GENOME RESEARCH

ESPERR: Learning strong and weak signals in genomic sequence alignments to identify functional elements

James Taylor, Svitlana Tyekucheva, David C. King, Ross C. Hardison, Webb Miller and Francesca Chiaromonte

Genome Res. 2006 16: 1596-1604; originally published online Oct 19, 2006;
Access the most recent version at doi:10.1101/gr.4537706

ESPERR: Learning strong and weak signals in genomic sequence alignments to identify functional elements

James Taylor,[1] Svitlana Tyekucheva, David C. King, Ross C. Hardison, Webb Miller, and Francesca Chiaromonte[1]

Center for Comparative Genomics and Bioinformatics, The Pennsylvania State University, University Park, Pennsylvania 16802, USA

Genomic sequence signals—such as base composition, presence of particular motifs, or evolutionary constraint—have been used effectively to identify functional elements. However, approaches based only on specific signals known to correlate with function can be quite limiting. When training data are available, application of computational learning algorithms to multispecies alignments has the potential to capture broader and more informative sequence and evolutionary patterns that better characterize a class of elements. However, effective exploitation of patterns in multispecies alignments is impeded by the vast number of possible alignment columns and by a limited understanding of which particular strings of columns may characterize a given class. We have developed a computational method, called ESPERR (evolutionary and sequence pattern extraction through reduced representations), which uses training examples to learn encodings of multispecies alignments into reduced forms tailored for the prediction of chosen classes of functional elements. ESPERR produces a greatly improved Regulatory Potential score, which can discriminate regulatory regions from neutral sites with excellent accuracy (~94%). This score captures strong signals (GC content and conservation), as well as subtler signals (with small contributions from many different alignment patterns) that characterize the regulatory elements in our training set. ESPERR is also effective for predicting other classes of functional elements, as we show for DNaseI hypersensitive sites and highly conserved regions with developmental enhancer activity. Our software, training data, and genome-wide predictions are available from our Web site (http://www.bx.psu.edu/projects/esperr).

[Supplemental material is available online at www.genome.org.]

Identification of functional elements within genome sequences often relies on specific characteristic signals, typically based on known biological examples. For instance, prediction of protein-coding exons and genes relies on knowledge of the genetic code and splicing signals. These predictions can be improved by incorporating evolutionary information from orthologous regions of other species through sequence alignments. In particular, insertions and deletions are rarely tolerated in coding regions, whereas substitutions at synonymous sites are frequently tolerated, and algorithms that effectively model these signals generate improved predictions (Korf et al. 2001; Siepel and Haussler 2004a). Knowledge of the rules for *cis*-regulatory modules is less complete, and hence prediction of these in individual or aligned sequences remains elusive. For functions where even less is known, such as replication origins or movement to appropriate locations within the nucleus, prediction is all the more challenging.

Signals currently used for identifying *cis*-regulatory modules include (1) specific sequence patterns, such as motifs associated with elements involved in protein–DNA interactions (e.g., transcription factor binding sites), (2) general sequence composition patterns, such as the high density of CpG dinucleotides found in most ubiquitous promoters, and (3) evolutionary patterns, particularly a high level of interspecies conservation, which should characterize functional regions under purifying selection.

While each of these signals is associated with some *cis*-regulatory modules, all of them have limitations (Tompa et al. 2005). Motif-based approaches can have high specificity, particularly when using a stringent consensus sequence, but when the patterns are degenerate (often the case with transcription factors), they can have both poor sensitivity and a very high false-positive rate. When the sites occupied by transcription factors in mammalian cells are identified in a relatively unbiased manner, such as by chromatin immunoprecipitation assayed on arrays of all nonrepetitive DNA (ChIP–chip), only a minority of the sites have a clear match to the binding site motif (Cawley et al. 2004; Bieda et al. 2006). These results suggest that, for a comprehensive set of binding sites, motif-based approaches have weak power.

Similarly, a complex relationship exists between function and evolutionary constraint. Many classes of functional elements do show significant association with constrained elements as a whole (Waterston et al. 2002). However a large number of functional genomic elements do not overlap constrained regions, and many constrained regions do not coincide with known functional elements (Bejerano et al. 2004; Siepel et al. 2005). These results suggest that interspecies sequence constraint also provides only weak power for comprehensive identification of functional elements.

Thus, it seems that while noncoding functional elements

[1] **Corresponding authors.**
E-mail james@bx.psu.edu; fax (814) 863-6699.
E-mail chiaro@stat.psu.edu; fax (814) 863-6699.

Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.4537706. Freely available online through the *Genome Research* Open Access option.

show association with various sequence and evolutionary characteristics, rarely will a single signal be sufficient for accurate and comprehensive prediction. While simple descriptive features can be very useful to better understand functional mechanisms, the effects of functional constraint on these elements are myriad—too complicated to be captured effectively by such features alone.

An alternative approach for identification of a class of functional elements for which training data are available is to apply a computational learning method with the potential to capture both the clear strong signals and the many subtle signals that characterize the class. Two major obstacles must be overcome to develop an effective method. First, the number of possible alignment columns increases exponentially with the number of sequences in an alignment. This number (>70,000 for a seven-species alignment) is much too large an "alphabet" to use to find patterns in alignment columns, and thus a reduced representation of the alignment is required. Second, the rules for distinguishing between functional classes based on patterns in alignments are not known a priori, and thus a training regimen is required. To solve these problems, we have designed a method trained on genomic sequence alignments, which contain information about the primary sequence of a set of species and the evolutionary relationships among them. Our method, denoted ESPERR (evolutionary and sequence pattern extraction through reduced representation), uses models capable of learning patterns both among the species at a given position (evolutionary patterns) and among aligned positions (and thus across the sequence). Underlying these models is a translation or "encoding" of alignments into a simplified representation that preserves a subset of the original information. This reduced representation should remove noise and irrelevant information but retain all the signals useful for characterizing a particular class of functional elements.

The key component of our method is the selection of such an encoding using (1) phylogenetic relationships to define a reasonable starting point, followed by (2) a heuristic search procedure that optimizes the encoding based on classification performance. Encodings produced by this procedure, and the models based on them, produce excellent classification performance on a variety of problems.

After explaining ESPERR in somewhat more detail, we turn to the prediction of *cis*-regulatory modules, using ESPERR to compute an improved Regulatory Potential (RP) score. This score, trained to discriminate regulatory regions from neutral sites, achieves a very good success rate of ~94% (the success rate is the fraction of training elements correctly classified using leave-one-out cross-validation) and shows excellent performance on a largely independent set of regulatory elements from the hemoglobin β gene cluster. To better understand the signals that contribute to this excellent performance, we explore the structure of the encoded word frequencies on which the score is based, and how this structure relates to RP. We find clear, strong signals (in particular, GC content and conservation) associated with a small number of encoded words, as well as diffuse, weak signals associated with combinations of many such words. Both kinds of signals contribute significantly to RP scores. We observe that the weak-signal component of RP may help to identify regulatory elements that lie far from any transcription start site.

Next, we apply ESPERR to detection of DNaseI hypersensitive sites. Training on DNaseI hypersensitivity data produced as part of the ENCODE project, we are able to effectively discriminate these regions, with a cross-validation success rate of ~80%. We show that this result compares favorably to an earlier approach based on support vector machines.

Finally, we consider the problem of identifying highly conserved regions that show developmental enhancer activity. The VISTA Enhancer Browser (http://enhancer.lbl.gov) describes 253 conserved regions that were tested for consistent enhancer activity in transgenic mouse embryos, 108 of which (~42%) show such activity. Using these tested regions for training, we are able to predict the assay result with ~83% accuracy. This indicates that ESPERR could greatly improve the effectiveness of the strategy employed to select regions to be tested for developmental enhancer activity.

Results

ESPERR

The ESPERR procedure finds an encoding from multiple alignment columns into a reduced alphabet that retains information useful for discriminating a chosen class of functional elements. The procedure consists of two stages, summarized graphically in Figure 1. In the first stage, we reduce the "alphabet" of alignment columns to a size where fitting classification models becomes tractable by grouping multiple alignment columns based on evolutionary similarity (Fig. 1A). We start by inferring the ancestral base probability distribution corresponding to each alignment column—for this we use an extended Hasegawa, Kishino, and Yano (HKY) substitution model, treating alignment gaps as a fifth base (see Methods). This inference provides a natural way to handle missing data; if data for a species are missing at a given position, it is not included in the inference (Fig. 1A, middle tree). In practice, the number of missing species allowed must be limited to ensure good inference. Next, we compute the frequency with which each ancestral distribution occurs in the training data and apply a novel clustering algorithm that forms groups of columns preserving both "neighborhood" (similarity of ancestral distributions) and frequency structure. Ancestral distributions correspond to points in a five-dimensional probability simplex, a two-dimensional projection of which is used to visualize the clustering step in Figure 1B.

The clusters resulting from the initial alphabet reduction in stage 1 provide an encoding that retains a substantial amount of information from the original alignment data, and reduced representations produced in this way can be used effectively for many applications. However, performance can be improved substantially by taking such an encoding as a starting point and then using classification performance to optimize the encoding for a particular problem. The second stage of the ESPERR procedure achieves this through an iterative search (Fig. 1C). At each stage of this search, candidate encodings are generated from the current encoding by either joining two groups or breaking a group into two (a random sample from each type of candidate is considered). Using the training data, cross-validation is run to evaluate the prediction performance of each candidate, and the candidate with the best performance is accepted as the new current encoding. After many iterations without seeing an improvement in performance the search is terminated, yielding an optimized encoding, usually to many fewer symbols (groups) than the starting point.

While this approach could be applied using any classification method, we generally use a log-odds classifier based on a type of variable-order Markov model (VOMM) (Buhlman and

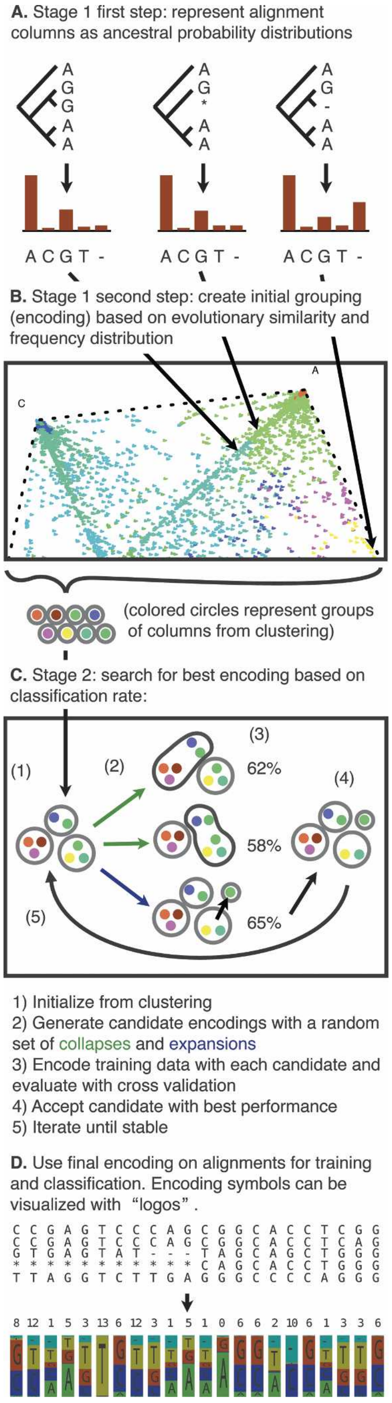
A. Stage 1 first step: represent alignment columns as ancestral probability distributions
A G G A A
A G * A A
A G - A A
A C G T -
A C G T -
A C G T -
B. Stage 1 second step: create initial grouping (encoding) based on evolutionary similarity and frequency distribution
A
C
(colored circles represent groups of columns from clustering)
C. Stage 2: search for best encoding based on classification rate:
(1)
(2)
(3)
(4)
(5)
62%
58%
65%
1) Initialize from clustering
2) Generate candidate encodings with a random set of collapses and expansions
3) Encode training data with each candidate and evaluate with cross validation
4) Accept candidate with best performance
5) Iterate until stable
D. Use final encoding on alignments for training and classification. Encoding symbols can be visualized with "logos".
C C G A G T C C C A G C G G C A C C T C G G
C C G A G T C C C A G C G G C A C C T C A G
G T G A G T A T - - - T A G C A G C T G G G
* * * * * * * * * * * C A G C A C C T G G G
T T A G G T C T T G A G G G C C C C A G G G
8 12 1 5 3 13 6 12 3 1 5 1 0 6 6 2 10 6 1 3 3 6

Figure 1. Overview of the ESPERR procedure.

Wyner 1998). These models capture variable-length dependencies among positions in sequences. Thus, when applied to strings of encoded alignment columns, VOMMs are able to capture sequence and evolutionary patterns that span multiple alignment columns.

Full details on ancestral distribution inference, clustering, the iterative search, and the fitting of VOMMs are provided in the Methods and Supplemental material.

Learning RP with ESPERR

Despite years of intense study, *cis*-regulatory elements remain difficult to predict. It has been shown previously (Elnitski et al. 2003; Kolbe et al. 2004) that an approach based on patterns in encoded alignments can be effective for discriminating these regions from ancestral repeats (a model for likely neutral regions). Applying ESPERR to learn an encoding for this problem yields a substantial improvement in discrimination over previous methods. The positive training data consists of a set of 97 experimentally validated regulatory elements (Elnitski et al. 2003). These were compiled from a diverse group of genes, including those expressed in muscle, liver, lung, and erythroid cells, and thus they contain binding sites for a wide variety of transcription factors. The negative training data are ancestral repeats, which are repetitive elements already present in the common ancestor of human, mouse, and dog. Alignments of seven species (human, chimpanzee, macaque, mouse, rat, cow, and dog) corresponding to the training regions were extracted from the UCSC Genome Browser (Karolchik et al. 2003). To improve the resolution of our cross-validation procedure, these alignments were chopped into 100-column pieces, and the ancestral repeats were randomly sampled to produce a training set equal in size to the positive set. We allowed alignment columns to be considered if they had no more than three missing species (and none missing among the three highest quality sequences: human, mouse, and dog), and required each 100-column segment to have at least 50 such columns. This resulted in positive and negative training sets containing 357 elements, covering ~31,000 human bases each. ESPERR with a log-odds classifier based on VOMMs (with a maximal order of 2) yielded a final encoding into 17 symbols, with a leave-one-out cross-validation success rate of ~94% on the training data.

This performance is a considerable improvement over previous RP scores (~82% for the scores of Kolbe et al. 2004, based on human–rodent alignments). Cumulative distributions of RP scores computed on the training sets and similarly prepared random samples of exonic and bulk genomic regions are shown in Figure 2A. RP scores do an excellent job discriminating regulatory regions from bulk and neutral DNA, as well as separating them from exons.

As an additional evaluation of RP performance, we considered 23 experimentally confirmed regulatory elements in the hemoglobin β gene cluster. These likely include most of the sequences with regulatory function for this extensively studied locus, and only five are part of our regulatory training set—providing reasonably exhaustive and independent test data for sensitivity and specificity assessments (King et al. 2005; see Supplemental data for technical details). The ROC plots (Fig. 2B) show that performance of the ESPERR-based RP scores on this data set, in terms of both sensitivity and specificity, is uniformly better than previous RP scores (from human–rodent alignments), GC content (measured for 100-bp windows), and two conservation scores: phastCons (Siepel et al. 2005) and MCS (Margulies et al. 2003).

ESPERR captures a variety of signals in regulatory elements

To begin unraveling the signals that contribute to the excellent performance of the RP score, we must examine the variability structure in the training data and how this structure relates to the score. We want to know which features of the training data lead to good performance, but this is a challenging prospect given

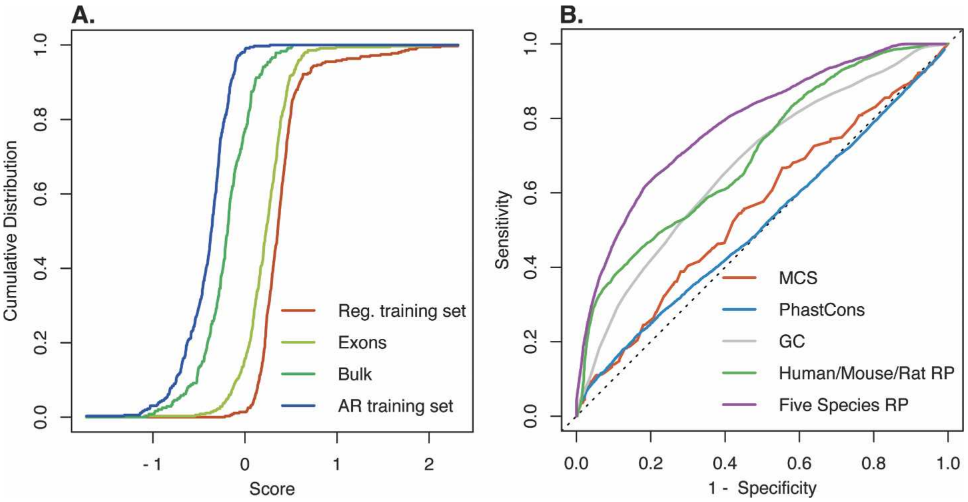

Figure 2. RP score performance demonstrated by cumulative distributions of scores on various genomic elements (*A*) and ROC plots for discrimination of 23 elements in the human β-globin locus (*B*).

that a very large number of alignment columns are grouped together by ESPERR for each reduced representation. Because RP is a log-odds score based on VOMMs with maximal order 2, we consider the frequencies of words of length 3 in the training data after applying the encoding learned by ESPERR. One approach for understanding the variability structure of a data set is principal component analysis (PCA), which finds a transformation of a data set to a new coordinate system in which the first component has the greatest variance, the second (orthogonal to the first) has the next greatest variance, and so on. Applying PCA to these word frequencies shows that a large amount of their variability is explained by the first few principal components (Fig. 3, top). However, a substantial amount of variability is spread across the many remaining components, consistent with the presence of both strong and weak signals in this data set.

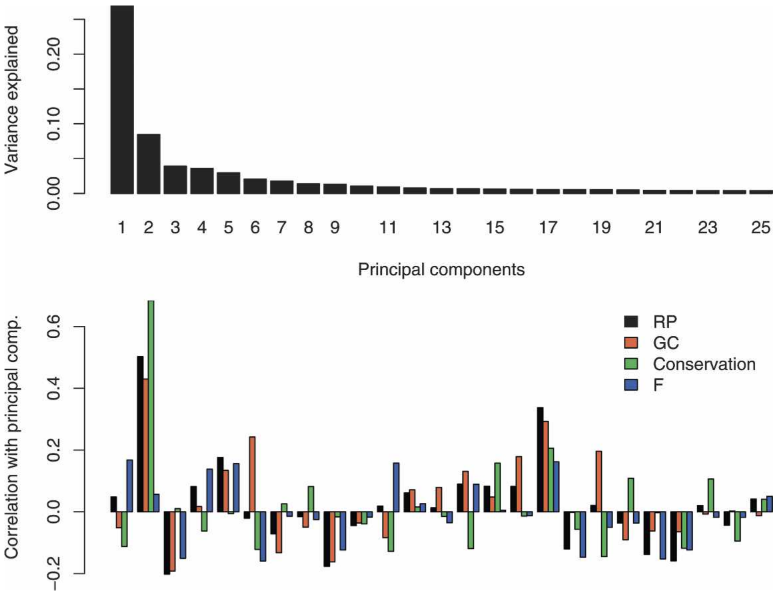

Figure 3. Share of variance explained by each of the first 25 principal components of the RP training data word frequencies (*top*) and correlation of RP score, GC content, conservation, and the residuals *F* with each principal component (*bottom*).

Our first insight into the nature of the strong signals comes from our analysis of the performance of RP scores. We note that while RP can discriminate regulatory elements better than conservation scores and GC content, exons can also have very high RP values. In fact, conservation and GC content are two signals traditionally associated with exons, as well as regulatory elements. Computing a regression of RP score on GC content and conservation (measured as the average phastCons score) shows that these two quantities alone explain ~68% of the variability in RP. Another factor typically associated with ubiquitous promoter regions is CpG dinucleotide density (Cooper et al. 2006); however, while CpG density does explain some within-class variability of the regulatory elements in our training set, it does not not independently contribute to RP (the percentage of RP variability explained does not increase when adding CpG density to the regression). Pinpointing the nature of other factors that systematically contribute to RP is complicated, because of the enormous reduction induced by our encoding and the random component involved in the search algorithm. Nevertheless, these factors are crucial for discrimination; about a third of RP is likely a composite of weaker signals. A practical way to measure this composite is to consider the residuals from the regression of RP on GC content and conservation, which we will denote as *F*. The bottom panel of Figure 3 shows the correlation of RP and each of these three quantities with the first 25 principal components. We see that the strongest component that has high correlation with RP also has high correlation with conservation and GC content; however, RP also shows substantial correlation with many of the weaker components, which are less exclusively dominated by the strong conservation and GC content signals.

To explore further the difference between these strong signals and the composite of weak, subtler signals represented by *F*, we correlate each of these three quantities with individual word frequencies in the training data. Figure 4 (bottom) shows box plots of these correlations. The positive correlation with both conservation and GC content are dominated by a small number of words, which are the outliers at the top of the distribution. In contrast, *F* shows far fewer dominant outliers and is associated with many different words. Further insight into the nature of these signals is obtained by examining the specific words that have the strongest positive correlation with each feature. Figure 4 (top) shows "logos" for the words most strongly correlated with each signal (the height of each character in the logo is determined by the ancestral probability distribution centroid for the columns encoded to that symbol). Again, conservation and GC content are dominated by words clearly associated with these signals (the search procedure has grouped fully conserved C and G columns together, so the symbol with strong G and C components shows up frequently in the highly conserved set). The words most strongly associated with *F* on the other hand are more diverse, consistent with indications that a variety of different patterns contribute to *F*.

RP weak components help to identify truly distal regulatory elements

Another way to gauge the role of strong and weak signals in RP scores is to examine these signals on an independent, complex set of regulatory elements. In particular, we would like to understand the signals RP captures in distal regulatory elements, which are less well characterized by conservation and GC content. To investigate this, we defined a collection of putative distal regulatory elements using various data sources available from the ENCODE Consortium.

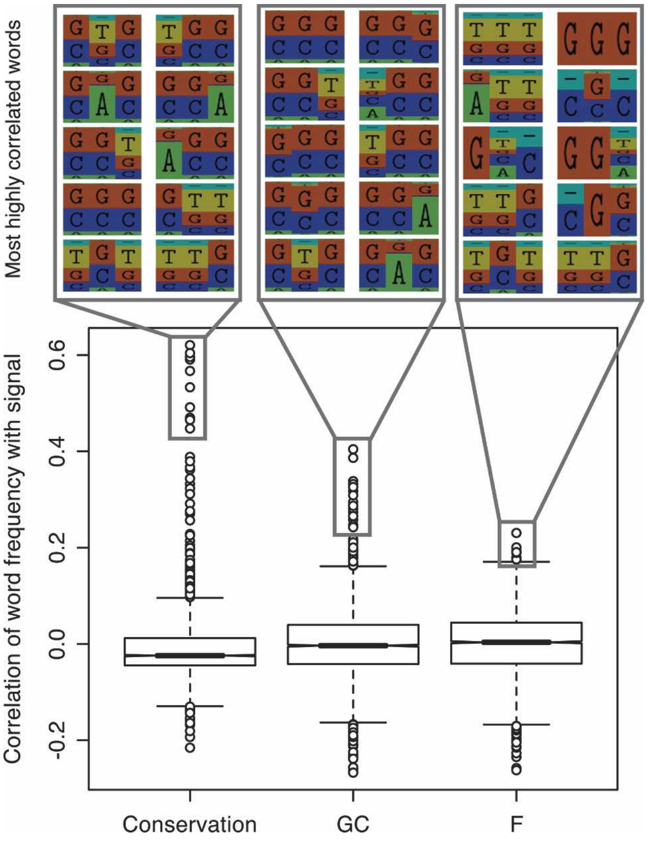

Figure 4. Distributions of the correlations between word frequencies in the RP training data and three component signals (GC content, conservation, and the residuals *F*). For each signal, representative logos of the most strongly correlated words are shown. (green, A; yellow, T; red, G; blue, C).

The ENCODE Transcriptional Regulation group used ChIP–chip to identify binding sites for a variety of factors (http://genome.ucsc.edu/encode/; Bieda et al. 2006; B. Ren, M. Snyder, and T. Gingeras, pers. comm.). We selected a subset of their experiments, emphasizing experimental platforms with high-resolution site identification and sequence-specific binding not exclusively associated with transcription start sites, and eliminated all sites overlapping repetitive regions or coding exons, expanding the remaining sites to cover at least 100 bp. To improve the quality of this set further, we restricted attention to sites supported by at least one additional line of experimental evidence suggesting regulation, such as ChIP–chip evidence for certain histone modifications associated with activation or factors associated with general chromatin modification, as well as DNaseI hypersensitivity and nucleosome depletion (http://genome.ucsc.edu/encode/; J. Stamatoyannopoulos and J. Lieb, pers. comm.). Finally, to focus on distal regulation, we removed sites falling within 2.5 kb of a transcription start site (http://genome.ucsc.edu/encode/; T. Gingeras, pers. comm.). This resulted in a set of 617 elements with multiple lines of evidence suggesting a distal regulatory function, 583 of which had sufficient aligning sequence to calculate the RP score, GC content, and average phastCons score.

Aggregate characteristics of these regions suggest that they are enriched for function; in particular, they show evidence of evolutionary constraint as measured by average phastCons scores (Siepel et al. 2005). This set may also contain some nonfunctional elements, as well as unannotated promoters and proximal elements, because there are likely transcription start sites that have not been identified.

For each of the three RP components (GC, conservation, and *F*), we examined the 50 highest scoring elements. Among those with high GC content, we see a strong enrichment for possible unannotated promoters: 21 overlap a ChIP–chip binding site for factors associated with transcription initiation (PolII, Taf250, TFIIB). Elements with high conservation also appear to contain possible unannotated promoters, with 12 regions overlapping such a binding site. In contrast, among the 50 putative distal elements with the highest *F*, only three overlap such a binding site. This suggests that, although strong signals such as GC content and constraint are still likely to play a role, true distal elements may be better characterized by the subtler, weaker signals proxied by *F*.

Discriminating ENCODE DNaseI hypersensitive sites with ESPERR

A large portion of the ENCODE regions in several cell lines have been assayed for hypersensitivity to the nuclease DNaseI, often used as marker for transcriptional regulatory elements (http://genome.ucsc.edu/encode; J. Stamatoyannopoulos, pers. comm.). From their data, we extracted a set of high-confidence positive calls (empirical *P*-value < 0.001 and plate quality > 0.5 in any cell line; 369 elements), and high-confidence negative calls (empirical *P*-value > 0.1 for all cell lines with plate quality > 0.5, and no overlap with other ENCODE functional elements as compiled by the ENCODE Multi-species Sequence Analysis group (http://genome.ucsc.edu/encode; E.H. Margulies, pers. comm.); 477 elements). Prior work on predicting DNaseI hypersensitive sites with a linear support vector machine (SVM) based on short motifs in the primary genomic sequence (length 1–6, ignoring strand) showed good performance (Noble et al. 2005). Using their methods and training data, we were able to confirm their reported success rate of ~85%. However, applying this approach to the ENCODE data set achieves a success rate of ~64%, suggesting that this more comprehensive set of sites is substantially more difficult to discriminate.

We applied ESPERR to this data set, using the same seven-species alignments as for the RP scores. Training data consisted of 319 positive elements and 379 negative elements with sufficient alignments, prepared as was done for RP scores (except that these elements were not chopped to 100-column segments because the training sets are larger and the elements are of less variable length). The procedure identified an encoding to 18 symbols, which achieved a success rate of ~80%. Thus, for this more comprehensive data set, the additional information available in multiple alignments and captured by ESPERR achieves substantially better performance than does a linear SVM using sequence motifs.

We also computed the ESPERR RP scores for these elements. Approximately 72% of the negative elements have a negative RP score; however, only ~54% of the positive elements have a positive RP score. This suggests that while hypersensitivity to DNaseI may be a marker for regulatory function, the sites identified by the ENCODE project contain regulatory elements that are substantially different from those in our RP training data, or perhaps elements of a different type entirely.

Identifying conserved regions with developmental enhancer activity

The VISTA Enhancer Browser (http://enhancer.lbl.gov) contains 253 conserved regions that have been tested for consistent enhancer activity in transgenic mouse embryos. A region was declared positive (validated) if at least three embryos showed the same pattern of expression for that element. Here, ESPERR produces a score to predict which of the numerous other conserved regions in the genome would be validated by this assay. For positive and negative training sets, we used 108 validated and 138 nonvalidated regions (a small number of regions with ambiguous results were excluded).

Because both the positive and negative training sets for this problem consist of highly conserved elements, alignments spanning a much deeper evolutionary tree were used as compared with the previous applications. Specifically, we used alignments of human, mouse, opossum, chicken, frog, zebrafish, and pufferfish. Training elements were not chopped, and alignment columns with at most three missing species were considered valid, with at least 50 such columns required for an element to be used, resulting in a positive set of 108 elements covering 143,688 human bases and a negative set of 134 elements covering 165,272 human bases. ESPERR identified an encoding to 15 symbols and yielded a very good cross-validation success rate of ~83%. Thus, using our method to score conserved elements for potential enhancer activity could greatly increase the rate of discovery and validation of new conserved embryonic enhancers.

Discussion

We have presented ESPERR, a method to learn encodings of multiple alignments that retain useful information for a chosen classification problem. We have shown excellent performance for predicting three different types of functional elements, each of which involves a binary (e.g., positive vs. negative) classification performed by log-odds, based on variable-order Markov models. Moreover, ESPERR can be used to find useful alignment encodings for other types of binary and nonbinary predictions. We have applied ESPERR using a multiway VOMM-based classifier to successfully discriminate among tissue-specific promoters, ubiquitous promoters, and nonpromoter regions (J. Taylor, N.D. Trinklein, R.C. Hardison, W. Miller, F. Chiaromonte, in prep.). We have also begun exploring the application of our method to gene prediction. Alignment encodings have already been used for gene prediction; for example, TWINSCAN encodes positions as match, other aligned, and unaligned, and estimates models over the encoded sequence (Korf et al. 2001). ESPERR may be able to find effective encodings of multiple "informant" species in related gene-prediction algorithms (e.g., N-SCAN, Gross and Brent 2006). The first stage of ESPERR could also be employed to create reduced representations of multiple alignments for analyses where there is not a natural performance metric for driving the iterative search. One such application is the identification of encodings for the unsupervised characterization of alignments from highly conserved sequences (Bejerano et al. 2004).

ESPERR-based RP scores have proven effective for identifying enhancer elements. Wang et al. (2006) identified 75 regions having a positive RP score as well as matches to the binding site motif for the essential erythroid transcription factor GATA-1. They tested these regions with reporter gene assays in transiently transfected human K562 cells and/or after site-directed integration into murine erythroleukemia cells, and found that regions with high RP score were validated frequently (at least 50%), with even higher validation rates at higher RP scores. In contrast, segments with low RP tended to be inactive.

ESPERR is most appropriate when the loci in question are under selection among the species examined, and at the very least requires that the loci can be aligned (although for all examples presented here we lose a small number of training sequence because of lack of sufficient alignment). For most applications, including those described in this paper, elements do not necessarily exhibit strict nucleotide-level conservation. For example, binding sites in regulatory elements may change relative order or motif (Ludwig et al. 1998; Dermitzakis and Clark 2002; Costas et al. 2003). Also, some elements may only be functional in a specific lineage—see, for instance, studies by Valverde-Garduno et al. (2004) on lineage-specific hypersensitive sites in the *GATA1* locus in humans and mice. However, as long as the elements retain sufficient alignability, ESPERR can still achieve very good performance: In fact, our method can tolerate some degree of local change and even capture such change if it occurs with a consistent pattern.

To infer the ancestral base distribution, ESPERR extends traditional nucleotide substitution models by treating gaps like a fifth "nucleotide." While this extension has been used effectively (McGuire et al. 2001), it is naive, in that it treats indels affecting multiple consecutive positions as multiple independent events, and thus is overly sensitive to all but very short indels. Nonetheless, the extended HKY model works well in ESPERR, most likely because it is combined with a classifier that incorporates context and thus captures dependencies among neighboring sites. More sophisticated modeling of indels for ancestral distribution inferences would integrate naturally into our procedure, and we expect this to become more important as we apply ESPERR using other classifiers, as well as to unsupervised classification (clustering) problems.

For the applications presented here we used ESPERR on alignments of at most seven species. Further increasing the number of species could add predictive power in some problems; our method can easily scale to incorporate more sequences, and because of the way we handle missing data, these could be picked from the many low-coverage genomic sequences currently becoming available. However, care must be taken in selecting what species to use. Very low coverage genomes may in some cases add more noise than exploitable signals, and, in general, the type of functional elements under consideration should dictate species selection (McAuliffe et al. 2005). For example, if elements are not expected to be under very strong constraint, comparisons should be restricted to closely related species.

The intense efforts to characterize and improve predictions of regulatory regions and other functional intervals in the genome are yielding many helpful resources. Biochemical assays of protein binding and chromatin modifications at high resolution, predictions of clusters of conserved transcription factor binding sites, refined estimates of nucleotides under constraint, and other experimental and computational efforts provide a plethora of resources from which investigators can build hypotheses to test. The approach described here (ESPERR) differs from other methods in its emphasis on training to discover both strong and weak signals in alignments, and in its broad applicability—as signals can be learned to discriminate potentially any functional classes for which training data are available. ESPERR can be applied to new sets of functional elements, such as those explored by the ENCODE project, to generate genome-wide predictions for many functional classes. Future efforts to better understand the many subtle signals discovered by ESPERR should provide new insights into the mechanisms underlying specific functions, which could then be tested experimentally. Another exciting challenge is to combine discriminatory methods like ESPERR with other bioinformatic predictions of functional regions to improve accuracy.

Methods

To infer the ancestral base probability distribution corresponding to a given alignment column we use Felsenstein's algorithm (Durbin et al. 1998; Mayrose et al. 2004). For all the applications presented here we allowed up to three species to be missing for any column, in which case those leaves of the tree were left out of the inference (treated as "Felsenstein wildcards"). To estimate the probability of possible substitutions over each branch of the tree (required for the inference), we assumed a continuous time Markov process in which a rate matrix specifies the instantaneous rate of each type of substitution event. We used the rate matrix parameterization provided by the HKY model of Hasegawa et al. (1985) consisting of equilibrium probabilities for the four bases (π_A, π_C, π_G, π_T), and the ratio between the rates of transitions and transversions (κ). We extended this model to accommodate gaps as if they were a fifth nucleotide, introducing an additional equilibrium probability (π_{Gap}) and rate ratio (gaps to transversions; σ). These parameters are estimated using the Expectation Maximization algorithm implemented in the PHAST software package (Siepel and Haussler 2004b). For the applications presented in this paper, we fix the tree topology as that determined by the ENCODE MSA group (http://genome.ucsc.edu/encode; E.H. Margulies, pers. comm.), and run the estimation on a random sample of genome-wide alignments. More details on ancestral distribution inference are provided in the Supplemental material.

The novel clustering algorithm underlying the first stage of ESPERR groups alignment columns agglomeratively, based on distance between corresponding ancestral distributions and their frequency (occurrence counts for columns create a frequency distribution over ancestral distributions). For distance calculations, each cluster is represented by a centroid defined as the "average ancestral distribution" (weighted with frequencies). To preserve the neighborhood structure, at each stage of the agglomeration we consider merging each cluster with its nearest neighbor (Euclidean distance between centroids). To preserve the frequency distribution, the merger that is accepted at each stage is the one that maximizes the mutual information between the distributions before and after merging (in practice this is equivalent to accepting the merger with the maximum entropy; see Supplemental material). Because the algorithm is based on entropy, clusters must not have zero frequency. Thus, we perform an initial preclustering, grouping columns that never occur or are very seldom in the training data (occurring less than five times) with their nearest neighbor (having five or more occurrences). The agglomeration is terminated once a desired number of clusters is reached; for all applications presented here we have stopped at 75; a small enough encoding for the search to fit VOMMs with some power, yet large enough to allow it substantial flexibility.

The second stage of ESPERR—the heuristic search—generates candidate encodings, accepts the best based on a figure of merit (FOM), and repeats until an optimal encoding is found. The FOM is the fraction of elements in the training data correctly classified under cross-validation and does not include "unclassifiable" elements (those falling between the highest scoring negative training element and the lowest scoring positive one; see Supplemental material). The search is initialized with the encoding determined by agglomerative clustering in the first stage. We refer to the symbols (groups) produced by clustering as "atoms," because they are never split during the search. In each search iteration, candidate encodings are generated from the current encoding by either merging two symbols or extracting an atom from one of the symbols. We evaluate only a random sampling of moves of each type (50 and 30, respectively, for the applications presented here), which reduces computations while still producing reasonable moves with high probability. To improve the efficiency of the search we introduce two heuristics. First, since large encodings require more parameters, they are more susceptible to overfitting and thus score more elements in the unclassifiable range, reducing the FOM. Consequently, the search has a strong preference for small encodings, and it is possible that evaluating single atom extractions will not be enough to by-pass local optima. To overcome this, if the FOM does not increase over 20 consecutive iterations, we consider only extractions for 5 consecutive steps, which allows us to move out of local optima through poorer performing, larger encodings. Second, it is possible for the search to make bad moves (mergers or extractions), which then take a long time to be reversed. To recover efficiency, we add a "restarting" heuristic: If we proceed for 50 iterations without reaching an encoding better than the best seen so far, we restart the search at that best encoding. Termination is similar but extends to a much larger number of iterations—we stop when 1000 iterations fail to find a better encoding and adopt that best encoding as the final one. More details are provided in the Supplemental material.

In all applications presented in this paper, elements are classified based on the sign of a log-odds score, which compares their probabilities under two VOMMs estimated on the positive and negative training set. VOMMs are similar to fixed-order Markov chains; however, they can use a variable number of previous positions ("context") when determining the transition probability at a given position in a string (up to a fixed maximal context length, here 2). Our implementation includes in the model contexts observed at least some number of times (here 5) in the training data (pruning). To allocate probability to patterns never seen in the training data we use a "discount" rule (smoothing). This and other details of the VOMM variant used in ESPERR are provided in the Supplemental material.

Training sets were prepared using the 17-species MultiZ alignments from the UCSC Genome Browser (Karolchik et al. 2003; Blanchette et al. 2004). For the computation of RP scores and hypersensitive site predictions we used the subset of mammalian species in these alignments with higher sequence quality; namely, human (hg17), chimpanzee (panTro1), macaque (rheMac2), mouse (mm7), rat (rn3), dog (canfam2), and cow (bosTau2). For prediction on highly conserved regions with embryonic enhancer activity, we used a subset of species spanning a larger evolutionary distance; namely, human (hg17), mouse (mm7), opossum (monDom2), chicken (galGal2), frog (xenTro1), zebrafish (danRer3), and pufferfish (fr1). Alignments corresponding to each element of a training set were extracted. Gaps between alignment blocks were annotated as such, and all other gaps (including complex insertion/deletion events and gaps to long N stretches in sequences) were annotated as missing data. For the RP score application only, the training sets were then chopped to 100-column alignment segments. For all applications, training elements were also required to have at least 50 good alignment columns (those having three or fewer missing species) to be included.

The ancestral base inference, agglomerative clustering, and iterative search were implemented in Python with performance-critical portions implemented in C—these include code for estimating VOMMs and scoring alignment segments, which are run to perform cross-validation over thousands of candidate encodings. The simple pruning and smoothing rules used in VOMM estimation are amenable to efficient implementation, making the iterative search tractable. The search can be spread over multiple cluster nodes using MPI. Running time for the search varies depending on specific data and the random component; for the RP application convergence is generally achieved in ~10,000 iterations, requiring a day on a 2-Ghz Athlon machine—substantially less on a small cluster.

Acknowledgments

This work was supported by NIH grants HG02238 (W.M.) and DK65806 (R.H.). We thank the ENCODE Consortium for making their data publicly available, the ENCODE Transcriptional Regulation, Multispecies Sequence Analysis, Genes and Transcripts, and Chromatin and Chromosomes groups, and the NISC Comparative Sequencing Program for specific data sets used. We also thank L. Elnitski for her work on the collection of human regulatory elements used in training RP scores; L. Pennacchio and E. Rubin for making their enhancer assay results publicly available; M. Brent, M. Dermitzakis, T. Columbo, T. Furey, N. Trinklein, and all of the participants of the 2nd Barbados Workshop on Genomics and Gene Regulation who provided comments, feedback, or discussion that improved this work. Finally, we wish to thank the referees of a previous version of this manuscript for their insightful and motivating comments.

References

Bejerano, G., Haussler, D., and Blanchette, M. 2004. Into the heart of darkness: Large-scale clustering of human non-coding DNA. *Bioinformatics* **20:** I40–I48.

Bieda, M., Xu, X., Singer, M.A., Green, R., and Farnham, P.J. 2006. Unbiased location analysis of E2F1-binding sites suggests a widespread role for E2F1 in the human genome. *Genome Res.* **16:** 595–605.

Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D., et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14:** 708–715.

Buhlmann, P. and Wyner, A. 1998. Variable length Markov chains. *Ann. Statist.* **27:** 480–513.

Cawley S., Bekiranov S., Ng H.H., Kapranov P, Sekinger E.A., Kampa D., Piccolboni A., Sementchenko V., Cheng J., Williams A.J., et al. 2004. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116:** 499-509.

Cooper, S.J., Trinklein, N.D., Anton, E.D., Nguyen, L., and Myers, R.M. 2006. Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. *Genome Res.* **16:** 1–10.

Costas, J., Casares, F., and Vieira, J. 2003. Turnover of binding sites for transcription factors involved in early *Drosophila* development. *Gene* **310:** 215–220.

Dermitzakis, E.T. and Clark, A.G. 2002. Evolution of transcription factor binding sites in Mammalian gene regulatory regions: Conservation and turnover. *Mol. Biol. Evol.* **19:** 1114–1121.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. 1998. *Biological Sequence Analysis*. Cambridge University Press, Cambridge, UK.

Elnitski, L., Hardison, R.C., Li, J., Yang, S., Kolbe, D., Eswara, P., O'Connor, M.J., Schwartz, S., Miller, W., and Chiaromonte, F. 2003. Distinguishing regulatory DNA from neutral sites. *Genome Res.* **13:** 64–72.

Gross, S.S. and Brent, M.R. 2006. Using multiple alignments to improve gene prediction. *J. Comput. Biol.* **13:** 379–393.

Hasegawa, M., Kishino, H., and Yano, T. 1985. Dating of the human–ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22:** 160–174.

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., et al. 2003. The UCSC Genome Browser Database. *Nucleic Acids Res.* **31:** 51–54.

King, D.C., Taylor, J., Elnitski, L., Chiaromonte, F., Miller, W., and Hardison, R.C. 2005. Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. *Genome Res.* **15:** 1051–1060.

Kolbe, D., Taylor, J., Elnitski, L., Eswara, P., Li, J., Miller, W., Hardison, R.C., and Chiaromonte, F. 2004. Regulatory potential scores from genome-wide three-way alignments of human, mouse, and rat. *Genome Res.* **14:** 700–707.

Korf, I., Flicek, P., Duan, D., and Brent, M.R. 2001. Integrating genomic homology into gene structure prediction. *Bioinformatics* **17:** S140–S148.

Ludwig, M.Z., Patel, N.H., and Kreitman, M. 1998. Functional analysis of eve stripe 2 enhancer evolution in *Drosophila*: Rules governing conservation and change. *Development* **125:** 949–958.

Margulies, E.H., Blanchette, M., NISC Comparative Sequencing Program, Haussler, D., and Green, E.D. 2003. Identification and characterization of multi-species conserved sequences. *Genome Res.* **13:** 2507–2518.

Mayrose, I., Graur, D., Ben-Tal, N., and Pupko, T. 2004. Comparison of site-specific rate-inference methods for protein sequences: Empirical Bayesian methods are superior. *Mol. Biol. Evol.* **21:** 1781–1791.

McAuliffe, J.D., Jordan, M.I., and Pachter, L. 2005. Subtree power analysis finds optimal species for comparative genomics. *Proc. Natl. Acad. Sci.* **102:** 7900–7905.

McGuire, G., Denham, M.C., and Balding, D.J. 2001. Models of sequence evolution for DNA sequences containing gaps. *Mol. Biol. Evol.* **18:** 481–490.

Noble, W.S., Kueh, S., Thurman, R., Yu, M., and Stamatoyannopoulos, J.A. 2005. Predicting the in vivo signature of human gene regulatory sequences. *Bioinformatics* **21:** 338–343.

Siepel, A. and Haussler, D. 2004a. Computational identification of evolutionarily conserved exons. *Proc. 8th Annual Int'l Conf. on Research in Computational Biology*, pp. 177–186. RECOMB, San Diego, CA.

Siepel, A. and Haussler, D. 2004b. Combining phylogenetic and hidden Markov models in biosequence analysis. *J. Comput. Biol.* **11:** 413–428.

Siepel, A., Bejerano, G., Pederson, J.S., Hinrichs, A., Hou, M., Rosenbloom, K., Clawson, J., Spieth, J., Hillier, L.W., Richards, S., et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15:** 1034–1050.

Tompa, M., Li, N., Bailey, T.L., Church, G.M., De Moor, B., Eskin, E., Favorov, A.V., Frith, M.C., Fu, Y., Kent, W.J., et al. 2005. Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.* **23:** 137–144.

Valverde-Garduno, V., Guyot, B., Anguita, E., Hamlett, I., Porcher, C., and Vyas, P. 2004. Differences in the chromatin structure and *cis*-element organization of the human and mouse GATA1 loci: Implications for *cis*-element identification. *Blood* **104:** 3106–3116.

Wang, H., Zhang, Y., Cheng, Y., Zhou, Y., King, D.C., Taylor, J., Chiaromonte, F., Kasturi, J., Petrykowska, H., Gibb, B., et al. 2006. Experimental validation of predicted mammalian erythroid *cis*-regulatory modules. *Genome Res.* (this issue).

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.

Received August 14, 2005; accepted in revised form July 31, 2006.