BIOINFORMATICS

Vol. 21 Suppl. 1 2005, pages i338–i343
doi:10.1093/bioinformatics/bti1047

Predicting the in vivo *signature of human gene regulatory sequences*

William Stafford Noble[1,]*, *Scott Kuehn*[2], *Robert Thurman*[2], *Man Yu*[2] *and John Stamatoyannopoulos*[3]

[1]*Department of Genome Sciences and Department of Computer Science and Engineering,* [2]*Division of Medical Genetics, University of Washington, Seattle, WA, USA and* [3]*Department of Molecular Biology, Regulome, 2211 Elliott Avenue, Suite 600, Seattle, WA 98121, USA*

Received on January 15, 2005; accepted on March 27, 2005

ABSTRACT

Motivation: In the living cell nucleus, genomic DNA is packaged into chromatin. DNA sequences that regulate transcription and other chromosomal processes are associated with local disruptions, or 'openings', in chromatin structure caused by the cooperative action of regulatory proteins. Such perturbations are extremely specific for *cis*-regulatory elements and occur over short stretches of DNA (typically ~250 bp). They can be detected experimentally as DNaseI hypersensitive sites (HSs) *in vivo*, though the process is extremely laborious and costly. The ability to discriminate DNaseI HSs computationally would have a major impact on the annotation and utilization of the human genome.

Results: We found that a supervised pattern recognition algorithm, trained using a set of 280 DNaseI HS and 737 non-HS control sequences from erythroid cells, was capable of *de novo* prediction of HSs across the human genome with surprisingly high accuracy determined by prospective *in vivo* validation. Systematic application of this computational approach will greatly facilitate the discovery and analysis of functional non-coding elements in the human and other complex genomes.

Availability: Supplementary data is available at noble.gs.washington.edu/proj/hs

Contact: noble@gs.washington.edu; jstam@regulome.com

1 INTRODUCTION

The vast majority of gene regulatory sequences in the human and other complex genomes remain undiscovered. In the living cell nucleus, DNA is packaged into chromatin fibers by non-specific association with the histone proteins that make up the nucleosome. Binding of activating proteins to regulatory DNA sequences requires cooperativity between the regulatory factors in order to displace a nucleosome, which in turn disrupts the local architecture of chromatin. This fundamental feature of eukaryotic *cis*-regulatory sequences was recognized nearly 25 years ago (Wu, 1980; Gross and Garrard, 1988), when it was discovered that such sequences were hypersensitive to cutting by the non-specific endonuclease DNaseI *in vivo*.

DNaseI hypersensitive sites (HSs) have since proven to be extremely reliable and generic markers of *cis*-regulatory sequences. Mapping of DNaseI HSs is a gold-standard approach for discovering functional non-coding elements involved in gene regulation and has underpinned the discovery of most experimentally established distal *cis*-acting elements in the human genome. In most cases, identification of functional elements marked by HSs significantly preceded the assignment of a specific functional role (enhancer, insulator, etc.) to those elements (Gross and Garrard, 1988; Li *et al*., 2002).

Comprehensive identification of DNaseI HSs in the human genome would be expected to disclose the location of all known classes of *cis*-regulatory sequences, including promoters, enhancers, silencers, insulators, boundary elements and locus control regions. Computational methods for the identification of the DNaseI HSs would therefore be expected to accelerate dramatically the functional annotation of the human genome.

Traditional approaches to computational prediction of *cis*-regulatory sequences in complex genomes have focused on identification and combinatorial analysis of short sequence motifs (presumed to represent regulatory factor binding sites) derived from examples of known sites, analysis of upstream regions of co-regulated genes (Sinha and Tompa, 2002; Berman *et al*., 2002), analysis of phylogenetic data or combinations thereof (Prakash *et al*., 2004). Unfortunately, the performance of even the most advanced algorithms is poor (Tompa *et al*., 2005), and the described methods generally lack biological validation, particularly in the context of the human genome. Even in the case of extensively characterized loci, such as the α- and β-globin domains, computational motif-based approaches have proven to be of little value for the discovery or annotation of HSs.

*To whom correspondence should be addressed.

The core sequences giving rise to HSs *in vivo* are anticipated to contain complex features that facilitate recognition by specific sets of regulatory factors interacting cooperatively over relatively short distances (150–250 bp) (Felsenfeld, 1996; Stamatoyannopoulos *et al*., 1995). However, it is not clear a priori whether recognition of such features is computationally tractable.

Conventional molecular approaches to the visualization of HSs have relied on an indirect method (Wu, 1980), and subsequent experimental localization of the core 150–250 bp activating sequences is extremely laborious (Lowrey *et al*., 1992; Talbot *et al*., 1990). As a result, relatively few HSs identified with traditional methods have been localized definitively to specific sequence elements, precluding systematic computational analyses. Recently, however, novel methods for large-scale sequence-specific discovery of DNaseI HSs have been described (Sabo *et al*., 2004; Dorschner *et al*., 2004), providing the basis for the recovery of larger numbers of DNaseI HSs sequences that can be utilized in computational models.

In this paper, we demonstrate that a sequence-based classification algorithm can learn to recognize DNaseI HSs with high accuracy. To train the algorithm, we take advantage of a collection of 280 validated erythroid HS sequences from throughout the human genome. We also use a set of 737 confirmed non-HS sequences of equivalent length. We employ a support vector machine (SVM) classifier, which learns by example to discriminate between two given classes of data (Vapnik, 1998; Cristianini and Shawe-Taylor, 2000). In a cross-validated test, the SVM achieves an accuracy of 85.24 ± 5.03% in predicting HSs. Furthermore, we perform a prospective *in vivo* experimental validation of the SVM predictions on previously untested regions of the human genome, using the assay described by Sabo *et al*. (2004) and Dorschner *et al*. (2004). Among HS predictions to which the SVM assigns probabilities >80%, 79.4% prove to be HSs on experimental validation in two hematopoietic cell types.

2 METHODS

2.1 Data

For training and cross-validation of the SVM, we use 280 validated erythroid HS sequences from throughout the human genome. These enabling sequences emerged from the recent description of a novel methodology for the identification of HSs via cloning based on their *in vivo* activity in K562 erythroid cells (Sabo *et al*., 2004). We also collected 737 sequences from around the genome (distributed proportionally among the autosomes and X chromosome but excepting the Y chromosome) that were non-hypersensitive when tested in the same cell type. Both K562 HS and non-HS sequences were similar in size (mean length 242.1 versus 242.8 bp, respectively). The complete dataset is available at noble.gs.washington.edu/proj/hs

We designed primers using Primer3 (Rozen and Skaletsky, 2000) with the following parameters: target amplimer size = 250 bp ± 50 bases; primer T_m (melting temperature) optimal = 60 ± 2°C; %GC = 50% optimal, range 40–80%; length = 24 bp optimal, range 19–27 bp; poly X maximum = 4.

We cultured erythroid cells (K562, ATCC) under standard conditions [37°C, 5% CO_2 in air, RPMI 1640 plus 10% FBS (Invitrogen, Carlsbad, CA, USA)]. We harvested the cultures at a density of 5×10^5 cells/ml. We performed DNaseI digestions following a standard protocol (Reitman *et al*., 1993). DNA was subsequently purified using the Puregene system (Gentra Systems, Minneapolis, MN, USA).

2.2 Support vector machine

We use the freely available Gist SVM implementation (Pavlidis *et al*., 2004). For each SVM optimization, we use the default parameters: a linear kernel function and a 2-norm soft margin with asymmetric penalties assigned to the positive and negative classes. Experiments with higher-order kernel functions and different soft margin settings yielded only very small changes in performance (data not shown).

The output of the SVM is a unit-free discriminant score; however, this score can be converted into a more useful probability by performing a sigmoid curve fit (Platt, 1999). This approach involves holding out a portion of the training set from the SVM optimization and fitting the sigmoid parameters using the discriminants from the held-out data. A probability score of 50% corresponds approximately to the hyperplane identified by the SVM, and increasing or decreasing probabilities are reflective (non-linearly) of increasing distance from the hyperplane (in positive or negative directions). The Gist software implements this curve fitting procedure.

2.3 Performance measure

We measure the overall quality of an SVM classifier using a receiver operating characteristic (ROC) curve (Hanley and McNeil, 1982). The trained SVM receives as input a list of candidate HS sequences and produces as output a ranked list of these sequences, with the confidently predicted HSs at the top of the list. Setting a threshold anywhere in this ranked list produces a particular rate of true and false positives with respect to that threshold. The ROC curve plots true positive rate as a function of false positive rate as the threshold varies from the top to the bottom of the ranked list. The ROC score is the area under this curve. A classifier that correctly places all of the HSs at the top of its ranked list would receive a ROC score of 1, whereas a random ranking would receive a score of ∼0.5.

3 RESULTS

The SVM algorithm learns to separate a set of labeled training data by placing the data in a high-dimensional space (a *feature* *space*) and discovering in that space a hyperplane that separates the two classes. Predicting the label of a new, unlabeled data point simply involves determining on which side of the hyperplane that point lies. SVMs boast powerful theoretical underpinnings (Vapnik, 1998) and wide applicability because of their use of kernel functions to represent data. The kernel function defines similarities between pairs of data points and allows the SVM to operate in an implicit vector space even for non-vector data, such as teer, graphs and strings. In computational biology, SVMs have been applied to a wide variety of problems (Noble, 2004), including the classification of several types of DNA sequence elements: translation start sites (Zien *et al*., 2000) and splice sites (Zhang *et al*., 2003).

Before the SVM classification of HS and non-HS sequences, we need to embed the sequences into a vector space. In this work, this embedding is accomplished by using the spectrum kernel (Leslie *et al*., 2002). We hypothesize that the difference between HS and non-HS sequences can be well characterized in terms of the presence of various short, motif-like sequence features. The spectrum kernel exhaustively enumerates all such features ('k-mers') of a given length (k) and represents each sequence as the frequency with which each k-mer appears in the sequence. For example, the sequence 'ACGT' contains three distinct 2mers ('AC,' 'CG' and 'GT'). The $k = 2$ spectrum kernel representation of this sequence is a 16-element vector (one entry for each possible dinucleotide), with 0.33 for the three k-mers listed above and 0 for all other entries. In general, we do not expect the k-mers to be strand-specific, so reverse complements are collapsed into a single feature. Thus, for $k = 2$, there are only 10 distinct dinucleotides. In the experiments reported here, we concatenate the feature vectors for $k = 1, \ldots, 6$. Thus, the feature vector representation of a sequence contains $2 + 10 + \ldots = 2772$ entries.

3.1 Cross-validation

We first tested the pattern recognition performance of the SVM via 10-fold cross-validation on the collection of 1017 (280 + 737) sequences. This test involves randomly dividing the sequence set into 10 equal-sized subsets, and then repeatedly training on 90% subsets of the data and testing the SVM's generalization performance on the held-out 10%. For this data set, the mean area under the ROC curve across 10-fold cross-validation was 0.842±0.021, indicative of excellent performance (Fig. 1). At the classification threshold selected by the SVM, the mean accuracy was 85.24 ± 5.03%.

We hypothesize that the DNaseI sequences that the SVM fails to identify during cross-validation represent a distinct, hard-to-identify subclass of DNaseI HSs. To test this hypothesis, we collected a set of 83 HSs that were incorrectly classified as non-HS during cross-validation. Removing the 83-member false-negative (FN) class from the training set and then retraining and cross-validating a new SVM (using the same 737 non-HS sequences) produced an ROC of 0.970 ± 0.0045. Conversely, a second SVM trained to discriminate between the 83 FN sequences and the remaining 934 sequences achieved an ROC of 0.635 ± 0.026. This result signifies a weaker classifier, though one which performs substantially better than chance ($p < 0.0000017$). Thus, learning accurately to recognize this smaller and potentially more diverse class of HSs may require a larger training set or a different collection of sequence features.

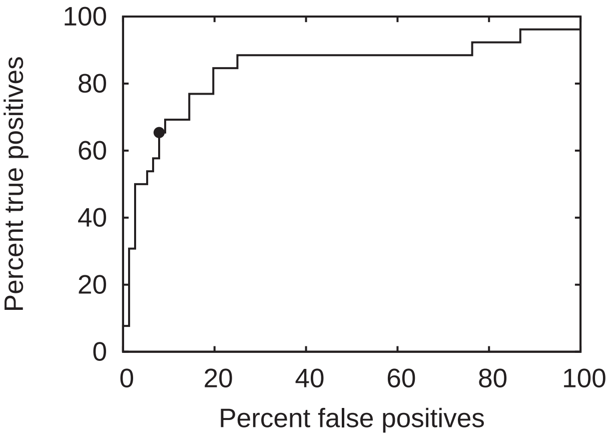

Fig. 1. Receiver operating characteristic curve for SVM discrimination of DNaseI HS versus non-HS sequences. The ROC was computed by training an SVM on a randomly selected 90% subset of a dataset comprising 280 HS and 737 non-HS sequences, followed by testing on the held-out 10%. The area under this particular curve is 0.84059, indicative of excellent performance. The dot marks the location of the decision boundary selected by the SVM. At this threshold, the SVM correctly identifies 17 HSs and 70 non-HSs, and makes 6 false positive and 9 false negative predictions.

3.2 Prospective experimental validation

Next we tested the ability of an SVM trained over a random 90% subset of the combined 1017 K562 HS and Non-HS examples to predict the *in vivo* DNaseI HS status in K562 cells of 60 000 non-repetitive sequences (as identified by the RepeatMasker track on the UCSC Genome Browser) with mean length 225 bp selected from throughout the human genome. The expected prevalence of HSs in this set of sequences is higher than random background but <10% (Sabo *et al*., 2004).

From the resulting SVM probabilities, we randomly selected for further testing sequences with assigned high probability (>80%; $n = 146$) and low probability (<20%; $n = 43$). Each sequence was tested for DNaseI hypersensitivity in K562 erythroid cells using a previously validated real-time quantitative PCR assay designed to discriminate DNaseI HSs with >95% confidence (Sabo *et al*., 2004). We found 108/146 of the high probability predictions to be DNaseI HSs when tested in K562 cells, yielding a positive-predictive value (PPV) for the SVM of 73.9% (Fig. 2). Testing of low probability predictions in the same cell type revealed that 39/43 were correctly

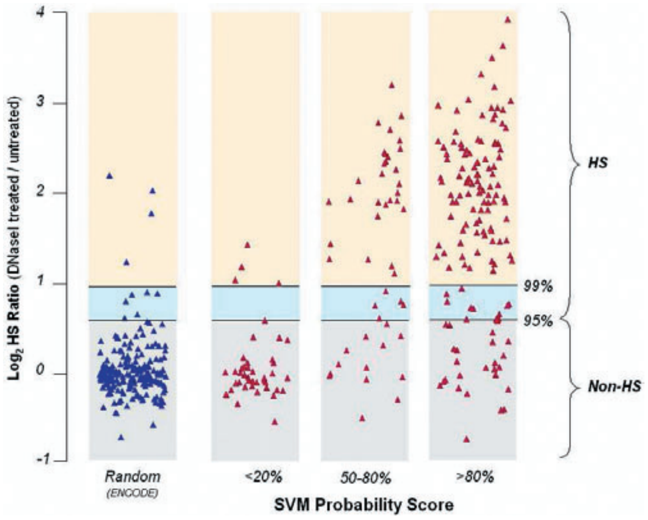

Fig. 2. DNaseI hypersensitivity testing of SVM predictions and control sequences in K562 cells. The y-axis plots the $\log_2$ of the DNaseI sensitivity ratio (copies remaining in DNaseI-untreated sample/DNaseI-treated sample) assayed by real-time quantitative PCR. Results from SVM predictions are stratified into low ($<20\%$), intermediate (50–80%), and high ($>80\%$) SVM-assigned probability groups. Means of six replicate measurements for each PCR amplified sequence ('amplicon') corresponding to an SVM prediction are shown with triangles. Results are classified as non-HS (gray shaded boxes) or HS [blue (95% confidence) and orange (99% confidence) shaded boxes] on the basis of quantitative DNaseI hypersensitivity measurements obtained with real-time PCR (Sabo *et al*., 2004; McArthur *et al*., 2001; Dorschner *et al*., 2004) using a validated model for K562 cells described by Sabo *et al*. (2004). Results from 186 randomly selected amplicons from the ENCODE regions (ENCODE Consortium, 2004) are also shown. The proportion of HS-positives in the random set (5.3%) is higher than expected for the genome at large, given the considerably higher gene and functional element density of the ENCODE regions. (Notably, HSs from the random set coincided with known or predicted regulatory sequences, including HS4 from the β-globin LCR and several promoters and CpG islands.)

classified as non-HSs, for a negative-predictive value (NPV) of 90.7%. We also examined 49 intermediate probability (50–80%) predictions, and found 33 (67.3%) to be positive. The cumulative PPV for all predictions with probability $>50\%$ was 70.6%. These results demonstrate the ability of the SVM to identify DNaseI HSs *in vivo* with high accuracy.

The high proportion of true-positive predictions within a single cell type suggests further that the elements identified by the SVM might represent a class of HSs that are active in many tissues or are even constitutive. Additionally, because some HSs are expected to be tissue or lineage-restricted, a proportion of predictions that yielded negative results in erythroid cells might prove to be HS in another tissue type. To address this possibility, we tested a subset ($n = 93$) of sequences with assigned probability $>50\%$ in another hematopoietic cell type, B-lymphoblastoid cells (EBV-transformed primary lymphoblast line GM0990, Coriell). Of 65 SVM-predicted HSs that were DNaseI hypersensitive in K562 cells, 58 (89.2%) were also HSs in lymphoblastoid cells. An exemplary SVM-predicted HS of this type lying upstream of the NF1 tumor suppressor gene is illustrated in Figure 3. Conversely, we found 8/28 (28.6%) sequences that tested negative in K562 cells were HS-positive in lymphoid cells. These results indicate that the overall PPV estimate for the SVM based on testing only in K562 cells represents a minimum value. More extensive testing in additional tissue types might reveal further SVM HS predictions to be correct.

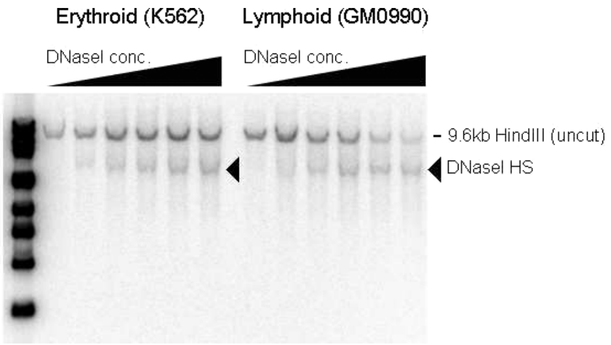

Fig. 3. Conventional DNaseI HS analysis of SVM predictions. To confirm further that SVM predictions correspond to classical DNaseI HSs, we selected positive predictions for conventional DNaseI HS assays employing the indirect end-label Southern blotting technique (Lowrey *et al*., 1992). Shown are exemplary results from an SVM prediction 400 bp upstream of the Nf1 tumor suppressor gene on chromosome 17 that coincides with a classical DNaseI HS in both erythroid (K562) and lymphoid (GM0990) cells. For each tissue type, lanes represent increasing (left to right) DNaseI treatment intensity (0, 1, 2, 4, 8 and 16 U DNaseI). A radiolabeled probe is targeted to the 5′ end of a 9.6 kb HindIII fragment encompassing the Nf1 transcriptional start site and upstream and downstream flanking sequences. As DNaseI concentration increases, the 9.6 kb parental band is cleaved specifically at the hypersensitive site, releasing the marked sub-band.

3.3 Genome-wide prediction

We then considered how frequently SVM-predicted sequences occur in the human genome. We first partitioned the human genome sequence (assembly hg16 = NCBI 34) into non-overlapping 225 bp segments and identified 4 217 066 segments lacking repetitive sequences. Next, we selected and scored all segments and applied a sigmoid fit to derive probabilities from the SVM discriminant scores. The SVM predicted 36 581 (0.89%) genomic segments to be HSs at a probability threshold of 50%; 19 429 (0.47%) had probability scores $>80\%$. At a cumulative minimum PPV level of 70.6% for DNaseI HSs *in vivo*, these results suggest that the human genome contains $>26\,500$ functional non-coding elements of the class predicted by the SVM. Analysis of the distribution of SVM predictions in relation to genes revealed strong clustering around annotated transcriptional start sites; however, 65% of predictions were located >5 kb distant from the nearest 5′ start site.

3.4 Feature analysis

In order to perform its classification, the SVM simultaneously exploits a large collection of simple k-mer sequence features. This collection does not correspond to traditional motifs, but encompasses them in the context of a rich feature space, which implicitly allows for mismatching and complex dependencies between sequence positions by combining many short k-mer features.

In order to gain some insight into this complex feature space representation, we analyzed the set of 83 improperly classified (FN) HSs from the initial training set for the presence of simple sequence features that distinguished them from the correctly recognized class. We observed that the CG dinucleotide frequency was significantly lower (1.3%) in the FN class than in the 197 correctly discerned HSs (6.8%), and that the AT dinucleotide frequency was also skewed, but to a lesser degree (6.2% versus 2.8%, respectively).

To examine whether the SVM had exploited these disparities in producing its initial classifications, we computed the Pearson correlation between the SVM discriminants and each of the 2772 sequence features. This analysis revealed that, during the initial training, the SVM had highlighted CG dinucleotides as the most important simple sequence feature, with a correlation of 0.916. Previous observations stemming from specific genes have suggested that certain CpG-rich sequences play a role in maintaining open chromatin structures (Tazi and Bird, 1990); however, the generality of this observation was unknown. A posteriori analysis of the 36 581 human genomic predictions revealed a sharply lower correlation (0.679), indicating that the SVM was integrating a complex array of additional features in performing predictions. Given the overlap between CpG islands and functionally important genomic locales, significant overlap between the SVM predictions and this feature is expected. However, 34% of the 36 581 predictions lie outside CpG islands, as defined by the CpG island track on the UCSC Genome Browser. Moreover, where overlap occurs, only a small fraction (13%) of the CpG sequence is highlighted by the SVM, suggesting that it is recognizing the functional core of these nebulously defined elements.

3.5 Enrichment in CTCF sites

Although most classes of regulatory sequences bind to a variety of regulatory proteins, insulator and chromatin domain boundary elements invariably contain recognition sites for the protein CTCF. Insulator and boundary elements organize the human genome by partitioning functional gene domains (Bell *et al*., 2001). These elements typically give rise to prominent DNaseI HSs that are manifest across a wide range of tissue types. We therefore hypothesized that CTCF sites should be significantly enriched in high versus low probability SVM predictions. We searched sets of sequences selected from the top 25% and bottom 25% of the SVM probability range for occurrences of the canonical CTCF binding motif CCGCNNGGNGGCAG. This search discovered 3462 CTCF sites that received positive log-odds scores in the top 25% set and only 335 such sites in the bottom 25% set. Using a more stringent log-odds threshold of 2, we found 548 CTCF sites in the top 25% and 29 sites in the bottom 25% set. Among the top 25% set, 3 CTCF sites perfectly match the consensus and 57 more match with a single mismatch. No sites match this well in the bottom 25%. The dramatic enrichment of CTCF sites in high probability SVM predictions suggests that a prominent subset of SVM-predicted HSs function *in vivo* as insulator or domain boundary elements.

4 DISCUSSION

Identification of DNaseI HSs is a gold-standard methodology for the identification of vertebrate *cis*-regulatory sequences and has facilitated the discovery of the vast majority of validated human *cis*-regulatory elements residing outside of core promoters. Although novel molecular approaches for large-scale mapping of DNaseI HSs have recently been described (Dorschner *et al*., 2004; Sabo *et al*., 2004), comprehensive annotation of human DNaseI HSs—even in the context of a single tissue—remains distant and will require substantial resources. In contrast, computational tools provide the basis for rapid coverage of the entire genome.

A priori, prediction of DNaseI HSs is expected to be an extremely challenging computational problem. The fact that it has proven tractable for a subclass of these elements is therefore quite surprising. Given the relatively modest size of the training sets employed here, the accuracy of the approach will probably improve with expanded numbers of examples. Although not every HS necessarily encodes a classical *cis*-regulatory element, most HSs do. It is therefore notable that the current level of predictive accuracy (PPV 70%) is substantially higher than that described for any computationally based methodology for identification of *cis*-regulatory sequences. Nor is attainment of 100% accuracy a requirement, given the potential for coupling of computational predictions to a platform for high-throughput biological validation, such as the high-throughput real-time PCR assay employed here for prospective examination of SVM annotations. Iterative application of the training-and-testing paradigm with additional HS sequences should enable generation of more powerful, accurate and diverse classifiers.

Although described and validated in the context of a single tissue (human erythroid cells), the approach described here is broadly applicable. Extension of this paradigm to other tissue types should enable recognition of additional classes of HSs and, thereby, delineation of large numbers of novel elements expected to play central roles in the transcriptional regulation of human genes. Because DNaseI HSs are a fundamental property of *cis*-regulatory sequences from a wide variety of organisms, the approach described here should be widely extensible to other vertebrate genomes, and to higher eukaryotic genomes generally.

In summary, our results demonstrate the feasibility of accurate, large-scale computational prediction of the *in vivo* signature of human *cis*-regulatory sequences and provide a powerful new tool for the annotation of complex genomes.

REFERENCES

Bell,A.C., West,A.G. and Felsenfeld,G. (2001) Insulators and boundaries: versatile regulatory elements in the eukaryotic genome. *Science*, **291**, 447–450.

Berman,B.P., Nibu,Y., Pfeifer,B.D., Tomancak,P., Celniker,S.E., Levine,M., Rubin,G.M. and Eisen,M.B. (2002) Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc. Natl Acad. Sci. USA*, **99**, 757–762.

Cristianini,N. and Shawe-Taylor,J. (2000) *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge, UK.

Dorschner,M.O. *et al*. (2004) High-throughput localization of functional elements by quantitative chromatin profiling. *Nat. Meth.*, **1**, 219–225.

ENCODE Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) project. *Science*, **306**, 636–640.

Felsenfeld,G. (1996) Chromatin unfolds. *Cell*, **86**, 13–19.

Gross,D.S. and Garrard,W.T. (1988) Nuclease hypersensitive sites in chromatin. *Annu. Rev. Biochem.*, **57**, 159–197.

Hanley,J.A. and McNeil,B.J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**, 29–36.

Leslie,C., Eskin,E. and Noble,W.S. (2002) The spectrum kernel: a string kernel for SVM protein classification. In Altman,R.B., Dunker,A.K., Hunter,L., Lauderdale,K. and Klein,T.E. (eds), *Proceedings of the Pacific Symposium on Biocomputing*, World Scientific, New Jersey, pp. 564–575.

Li,Q., Peterson,K.R., Fang,X. and Stamatoyannopoulos,G. (2002) Locus control regions. *Blood*, **100**, 3077–3086.

Lowrey,C.H., Bodine,D.M. and Nienhuis,A.W. (1992) Mechanism of DNase I hypersensitive site formation within the human globin locus control region. *Proc. Natl Acad. Sci. USA*, **89**, 1143–1147.

McArthur,M., Gerum,S. and Stamatoyannopoulos,G. (2001) Quantification of DNaseI-sensitivity by real-time PCR: quantitative analysis of DNaseI-hypersensitivity of the mouse β-globin LCR. *J. Molec. Biol.*, **313**, 27–34.

Noble,W.S. (2004) Support vector machine applications in computational biology. In Schölkopf,B., Tsuda,K. and Vert, J.-P. (eds) *Kernel Methods in Computational Biology*, MIT Press, Cambridge, MA, pp. 71–92.

Pavlidis,P., Wapinski,I. and Noble,W.S. (2004) Support vector machine classification on the web. *Bioinformatics*, **20**, 586–587.

Platt,J.C. (1999) Probabilities for support vector machines. In Smola,A., Bartlett,P., Schölkopf,B. and Schuurmans,D. (eds), *Advances in Large Margin Classifiers*, MIT Press, Cambridge, MA, pp. 61–74.

Prakash,A., Blanchette,M., Sinha,S. and Tompa,M. (2004) Motif discovery in heterogeneous sequence data. In *Proceedings of the Pacific Symposium on Biocomputing*, Vol. 9, pp. 348–359.

Reitman,M., Lee,E., Westphal,H. and Felsenfeld,G. (1993) An enhancer/locus control region is not sufficient to open chromatin. *Molec. Cell. Biol.*, **13**, 3990–3998.

Rozen,S. and Skaletsky,H. (2000) Primer3 on the WWW for general users and for biologist programmers. In Krawetz,S. and Misener,S. (eds), *Bioinformatics Methods and Protocols*, Methods in Molecular Biology, Human Press, Totowa, NJ: pp. 365–386.

Sabo,P.J., Hawrylycz,M., Wallace,J.C., Humbert,R., Yu,M., Shafer,A., Kawamoto,J., Hall,R., Mack,J., Dorschner,M.O., McArthur,M. and Stamatoyannopoulos,J.A. (2004) Discovery of functional noncoding elements by digital analysis of chromatin structure. *Proc. Natl Acad. Sci. USA*, **101**, 16837–16842.

Sabo,P.J., Humbert,R., Hawrylycz,M., Wallace,J.C., Dorschner,M.O., McArthur,M. and Stamatoyannopoulos,J.A. (2004) Genome-wide identification of DNase1 hypersensitive sites using active chromatin sequence libraries. *Proc. Natl Acad. Sci. USA*, **101**, 4537–4542.

Sinha,S. and Tompa,M. (2002) Discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res.*, **30**, 5549–5560.

Stamatoyannopoulos,J.A., Goodwin,A., Joyce,T. and Lowrey,C.H. (1995) NF-E2 and GATA binding motifs are required for the formation of DNase I hypersensitive site 4 of the human beta-globin locus control region. *EMBO J.*, **14**, 106–116.

Talbot,D., Philipsen,S., Fraser,P. and Grosveld,F. (1990) Detailed analysis of the site 3 region of the human beta-globin dominant control region. *EMBO J.*, **9**, 2169–2177.

Tazi,J. and Bird,A. (1990) Alternative chromatin structure at CpG islands. *Cell*, **60**, 909–920.

Tompa,M. *et al.* (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, **23**, 137–144.

Vapnik,V.N. (1998) Adaptive and learning systems for signal processing, communications, and control. In *Statistical Learning Theory*. Wiley, New York.

Wu,C. (1980) The 5′ ends of Drosophila heat shock genes in chromatin are hypersensitive to DNase I. *Nature*, **286**, 854–860.

Zhang,X.H.-F., Heller,K.A., Hefter,I., Leslie,C.S. and Chasin,L.A. (2003) Sequence information for the splicing of human pre-mRNA identified by support vector machine classification. *Genome Res.*, **13**, 2637–2650.

Zien,A., Rätch,G., Mika,S., Schölkopf,B., Lengauer,T. and Müller,K.-R. (2000) Engineering support vector machine kernels that recognize translation initiation sites. *Bioinformatics*, **16**, 799–807.