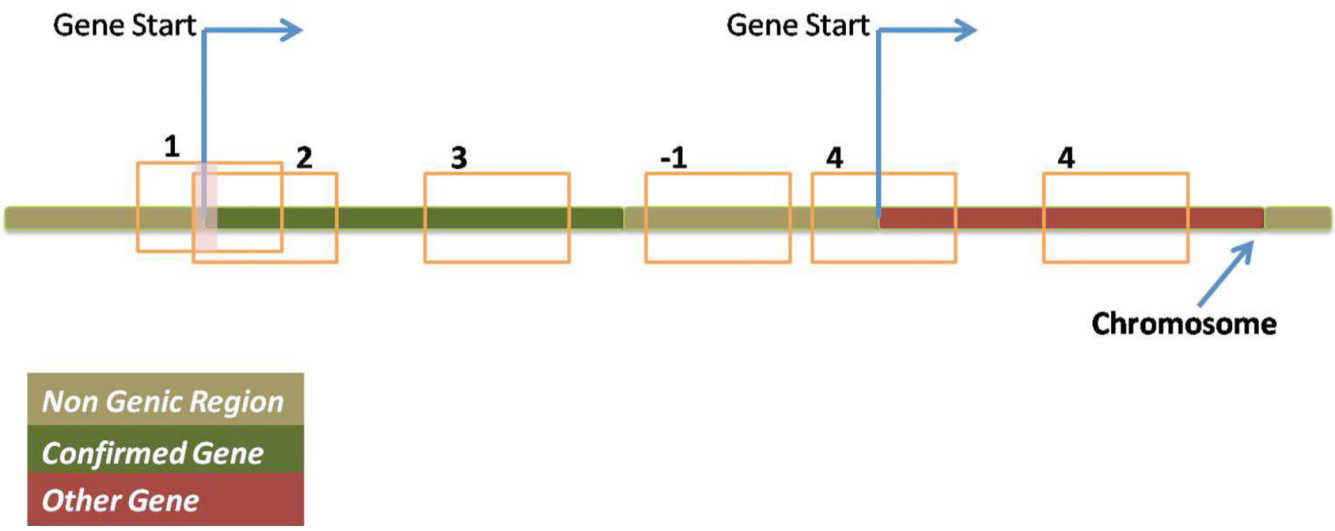

Fig. 4.
Labels associated with the windows extracted from a chromosome.

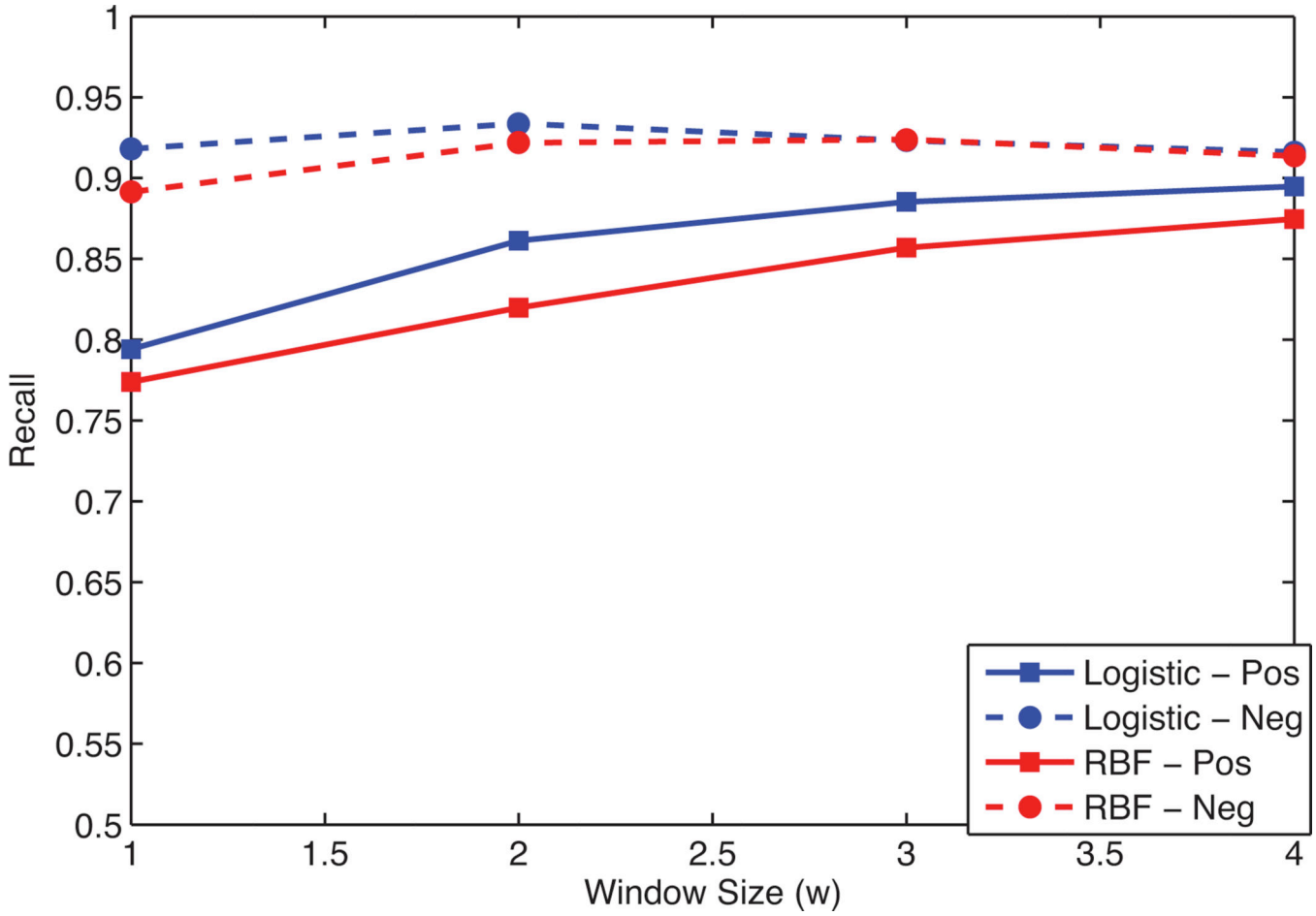

Fig. 5.

Recall for the positive and negative classes using logistic regression and RBF network classifier for varying window sizes (in thousand base pairs) and margin width $m = 50$.

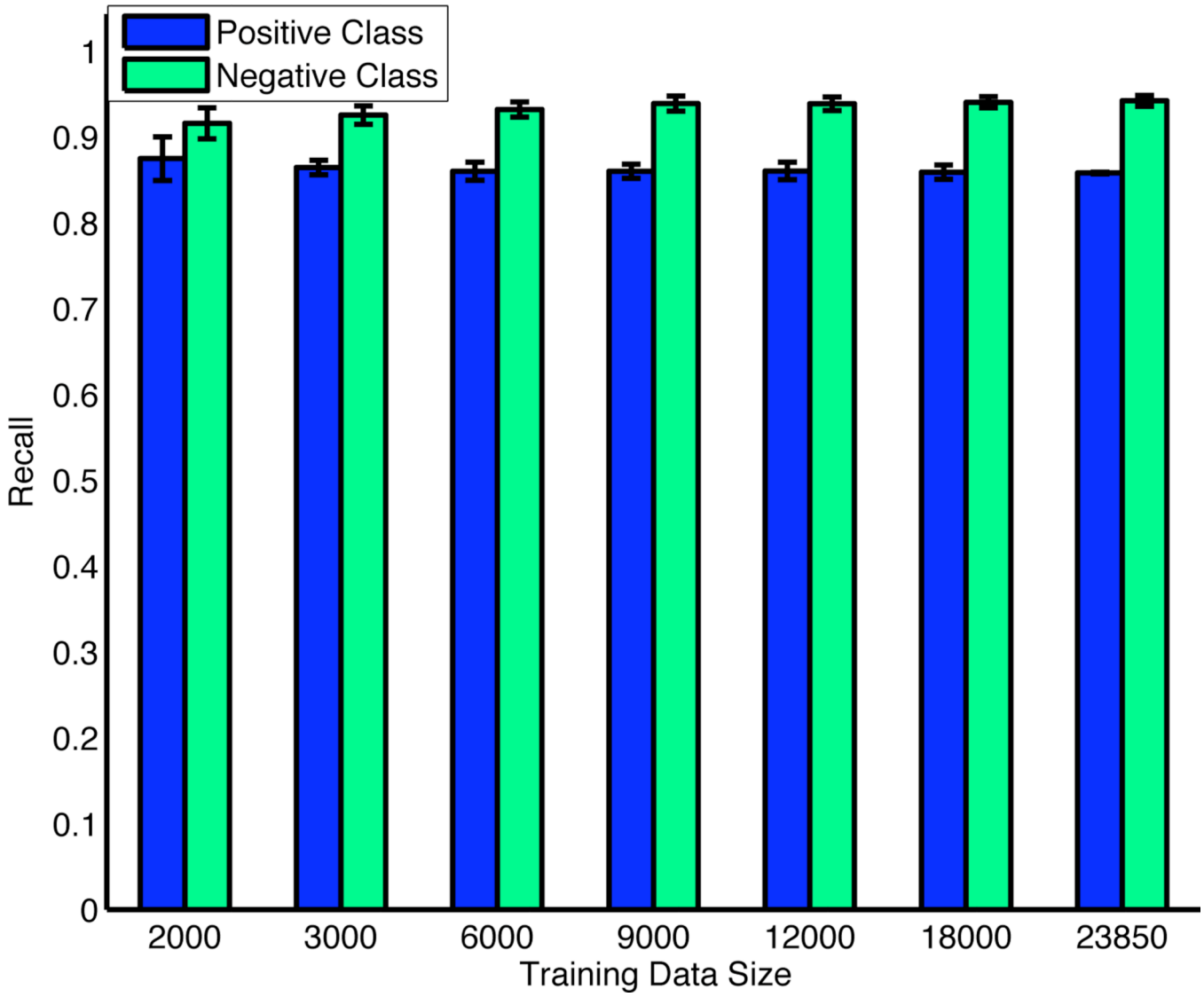

Fig. 6.

Recall for the positive and negative classes using logistic regression with training sample sizes for window size $w = 1000$ and margin width $m = 50$.

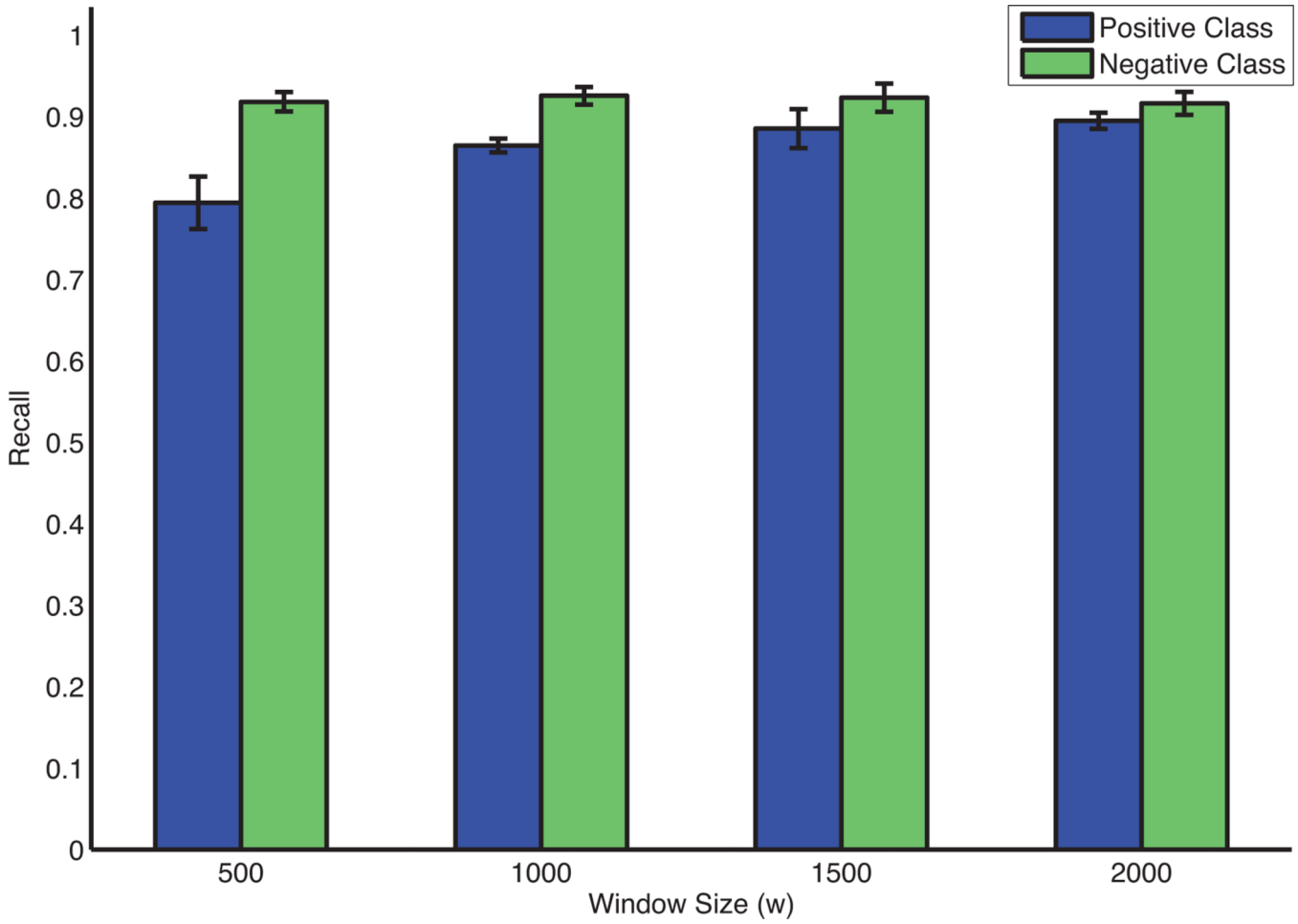

Fig. 7.

Recall for the positive and negative classes using logistic regression with varying window sizes and margin $m = 50$.

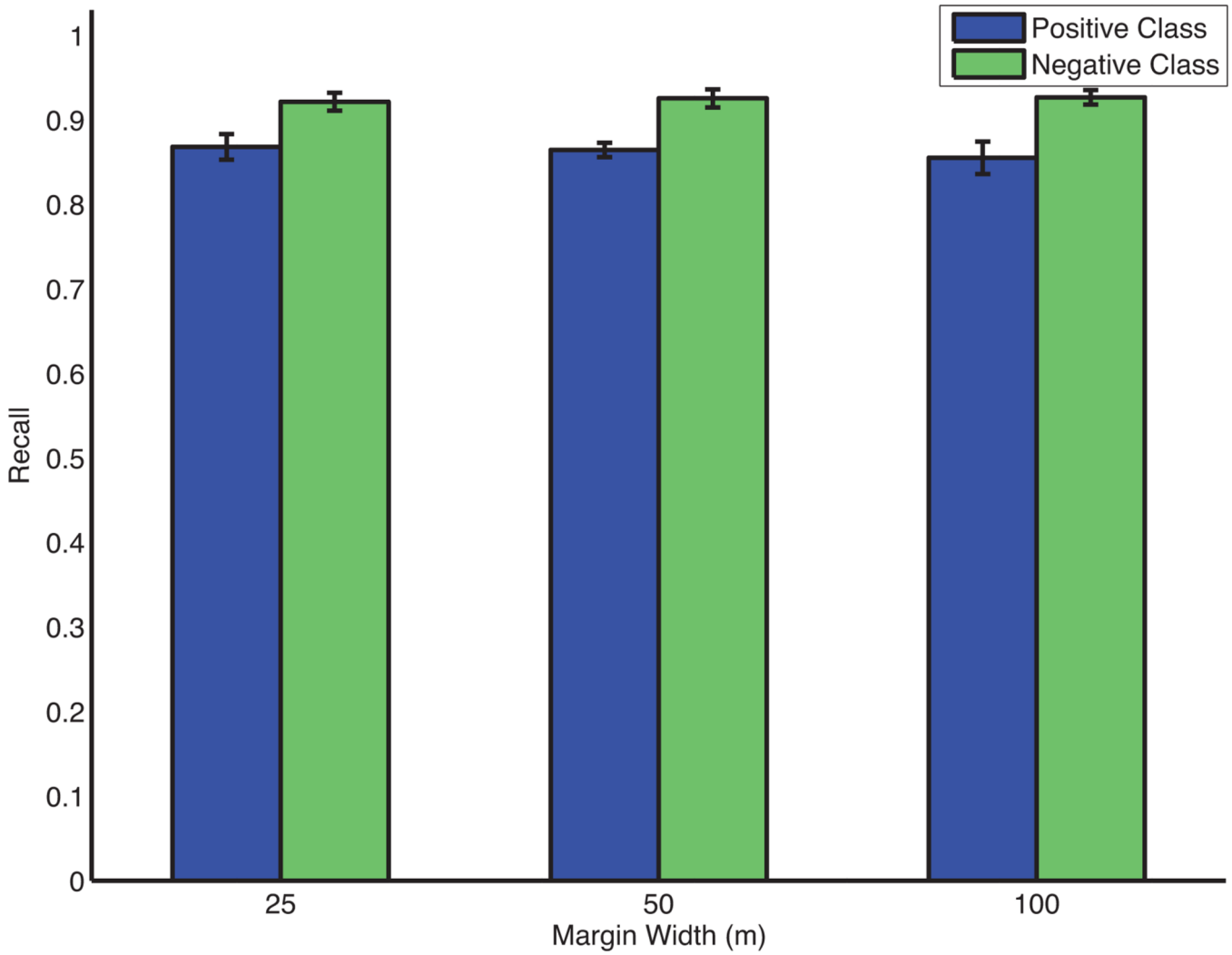

Fig. 8.

Recall for the positive and negative classes using logistic regression with varying margin widths and window size $w = 1000$.

Table 1

Confusion matrices for our classifier's predictions versus the actual labels associated with the windows on all fourteen chromosomes. Cells in boldface highlight the % of windows that contain potential new genes.

Chromosome 1

	Actual Labels				
Pred	**−1**	**1**	**2**	**3**	**4**
1	**0.09**	0.80	0.55	0.54	0.44
−1	0.91	0.20	0.45	0.46	0.56

Chromosome 2

	Actual Labels				
Pred	**−1**	**1**	**2**	**3**	**4**
1	**0.09**	0.74	0.69	0.53	0.55
−1	0.91	0.26	0.31	0.47	0.45

Chromosome 3

	Actual Labels				
Pred	**−1**	**1**	**2**	**3**	**4**
1	**0.05**	0.90	0.65	0.51	0.43
−1	0.95	0.10	0.35	0.49	0.57

Chromosome 4

	Actual Labels				
Pred	**−1**	**1**	**2**	**3**	**4**
1	**0.08**	0.80	0.69	0.64	0.53
−1	0.92	0.20	0.31	0.36	0.47

Comput Syst Bioinformatics Conf. Author manuscript; available in PMC 2014 July 28.

Chromosome 5

Pred	*Actual Labels*			
	−1	**1**	**2**	**3**	**4**
1	**0.06**	0.82	0.62	0.53	0.42
−1	0.94	0.18	0.38	0.47	0.58

Chromosome 6

Pred	*Actual Labels*			
	−1	**1**	**2**	**3**	**4**
1	**0.07**	0.89	0.75	0.55	0.54
−1	0.93	0.11	0.25	0.45	0.46

Chromosome 7

Pred	*Actual Labels*			
	−1	**1**	**2**	**3**	**4**
1	**0.08**	0.89	0.63	0.53	0.42
−1	0.92	0.11	0.37	0.47	0.58

Chromosome 8

Pred	*Actual Labels*			
	−1	**1**	**2**	**3**	**4**
1	**0.12**	0.96	0.79	0.59	0.53
−1	0.88	0.04	0.21	0.41	0.47

Chromosome 9

Pred	*Actual Labels*			
	−1	**1**	**2**	**3**	**4**
1	**0.04**	0.89	0.63	0.52	0.40
−1	0.96	0.11	0.37	0.48	0.60

Chromosome 10

Pred	*Actual Labels*			
	−1	**1**	**2**	**3**	**4**
1	**0.08**	0.91	0.67	0.61	0.55
−1	0.92	0.09	0.33	0.39	0.45

Chromosome 11

Pred	*Actual Labels*			
	−1	**1**	**2**	**3**	**4**
1	**0.06**	0.87	0.66	0.48	0.41
−1	0.94	0.13	0.34	0.52	0.59

Chromosome 12

Pred	*Actual Labels*			
	−1	**1**	**2**	**3**	**4**
1	**0.06**	0.91	0.74	0.59	0.53
−1	0.94	0.09	0.26	0.41	0.47

Chromosome 13

Pred	*Actual Labels*			
	−1	**1**	**2**	**3**	**4**
1	**0.05**	0.85	0.61	0.48	0.38
−1	0.95	0.15	0.39	0.52	0.62

Chromosome 14

Pred	*Actual Labels*			
	−1	**1**	**2**	**3**	**4**
1	**0.07**	0.90	0.71	0.61	0.48
−1	0.93	0.10	0.29	0.39	0.52

Table 2

A chromosome-by-chromosome summary of the number of candidate segments for the start of the genes, the proportion of those that matched an EST, and the source of the matching ESTs.

Chr. #	# Segments	# Segments with matches in EST	Matching Species		
			P. falciparum	Other *Plasmodium*	*Others*
1	4	3	1	1	1
2	28	13	11	2	0
3	4	3	1	1	1
4	17	7	5	1	1
5	9	5	2	1	2
6	12	7	5	2	0
7	11	7	5	0	2
8	19	15	15	0	0
9	4	4	4	0	0
10	28	19	18	1	0
11	16	10	8	2	0
12	24	20	17	3	0
13	17	12	8	0	4
14	30	21	18	1	2
Total:	**223**	**146**	**118**	**15**	**13**

Table 3

A chromosome-by-chromosome summary of the number of candidate segments for the end of the genes, the proportion of those that matched an EST, and the source of the matching ESTs.

Chr. #	# Segments	# Segments with matches in EST	Matching Species		
			P. falciparum	Other *Plasmodium*	*Others*
1	5	3	2	1	0
2	9	4	3	1	0
3	2	2	0	1	1
4	9	5	3	1	1
5	9	6	3	1	2
6	8	5	3	2	0
7	19	16	13	1	2
8	7	7	3	1	3
9	5	4	4	0	0
10	17	11	10	1	0
11	21	16	12	3	1
12	9	6	4	2	0
13	19	13	10	0	3
14	22	15	13	1	1
Total:	**161**	**113**	**83**	**16**	**14**