Predicting site-specific human selective pressure using evolutionary signatures

Javad Sadri1,2, Abdoulaye Banire Diallo3 and Mathieu Blanchette1,∗

1School of Computer Science, McGill University, 3630 University, Montreal, QC, Canada H3A 2B2, 2Department of Computer Engineering, Faculty of Engineering, University of Birjand, Birjand, Iran and 3Department of Computer Science, Université du Québec à Montréal, Montreal, QC, Canada H3C 3P8

ABSTRACT

Motivation: The identification of non-coding functional regions of the human genome remains one of the main challenges of genomics. By observing how a given region evolved over time, one can detect signs of negative or positive selection hinting that the region may be functional. With the quickly increasing number of vertebrate genomes to compare with our own, this type of approach is set to become extremely powerful, provided the right analytical tools are available.

Results: A large number of approaches have been proposed to measure signs of past selective pressure, usually in the form of reduced mutation rate. Here, we propose a radically different approach to the detection of non-coding functional regions: instead of measuring past evolutionary rates, we build a machine learning classifier to predict current substitution rates in human based on the inferred evolutionary events that affected the region during vertebrate evolution. We show that different types of evolutionary events, occurring along different branches of the phylogenetic tree, bring very different amounts of information. We propose a number of simple machine learning classifiers and show that a Support-Vector Machine (SVM) predictor clearly outperforms existing tools at predicting human non-coding functional sites. Comparison to external evidences of selection and regulatory function confirms that these SVM predictions are more accurate than those of other approaches.

Availability: The predictor and predictions made are available at http://www.mcb.mcgill.ca/~ Blanchem/sadri.

Contact: blanchem@mcb.mcgill.ca

1 INTRODUCTION

One of the central goals of comparative genomics is to use the comparison of genomes from different species to delineate functional regions in those genomes. Since most functional regions are under various degrees of negative selection, they tend to exhibit higher interspecies sequence conservation than their flanking neutral regions. Although mutations occur randomly in the genome, the rate at which they become fixed in a population depends, among other things, on the fitness of the mutated individuals (Moran and Pierce, 1962). Whereas non-functional regions of a given genome based on the observed number of mutations they have undergone during the evolution of a set of species. This approach has been used with success to identify all kinds of functional regions of the human and drosophila genomes (among others), including protein-coding genes (Dewey et al., 2004; Gross and Brent, 2006), non-coding RNA genes (Dowell and Eddy, 2006; Pedersen et al., 2006) and transcription factor binding sites (Loos and Ovcharenko, 2004; Moses et al., 2004).

A number of generic approaches have been developed to identify sites that appear to be under negative selection based on comparative genomics. After identifying and aligning orthologous regions from two or more related species [e.g. using Mulitz (Blanchette et al., 2004a) or MLAGAN (Brudno et al., 2003)], one can scan the alignment to identify regions where the sequence conservation is higher than expected. Early approaches dealt with a small number of genomes and evaluated conservation based on a sliding window strategy (Boffelli et al., 2003; Frazer et al., 2004; Margulies et al., 2003). More recent approaches such as the very popular PhastCons approach (Siepel et al., 2005) use a tree hidden-Markov model (HMM) to assign sites to one of several rate categories. A strength of this type of approach is that it can take advantage of the fact that most functional regions involve several consecutive sites, while avoiding the drawback of using a fixed-size window. This family of approaches has been used to identify likely functional regions in all kinds of species, including vertebrates (Margulies et al., 2003, 2007; Thomas et al., 2003), yeast (Kellis et al., 2003) and drosophila (Stark et al., 2007), among others, although most of these regions remain functionally uncharacterized to date.

With the number of genomes being sequenced quickly increasing, the prospect of accurately measuring evolutionary rates and selective pressure at individual sites became an achievable goal. Several approaches were developed for this purpose, including SCONE (Asthana et al., 2007) and GERP (Cooper et al., 2005), which both attempt to evaluate selective pressure on a site-by-site basis. What these approaches lose in specificity because of the fact that they do not combine signals from neighboring sites, they gain in sensitivity, since they are able to detect very small regions under selection.

A significant drawback of all these approaches is their assumption that the mutation rate in a given region or at a given site has not changed during the evolution of the set of species considered. Although negative selection in species ancestral to human and in sister species is clearly informative about selection in human, it is only an indirect predictor of it. Past results do not guarantee more mutations become fixed in non-functional regions than in functional regions. This principle is the foundation of phylogenetic footprinting, whereby one can hope to distinguish functional from non-functional regions of a given genome based on the observed number of mutations they have undergone during the evolution of a set of species. This approach has been used with success to identify all kinds of functional regions of the human and drosophila genomes (among others), including protein-coding genes (Dewey et al., 2004; Gross and Brent, 2006), non-coding RNA genes (Dowell and Eddy, 2006; Pedersen et al., 2006) and transcription factor binding sites (Loos and Ovcharenko, 2004; Moses et al., 2004).

A number of generic approaches have been developed to identify sites that appear to be under negative selection based on comparative genomics. After identifying and aligning orthologous regions from two or more related species [e.g. using Mulitz (Blanchette et al., 2004a) or MLAGAN (Brudno et al., 2003)], one can scan the alignment to identify regions where the sequence conservation is higher than expected. Early approaches dealt with a small number of genomes and evaluated conservation based on a sliding window strategy (Boffelli et al., 2003; Frazer et al., 2004; Margulies et al., 2003). More recent approaches such as the very popular PhastCons approach (Siepel et al., 2005) use a tree hidden-Markov model (HMM) to assign sites to one of several rate categories. A strength of this type of approach is that it can take advantage of the fact that most functional regions involve several consecutive sites, while avoiding the drawback of using a fixed-size window. This family of approaches has been used to identify likely functional regions in all kinds of species, including vertebrates (Margulies et al., 2003, 2007; Thomas et al., 2003), yeast (Kellis et al., 2003) and drosophila (Stark et al., 2007), among others, although most of these regions remain functionally uncharacterized to date.

With the number of genomes being sequenced quickly increasing, the prospect of accurately measuring evolutionary rates and selective pressure at individual sites became an achievable goal. Several approaches were developed for this purpose, including SCONE (Asthana et al., 2007) and GERP (Cooper et al., 2005), which both attempt to evaluate selective pressure on a site-by-site basis. What these approaches lose in specificity because of the fact that they do not combine signals from neighboring sites, they gain in sensitivity, since they are able to detect very small regions under selection.

A significant drawback of all these approaches is their assumption that the mutation rate in a given region or at a given site has not changed during the evolution of the set of species considered. Although negative selection in species ancestral to human and in sister species is clearly informative about selection in human, it is only an indirect predictor of it. Past results do not guarantee
future performance in the stock market—neither does past selection necessarily implies current selection in human. There is considerable evidence that certain types of short functional regions such as transcription factor binding sites indeed turn-over quickly, causing a given locus to evolve neutrally in some species and under selection in another (Moses et al., 2006). Recently, Stiedl et al. introduced PhyloP (Pollard et al., 2010), an impressive package that allows the detection of sites under negative or positive selection, while allowing changes in evolutionary rates over the branches of the phylogenetic tree (Stiedl et al., 2006). To our knowledge, this represents the best approach available to date for identifying individual sites under selection.

In this article, we propose a very different approach to the identification of functional sites in a given reference genome. Instead of grouping previously selected sites at a site, we use the evolutionary history at a site (and at neighboring positions) to predict current (or very recent) selective pressure. More specifically, using whole-genome multiple alignments for a collection of 44 vertebrates (Margulies et al., 2007; Miller et al., 2007), we first reconstruct ancestral sequences (Blanchette et al., 2004b; Diallo et al., 2010) and identify evolutionary events at each site and along each branch of the tree. We then train machine learning classifiers to use the evolutionary history of a region to estimate the likelihood that it will have undergone a substitution in recent human history (since human-chimp divergence). Sites that are strongly predicted to remain conserved are likely functional sites in human. This approach has a number of key advantages. First, it does not require modeling or making assumptions about the evolution of functional regions of the genome process. Second, it is free to use whichever features of evolutionary history, and to weight each feature as it wishes, in order to maximize the accuracy of the selective pressure prediction in human. A predictor may for example weigh more heavily conservation along primate lineages than along more distantly related lineages. It may also weigh more or less heavily conservation at neighboring sites.

This article is organized as follows. We first study how individual evolutionary events along specific branches of the phylogenetic tree relate to the probability of substitution in human. We then propose a set of machine learning classifiers, including a Naive Bayes classifier, a support-vector machine (SVM) classifier and show that some of them outperform the Bayes classifier, a context-dependent substitution model. Blanchette et al. (2004b) and Diallo et al. (2007) have previously shown, using simulations, that the expected accuracy of this reconstruction can be as high as 99% base-by-base accuracy for early eutherian ancestors such as the Bosromeria ancestor, and that it is above 90% for almost all other ancestral nodes of the eutherian phylogeny.

2.2 Selection of unambiguously human conserved and mutated sites

A subset of human non-protein-coding genomic sites that could be unambiguously labeled as having undergone a substitution along the human branch since the human–chimp ancestry were identified. Several filters were applied to ensure that this set of sites is as enriched as possible for bona-fide mutated sites, and not the result of alignment errors. For site i to be considered as eligible, the following rules were applied:

- Site i must be conserved along the branches leading to the orangutan, gorilla and chimp.
- Sites i−1 or i+1 must be perfectly conserved between the human–chimp ancestor and both human and chimp.
- If site i is not conserved from the human–chimp ancestor to human, then it must be a transition (purine-to-purine or pyrimide-to-pyrimidine).
- Transversions are not considered, as they occur at different rates than transitions, which may introduce biases.
- If the human–chimp ancestral nucleotide is C, then it must not be followed by a G. This avoids C → TpG substitutions, whose elevated rate (Siepel and Haussler, 2004) may bias our analyses.

Finally, sites that satisfy all the above requirements are either called conserved or mutated, based on the event that took place between the human and chimp branches.

When applied to human chromosome 22, these filters resulted in the identification of ~41 600 mutated non-coding sites, and roughly one hundred times more non-coding conserved sites. A subset of 41 600 conserved sites was then selected randomly from all conserved sites, to form a balanced set of 83 200 examples, which was divided into a 50 000-example training set and a 33 200-example test set.

2.3 Feature set definition and extraction

Various approaches were considered to encode the history matrix \( H \) into a set of features that can be used to train classifiers. Those that resulted in the best results were the following.

**Feature Set 1:** This contains the number of conservations and substitutions at each position within the window: \( F_i^a(p) = \sum I_{ihp, p, c} \), \( F_i^s(p) = \sum I_{ihp, p, s} \). This feature set thus contains \( 2(2n+1) \) integer features. Insertions and deletions are not explicitly accounted for in this feature set, but the presence of a large number of these events reduces the counts of conservations and substitutions and thus impacts the values of the features. An alternate feature set where all five types of events were counted separately was also evaluated but produced slightly worse results, probably because the increase in the size of the feature set is not counterbalanced by the informativeness of the new features.

**Feature Set 2:** This takes an orthogonal approach and counts events for each branch of the tree rather than for each site. Specifically, \( F_i^b(b) = \sum p \sum W(p, b) I_{ihp, p, b} \). This feature set thus includes \( 85 \times 2 = 170 \) integer features, irrespective of window size.

A number of alternate feature sets have been considered including using the matrix \( H \) itself (or part of it corresponding to a smaller window around
the site). However, no predictor was able to take advantage of the richness of the provided information, probably in part because of the huge feature space involved.

2.3.1 Other measures of sequence conservation. The PhyloP package (Pollard et al., 2010) includes implementations of PhastCons (Siepel et al., 2005), GERP (Cooper et al., 2005), evolutionary rate likelihood-ratio test (LRT), as well as the PhyloP-SCORE. Each was run on the MultiZ 45-species multiple alignment obtained by removing the human sequence from the alignment, so that human conservation/mutations do not affect the scores produced. Default parameters were used for each algorithm. Each program outputs a score for each site in the dataset.

2.4 Classifiers
Our Naive Bayes approach was implemented in the straightforward manner, using uniform pseudocounts to estimate posterior probability features of the given class label. This classifier can be trained in a few seconds and makes predictions equally quickly.

The k-nearest neighbor approach used the distance between feature vectors to identify neighbors. We obtained best results on Feature Set 2, using $k=0$. Fairly large values of $k$ (number of neighbors) produced better results, $k=400$ was used for the results reported here. This approach is substantially slower as the running time required to classify a single test example is proportional to the number of training examples. Nonetheless, it runs in a few minutes on a standard desktop computer.

We used the SVMlight package (Joachims, 1999) from a Matlab interface to train classifiers for Feature sets 1 and 2. While our efforts to reduce overfitting failed for Feature set 2 (which contains a large number of features), we obtained our best results on Feature Set 1, with window size $w=0$ and $k=1$. We obtained our best results using a radial basis function (RBF) with $\gamma = 1.5$ (kernel parameter), $C=100$ (trade-off between training error and margin) and $J=0.7$ (relative cost of errors on positive examples). This is type of classification problem where the examples are fairly poorly separable results in a very large number of support vectors (more than half the training examples are retained as support vectors), which impacts running time and generalization. However, for Feature Set 1, the training and testing errors were essentially equal. Training on our 50,000 example took ~1h, and predicting on the 33,200 test examples took <10 min.

3 RESULTS
Our goal is to develop a machine learning predictor that will estimate the probability that a given site of the human–chimpanzee ancestor will undergo a substitution along the branch leading to human, given the complete set of evolutionary events that took place in that region during vertebrate evolution (but excluding recent human evolution). Because the only way to predict positions where substitutions will become fixed is to predict their effect on fitness, sites that are predicted to remain conserved with high probability are likely to be functional ones.

It may seem counter-intuitive that in order to predict the selective pressure on a site in human, we purposefully ignore evolutionary events having taken place along that branch. In fact, although events along the human branch are excluded from our feature set, they play a very important role in our training set, as they form the label of each training example. We also underscore the fact that, although a substitution along the human branch is a very strong indicator of the absence of selective pressure at that site, such events are also extremely rare (~0.5% of sites). Thus, events along the human branch are much more productively used as labeled of a (artificially balanced) training set than as features.

Fig. 1. Phylogenetic tree of the 44 vertebrate species considered to predict regions under selective pressure in human.

3.1 Training data
Our study is based on a balanced dataset consisting of all 41,600 non-coding sites from human chromosome 22 with unambiguous substitution along the human branch since the human–chimp ancestry and equally many non-coding sites with no substitutions, randomly sampled from the same chromosome. Coding regions were excluded from consideration because they can accurately detected using a variety of approaches (see [Siepel, 2007] and references therein) and they obey fairly different rules than non-coding sites. A set of rules were applied to ensure that apparent human substitutions are not simply due to alignment errors or to increased substitution rate caused by sequence context effects (see Section 2). We then inferred the evolutionary history of each site, together with a 501 bp window centered on it, using a multiple sequence alignment of the genomes of 44 vertebrate species (Miller et al., 2007) (Fig. 1) and a maximum likelihood ancestral sequence inference approach for both substitutions and indels (Diallo et al., 2007, 2010) (see Section 2). The full history of a site $i$ was then encoded as a matrix $H_i$, with 85 columns (corresponding to the 85 branches of the phylogenetic tree, excluding that leading to human) and 501 rows (corresponding to the 501 human sites in the window surrounding site $i$, where the entry $H_i(b,b)$ for branch $b$ at relative position $j \in \{-250, \ldots, -1, 0, 1, \ldots, 250\}$ corresponds to the evolutionary event inferred on that branch at position $i+t$: C(onservation), S(ubstitution), I(nsertion), D(eletion) or G(ap). The ‘Gap’ event denotes the presence of a gap in both the ancestor and descendant of branch $b$. Our goal is to assess the extent to which the fate of site $i$ along the human branch can be predicted from the matrix $H_i$.

3.2 Individual feature informativeness
We first measured how informative are individual events along each branch of the tree. This information can be measured by several means. First, we consider the question of whether the presence of orthologous bases in a given species (extant or ancestral) affects the likelihood of a conservation event along the human branch. A human site may have no detectable ortholog in a given
species \( s \) for several reasons: (i) Site \( i \) was inserted after the last common ancestor of \( x \) and human (denoted LCA(\( x, \text{human} \)); (ii) Site \( i \) was deleted since the LCA(\( x, \text{human} \)) along the lineage leading to \( x \); (iii) Site \( i \) actually has an ortholog in \( x \), but that and the surrounding sequence have diverged to the point where orthology cannot be detected (or, in the case of ancestral sequences, none of its descendant has a detectable ortholog).

Figure 2a plots the likelihood ratio of human conservation in the presence or absence of an orthologous base on branch \( b \) at site \( i+\delta \):

\[
\log \left( \frac{Pr[H(\text{human},0)=a,H(\text{b,\delta})=b]}{Pr[H(\text{human},0)=a,H(\text{b,\delta})=b]} \right)
\]

As expected, one observes that detectable orthology is relatively uninformative, even for small \( \delta \). This observation is new and quite significant.

Next, we ask whether the actual event (conservation or substitution) taking place at site \( i+\delta \) along branch \( b \) brings any information on the fate of site \( i \) in human:

\[
\log \left( \frac{Pr[H(\text{human},\delta)=b]}{Pr[H(\text{human},\delta)=b]} \right)
\]

Note that the probabilities at different \( \delta \) are the same. For example, in Feature Set 1, the total number of 'C' and 'S' observed along branch \( b \) is the same for all sites, even for large \( \delta \). It appears that the presence of bases with a human ortholog even located 250 bp away from the current site is only marginally less informative than considering orthology at the site itself. This is due to the fact that functional regions and detectable orthology blocks are generally quite large.

We then consider the actual event (conservation or substitution) taking place at site \( i+\delta \) along branch \( b \). For each position \( i \), we record the total number of 'C' and 'S' observed at site \( j \) summed over all branches.

• Feature Set 1 contains 2 \( (2w+1) \) features: for each position \( j \in W(i,w) \), we record the total number of 'C' and 'S' observed at site \( j \) summed over all branches.

• Feature Set 2 contains 2 \( \times 85 \) features: for each branch \( b \), we record the total number of 'C' and 'S' observed along branch \( b \), summed over all sites in \( W(i,w) \).

The decision to exclude counts of 'T', 'D' and 'G' events was made on the basis that these events are relatively rare in mammals and contribute little to the prediction problem while greatly increasing the number of features. Note, however, that although the counts of 'T', 'D' and 'G' events are not individually present as features, their presence is nonetheless reflected in the feature set through the counts of 'C' and 'S'. For example, in Feature Set 1, the total number of 'T', 'D' and 'G' events at a site is given by \( 85 - m(C) - m(S) \).

3.4 The difficulty of the classification problem

We start by illustrating the difficulty of the classification task at hand, in order to calibrate our expectations. The fraction of human sites under selection has been estimated (based on human-mouse alignments) to be at least 5% (The International Mouse Genome
Fig. 2. Individual feature informativeness. (a) Log-likelihood ratio of human conservation in the presence or absence of orthologous bases in the given species and at the given offset from the considered position. (b) Log-likelihood ratio of human conservation in the presence of a conservation versus a substitution along the given branch and at the given offset. For both (a) and (b), ratios for non-mammalian species are too noisy and are not shown. (c) Mutual information between the presence of human conservation and the event along each branch at each offset.
Sequencing Consortium, 2002); more recent estimates place it between 4% and 7% (Margulies et al., 2007). If we assume that the substitution rate in regions under selection is on average half the rate in neutral regions and that the probability of a neutral site mutating between the human-chimp ancestor and human is 0.5%, we obtain that roughly 2.56% of mutated sites are under selection, while this fraction jumps to ~5.01% among conserved sites. Thus, both mutated and conserved sites are rich in non-functional sites, but conserved sites are almost two times richer in selected sites.

Consider, for the sake of example, a balanced training set of 1000 human-conserved and 1000 human-conserved sites. We expect the mutated sites to contain ~975 neutral sites and 25 selected sites, while the conserved sites should contain ~95 neutral sites and 50 selected sites. This has significant implications on the accuracy that our human-conservation predictor can be expected to obtain. For example, consider a classifier \( \Omega \) that is able to distinguish selected from non-selected sites with 100% accuracy and that bases its prediction of human conservation prediction on this. The classifier \( \Omega \) would predict 25 + 50 = 75 sites as conserved, of which 50 would be correct (67% positive predictive value). It would predict 975 + 95 = 1925 sites as mutated, of which 975 would be correct (50.64% negative predictive value). This is the best classification accuracy we can expect (under the assumption that the substitution rate at selected sites is half that at neutral sites).

Remember, however, that our true goal is not to train a predictor to have similar rates. It has been shown to perform very well at identifying various types of functional sites, in particular larger ones such as exons, enhancers and RNA genes.

### 3.5 Accuracy of predictors of human selection

Our set of 83 200 examples was divided into a training set of 50 000 examples and a test set of 33 200 examples, both with exactly the same number of human-conserved and human-mutated sites. The results reported in this section describe the performance on the test set, which was not used during the training phase.

Several types of classifiers were trained on each type of feature set (see Section 2 for details on training and parameter tuning):

- **Naive Bayes**: classifiers calculate the posterior probability of the example label conditional on the set of observed features, under the (often unrealistic) assumption that features are independent. Despite their relative simplicity, Naive Bayes classifiers have proved remarkably useful in several settings (Rish, 2001).

- **k-nearest neighbor predictors (KNN)** (Shakhnarovich and Indyk, 2005): This simply obtain the probability of an example being positive based on a voting procedure using the \( k \)-nearest (most similar) training examples. This type of classifier can be very accurate for very large datasets, if an appropriate measure of similarity is available.

- **SVM** (Vapnik, 1995): This construct the optimal separating hyperplane between the two input classes to maximize the margin between the examples of the two classes. In the case of non-linearly separable classification problems, SVM uses kernel functions in order to implicitly map the feature vectors to a very high-dimensional feature space to obtain non-linear boundaries. After training SVM, support vectors are identified as those training examples that best define the boundary between the two classes. In the past few years, SVM have shown very good performance in a number of machine learning and pattern recognition problems (Cristianini and Shawe-Taylor, 2000).

All three types of predictors produced poor results when large values of \( w \) were used. The Naive Bayes and KNN classifiers performed best on feature set 2 (with \( w = 0 \)), whereas the SVM-based approaches were unable to handle the large number of features this set contains. On the other hand, the SVM approach was very effective at using the smaller number of features from Feature Set 1 and produced good results for both \( w = 0 \) and 1. Figure 3 shows the positive predictive values (PPV, defined as the ratio of the number of true positive predictions to the number of positive predictions) obtained for each of these classifiers. Because we expect the fraction of functional sites in our balanced training and testing sets to be relatively small (probably around 5 to 10%), we only plot PPVs for prediction thresholds resulting in up to 20% of the test examples being predicted positive. The two SVM predictors (Feature Set 1, \( w = 0 \) or 1) clearly outperforms all other approaches over much of the range of prediction threshold. The Naive Bayes and KNN predictors perform relatively poorly for high-confidence predictions, although they become competitive with the two SVM predictors at lower confidence calls.

In addition to these three types of classifiers trained on our two feature sets, we considered four previously proposed measures of sequence conservation, all implemented in the PhyloP package (Pollard et al., 2010), that aim at detecting sites under selection (although not specifically along the human lineage):

- **PhastCons** (Siepel et al., 2005): this approach identifies genomic regions with reduced mutation rate using a tree-HMM approach that assumes that neighboring sites are highly likely to have similar rates. It has been shown to perform very well at identifying various types of functional sites, in particular larger ones such as exons, enhancers and RNA genes.

- **Genomic Evolutionary Rate Profiling (GERP)** (Cooper et al., 2005): this approach assigns a conservation score to each site in an alignment, independently of neighboring sites. It measures the range of prediction threshold. The Naive Bayes and KNN predictors produce poor results when using the best classification accuracy we can expect (under the assumption that the substitution rate at selected sites is half that at neutral sites).

![Fig. 3. Performance of various previously published measures of sequence conservation (PhastCons, PhyloP-SCORE, GERP, PhyloP-LRT), compared with predictors developed in this article. X-axis: fraction of test examples predicted as positive; Y-axis: positive predictive value (fraction of human-conservation predictions that are indeed human conserved).](https://academic.oup.com/bioinformatics/article-abstract/27/13/i266/182309)
the ‘number of rejected substitutions’, defined as the expected number of substitution per site minus the observed number.

- Likelihood-ratio test (LRT) (Margulies et al., 2007; Pollard et al., 2010): this approach assigns a P-value to the difference in the likelihoods of an observed alignment column under a null model of neutral evolution versus a model with an additional rate parameter that is estimated from the column.

- PhyloP-SCORE (Pollard et al., 2010): similarly to the LRT, the SCORE test compares the hypotheses of neutrality to that of reduced or accelerated rate, but without the need to fit the rate parameter of the alternate hypothesis.

Each of these methods assigns a conservation score to each site \(i\), based either on the evolutionary history of that site alone (in the case of GERP, LRT and SCORE) or based on the evolutionary history of site \(i\) and its surrounding sites (in the case of PhastCons). These conservation scores were calculated after excluding the human sequence from the alignment, to ensure that our class labels (event along human branch) do not taint our feature set.

The accuracy of the predictions made by these four measures of conservation are shown in Figure 3. PhastCons, the only measure that integrates signals from several consecutive positions, performs best for highly confident predictions (top 5% predicted as positive) but its accuracy decreases and quickly becomes worse than the other approaches at slightly more lenient thresholds. This is likely due to its excellent ability to detect relatively large regions under selection, but its inability to detect smaller ones or to identify weakly conserved sites within highly conserved regions. At more lenient thresholds (3–20% of test examples being called positive), PhyloP-SCORE outperforms the other three approaches. Note, however, that the two SVM-based predictors clearly outperform both PhastCons and PhyloP-SCORE over the full range of prediction threshold, often by fairly substantial margins.

### 3.6 Comparison of predictions

Figure 4 shows an example of the predictions made by our SVM predictor (FeatureSet 2, \(w=1\)), compared with PhyloP-SCORE and PhastCons. The most obvious difference is the increased resolution of the SVM and PhyloP approaches, which evaluate each base independently (though in the context of the two flanking bases in the case of SVM), compared with the HMM-based approach of PhastCons, which produces much smoother estimates. Whereas the PhastCons scores are extremely useful to define broad regions of conservation such as exons or enhancers, they are essentially useless when it comes to determining which bases within those regions are actually functional. For example, not all bases within an enhancer are equally constrained, because some are bound by a transcription factors and others not, and because not all bases within a TF binding site are equally important. Thus, PhyloP and SVM predictions provide an extremely valuable finer-grain estimation of selective pressure.

While it is difficult to compare the merits of each predictor based on a single example, Figure 5 provides a comprehensive comparison of the three sets of predictions, based on the ENCODE regions of human chromosome 22 (1.69 Mb in total) (ENCODE-Project-Consortium et al., 2007). The top 10% of the sites predicted by each of the three methods was compared (this fraction is deliberately set slightly higher than the current estimates of the fraction of the genome under selection, to assess the ability of the methods to detect weak selection or selection on isolated sites). Clearly, the PhyloP and SVM approaches yield similar predictions (59% of the sites overlap), and quite different from the PhastCons sites (30–35% overlap only).

The predictions made by each method can be compared based on various external evidences of selection or function that were not part of their training. First, we compared the rate of human polymorphisms [HapMap project, phase III (International HapMap Consortium, 2007)] at sites predicted by each method or combination of methods. As expected, this rate is lowest for...
These regions can be identified by DNAseI digestion followed by PhyloP-SCORE and SVM. The reason for this depletion, which contradicts the relatively poor positive predictive value of Phastcons compared with the other two predictors (Fig. 3), may lie in PhastCons’s ability to identify fairly large regions under weak selection, such as non-coding RNA genes. We also note that the polymorphism rate is slightly lower in sites predicted by the SVM approach only, compared with those predicted by PhyloP, suggesting stronger selection on the former than the latter.

Second, we evaluated the evidence for a regulatory function of the predicted sites. Two types of experimental evidence were considered. Regions of the genome where the chromatin is open makes it possible for transcription factors to bind DNA. These regions can be identified by DNAseI hypersensitive regions, a larger percentage than for PhyloP-SCORE-only sites or PhastCons-only sites. Another type of evidence of regulatory function is the presence/absence of predicted transcription factor binding sites. It is difficult to imagine a more challenging yet fascinating classification problem!

Evolution has been conducting site-specific functionality assays for hundreds of millions of years. The ability to decipher the results of these experiments has and will continue to provide us with a wealth of information about our genome and the impact of mutations therein.

ACKNOWLEDGEMENTS

We thank Tomi Pastinen for useful discussions and the reviewers for their excellent suggestions.

Funding: Genome Canada and NSERC grants (to M.B.); FQRNT scholarship (to J.S.).

Conflict of Interest: none declared.

REFERENCES

Asthana, S. et al. (2007) Analysis of sequence conservation at nucleotide resolution. PLoS Comput. Biol., 3, e254.

Blanchette, M. et al. (2004a) Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res., 14, 708–715.
Margulies,E.H. et al. (2005) An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. Proc. Natl Acad. Sci. USA, 102, 4795–4800.

Margulies,E. et al. (2007) Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. Genome Res., 17, 760–774.

Miller,W. et al. (2007) 28-way vertebrate alignment and conservation track in the UCSC genome browser. Genome Res., 17, 1797–1808.

Moran,P. and Pirtце,A. (1962) The Statistical Processes of Evolutionary Theory. Oxford, Clarendon Press.

Moses,A. et al. (2004) MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding sitespecific evolutionary model. Genome Biol., 5, R9.

Moses,A. et al. (2006) Large-scale turnover of functional transcription factor binding sites in drosophila. PLoS Comput. Biol., 2, e130.

Pedersen,J. et al. (2006) Identification and classification of conserved RNA secondary structures in the human genome. PLoS Comput. Biol., 2, e3.

Pollard,K. et al. (2010) Detection of nonsynonymous substitution rates on mammalian phylogenies. Genome Res., 20, 110–121.

Rish,I. (2001) An empirical study of the naive bayes classifier. IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, Seattle.

Sabo,P. et al. (2006) Genome-scale mapping of dna i sensitivity in vivo using tiling dna microarrays. Nat. Methods, 3, 511–518.

Shakhnovich,E.D. and Indyk,P. (2005) Nearest-Neighbor Methods in Learning and Vision. MIT Press, Cambridge, MA.

Siepel,A. and Haussler,D. (2004) Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. Mol. Biol. Evol., 21, 468–488.

Siepel,A. et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res., 15, 1034–1050.

Shendure,J. et al. (2008) DNA-seq: accurate detection of nucleotide variations in clinical samples on a genome-wide scale. Science, 321, 865–869.

Siepel,A. et al. (2006) New methods for detecting lineage-specific selection. In Proceedings of the 10th International Conference on Research in Computational Molecular Biology, Venice, Italy, pp. 190–205.

Siepel,A. et al. (2007) Targeted discovery of novel human exons by comparative genomics. Genome Res., 17, 1763–1773.

Stark,A. et al. (2007) Discovery of functional elements in 12 drosophila genomes using evolutionary signatures. Nature, 450, 219–232.

The International Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. Nature, 420, 520–526.

Thomas,J.W. et al. (2003) Comparative analyses of multi-species sequences from targeted genomic regions. Nature, 424, 788–793.

Vapnik,N. (1995) The Nature of Statistical Learning Theory. Springer, New York, USA.