

Scoring Two-Species Local Alignments to Try to Statistically Separate Neutrally Evolving from Selected DNA Segments *

Krishna M. Roskin
Center for Biomolecular
Science and Engineering
Baskin School of Engineering,
UCSC
krish@soe.ucsc.edu

Mark Diekhans
Center for Biomolecular
Science and Engineering
Baskin School of Engineering,
UCSC
markd@soe.ucsc.edu

David Haussler
Howard Hughes Medical
Institute
Baskin School of Engineering,
UCSC
haussler@soe.ucsc.edu

ABSTRACT

We construct several score functions for use in locating unusually conserved regions in a genome-wide search of aligned DNA from two species. We test these functions on regions of the human genome aligned to the mouse genome. These score functions are derived from properties of neutrally evolving sites on the mouse and human genome, and can be adjusted to the local background rate of conservation. The aim of these functions is to try to identify regions of the human genome that are conserved by evolutionary selection, because they have an important function, rather than by chance. We use them to get a very rough estimate of the amount of DNA in the human genome that is under selection.

Categories and Subject Descriptors

J.3 [Computer Applications]: Life and Medical Sciences—*biology and genetics*; I.5.1 [Pattern Recognition]: Models—*statistical*

General Terms

Algorithms, Performance, Design

Keywords

neutral evolution, evolutionary models, ancestral repeat, comparative genomics, mouse-human alignments, dinucleotide dependence, mutual information, CpG effect, context-dependent base substitutions, fraction of human genome under selection

*Funding for this project was provided by NHGRI Grant 1P41HG02371.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

RECOMB'03, April 10–13, 2003, Berlin, Germany.
Copyright 2003 ACM 1-58113-635-8/03/0004 ...\$5.00.

1. INTRODUCTION

As part of the Mouse Genome Project, groups at several universities have been studying alignments between the genomes of human and mouse [11]. Here we report on the designed of several score functions, described below, that could be applied to short aligned regions (tens to thousands of bases) to measure how diverged they were between the two species (early results from this project were reported in Roskin et al. [14]). Emphasis was on counting the number of observed base substitutions in various ways, although gaps are also considered in some versions of the score functions. We have been especially interested in looking at the distributions of these score functions on regions of aligned DNA that we have reason to believe are not under selection, but rather are evolving neutrally. We looked at two types of “neutral” sites:

1. 4D-sites: 3rd bases in the 8 four-fold degenerate codons (sites marked “x” in the codons GCx (ALA), CCx (PRO), TCx (SER), ACx (THR), CGx (ARG), GGx (GLY), CTx (LEU), GTx (VAL) that can be any base without changing the amino acid)
2. AR-sites: “ancient repeat” sites from retrotransposons or DNA transposons that were inserted in the genome before the human-mouse split and appear in syntenic positions in both species.

We use these sites to construct some simple score functions for human-mouse aligned regions. We noticed that substitutions at a given site are dependent on the flanking bases, so some of our score functions take this into consideration. The score functions are:

- normalized divergence (Section 2)
- *I*-score (Section 3)
- context-dependent *I*-score (Sections 4 and 6)
- context-dependent *I*-score with gap penalties (Sections 5)

We first define these functions for gap-less aligned regions only, then we discuss ways of extending them to include gap costs. In the initial results below, to apply the score function to a gapped alignment, we just remove the gaps and indels first (see example in Section 5 below).

Then we test the ability of the context-dependent I -score with gap penalties to separate neutrally evolving from selected DNA segments using coding regions as models of selected DNA regions. We use a linear classifier based on the score function. The results are compared to a linear classifier based on the percent identity. Three-fold cross-validation is used to ensure that no over-fitting occurs.

In the final section we use one of our score functions (the context-dependent I -score with gap penalty) to get a crude estimate of the fraction of the human genome that is under selection. To do this, we scored all non-overlapping 100bp windows with at least 30 aligned bases in the human genome, and plotted the empirical distribution of the scores we obtained (see Figure 8). We noticed an extra mass in the region where the scores for more highly conserved windows lie. This extra mass is absent when we plot the distribution of the scores from only the windows from ancient repeats, which are our model for typical scores from neutrally evolving DNA (see bell-shaped curve in Figure 8 representing the score distribution for windows of neutrally evolving DNA). We suspect but cannot prove that this extra mass represents windows containing DNA that is under selection. Indeed, windows containing coding exons and known regulatory elements do tend to have scores in the range where we see this extra mass in the genome-wide score distribution [5] (see Figure 6(a) and Figure 6(b)). We obtain a crude estimate of the size of this extra mass by simply scaling the curve in Figure 8 for the density of the neutral distribution to fit within the overall density for the genome-wide scores, using the value at the neutral mean. The neutral density is symmetric about this value, and the fit to the genome-wide density for all windows is quite good on the side representing scores from highly diverged regions, nearly all of which are likely to be neutral. On the side of more highly conserved regions, this scaling of the neutral density leaves out the extra mass that may represent windows that are under selection (it could also contain other neutral DNA that is conserved for other, unknown reasons). By subtracting the two densities, we find that this “selected” mass represents about 5.56% of the human genome.

It is clear that considerable further work needs to be done to validate and improve this estimate, including experimental validation of function for regions predicted to be under selection, more sophisticated analysis of the densities (as done in the paper by the Mouse Genome Sequencing Consortium[11]), better and more complete assemblies of the genomes of both species, more exploration of the sensitivity of the method to the choice of windows and score functions, and a better understanding of the properties of neutrally evolving DNA, so that it may be more precisely distinguished from DNA under selection. Extending these methods to alignments of multiple species would sharpen the results considerably allowing for more specific statistical model of neutral evolution to be used. We suspect that this will be required to more reliably distinguish selected regions from neutral regions.

2. DIVERGENCE

Let $A = (a_1, \dots, a_n; b_1, \dots, b_n)$ be a gap-less alignment where a_j is the human base aligned to the mouse base b_j . Define

$$X_j = \begin{cases} 0 & \text{if } a_j = b_j, \\ 1 & \text{if } a_j \neq b_j. \end{cases} \quad (1)$$

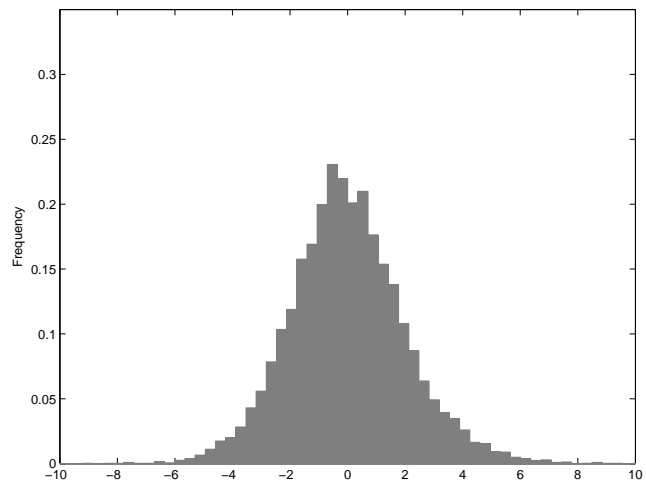


Figure 1: Histogram of the normalized divergence score for the aligned 4D-sites from approximately 8,600 pairs of orthologous genes between human and mouse.

Let

$$X = \sum_{j=1}^n X_j. \quad (2)$$

Then $D = X/n$ is the observed divergence in the alignment, i.e. the fraction of positions where the bases differ.

The score D is highly dependent on the length of the alignment, making it difficult to compare scores from alignments of different lengths. We convert D to a normalized divergence score (a “Z-score”) in the standard way. Let m be the fraction of bases that differ in a global “reference” set of alignments representing neutral evolution, e.g. all 4D-sites or all AR-sites. To define a normalized divergence score for the alignment A under a model for neutral evolution induced by this reference set, we assume that random variables X_1, \dots, X_n are independent, and that they have a common mean m , that is we assume the X_j are i.i.d. The normalized divergence score for the alignment A is then:

$$Z = \frac{X - E(X)}{\sqrt{\text{Var}(X)}} = \frac{X - mn}{\sqrt{(1-m)mn}} \quad (3)$$

If the i.i.d. assumption is satisfied, then the random variable Z should, by the central limit theorem, be approximately normal for large enough n . (Already for $n \geq 20$ the fit is not too bad if m is not too close to 0 or 1.)

In real human-mouse alignment data, the empirical distribution of Z over the sets of alignments representing neutral evolution that we have examined has been far from normal, exhibiting a variance much larger than 1. Figure 1 is the empirical distribution of the variable Z for 4D-sites from approximately 8,600 pairs of orthologous genes between human and mouse. For each orthologous pair, we formed an alignment A as defined above consisting only of the 4D-sites for this pair of genes. Figure 1 shows the histogram of this score for all pairs of genes with $n \geq 60$ 4D-sites. The variance of this empirical distribution is 4.0344.

We repeated this analysis for the approximately 730,000 alignments of ancient repeats with at least 60 aligned bases

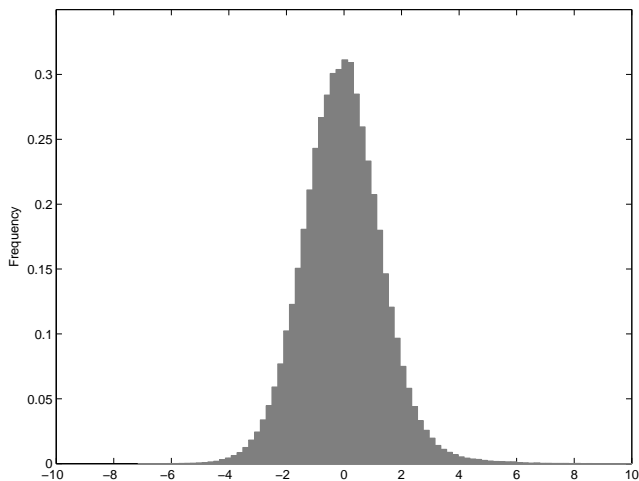


Figure 2: Histogram of the normalized non-context sensitive divergence score of the approximately 730,000 ancient aligned repeats in human and mouse.

produced by Scott Schwartz and Webb Miller from Penn State University [15]. We obtained a smaller variance of 1.8863 (see Figure 2), but still much too large for the score to be normal. The assumption that substitutions are i.i.d. is clearly rejected.

Before addressing the problem of modeling dependence between observed changes in an alignment so that we can obtain a properly normalized score, we first develop score functions that are based more directly on simple probability models of observed changes.

3. I-SCORE

One problem with the divergence score is that it treats equally all observed base changes, whereas in reality transitions are more than three times as frequent as transversions in the human-mouse data. In fact, all 16 possible observed changes (including the 4 identities) occur with different probabilities. It is customary to use loglikelihood ratios derived from these probabilities in constructing an alignment score function, so that each of the observed changes has its own “weight” in the overall score function [1, 4, 17].

Given a gap-less alignment $A = (a_1, \dots, a_n; b_1, \dots, b_n)$ as above, let

$$X_j = \log \frac{Q(a_j)R(b_j)}{P(a_j, b_j)} \quad (4)$$

where $Q(a_j)$ is the fraction of times the base a_j occurs in the human sequences in the set of reference alignments (“human background probability” of a_j), $R(b_j)$ is the fraction of times the base b_j occurs in the mouse sequences (“mouse background probability” of b_j), and $P(a_j, b_j)$ is fraction of times the aligned pair (a_j, b_j) occurs in the reference alignments (“the paired background probability” of (a_j, b_j)).

Since $Q(a_j)$ and $R(b_j)$ are the marginal distributions of $P(a_j, b_j)$, if we compute $-E(X_j)$, the negative expectation of X_j with respect to $P(a_j, b_j)$, we get the mutual information between a_j and b_j [3]. Thus $-E(X_j)$ is the average information that a mouse base gives about an aligned hu-

man base (and vice-versa, since mutual information is symmetric). Mutual information is always positive by Jensen’s inequality [3]. It follows that $E(X_j)$ is always negative for any probability distribution $P(a_j, b_j)$. Normally, X_j is negative when $a_j = b_j$ and positive otherwise, so it can also be easily viewed as a weighted measure of observed divergence.

Let A be the alignment of $a = a_1 \cdots a_n$ and $b = b_1 \cdots b_n$, and $X = \sum_j X_j$, where X_j is defined as in Equation 4 above. Then X has a simple probabilistic interpretation as well. Let M_0 denote the null hypothesis that the bases of a and b are independent and identically distributed according to the human and mouse background probabilities, respectively. Let M_1 denote the hypothesis that the aligned base pairs of a and b are independent, but within a pair, the two bases are dependent and distributed according to the paired background probabilities. Then it is easy to see that X is the loglikelihood ratio for the alignment A , given by

$$X = \log \frac{P(a, b|M_0)}{P(a, b|M_1)}. \quad (5)$$

Define

$$E(X) = E(X|M_1), \quad (6)$$

the expectation with respect to model M_1 . Here, in analogy with the properties of $E(X_j)$ discussed above, since model M_0 is the product of the two marginal distributions of M_1 for sequences a and b , the quantity $-E(X)$ is the mutual information between a and b , and hence $E(X)$ is always negative.

If we continue with the above assumptions, then X is a sum of i.i.d. random variables, and thus we can reasonably define the Z -score from X as we did for the observed divergence in Equation 3 above,

$$Z = \frac{X - E(X)}{\sqrt{\text{Var}(X)}} = \frac{X - \sum_j E(X_j)}{\sqrt{\sum_j \text{Var}(X_j)}}. \quad (7)$$

We call this the I -score of the alignment A . It is a normalized measure of divergences. Thus the negative I -score is a measure of conservation. The higher the negative I -score is, the more conserved the alignment. Since we are assuming that the X_j are i.i.d., the expressions $E(X_j)$ and $\text{Var}(X_j)$ are just constants independent of j that can be estimated from a global “reference” set of alignments representing neutral evolution, e.g. all 4D-sites or all AR-sites, as we did for the mean m in calculating the normalized divergence score.

A histogram of the negative I -score for all of the ancient repeats with at least 60 aligned base pairs is given in Figure 3. This data set has a variance of 1.9840. A histogram of the negative I -score for the 4D-sites of genes with at least sixty 4D-sites is given in Figure 4. This has a variance of 4.0424. Again, the variances larger than 1 indicate that the assumption of independence of the observed changes, and hence independence of the X_j , is violated in actual alignments.

4. CONTEXT-DEPENDENT I-SCORE

One aspect of the dependence of the X_j in the above two score functions is the strong effect of flanking bases on observed base changes. This includes the “CpG” effect that has been heavily studied [16], and other strong observed effects. We can construct a version of the I -score based on a more realistic model of the probabilities of observed base

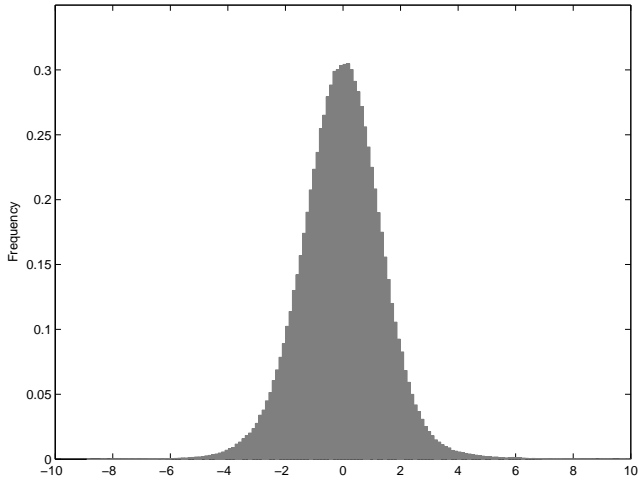


Figure 3: Histogram of the negative I -Scores of ancient repeats having a minimum of 60 aligned base pairs.

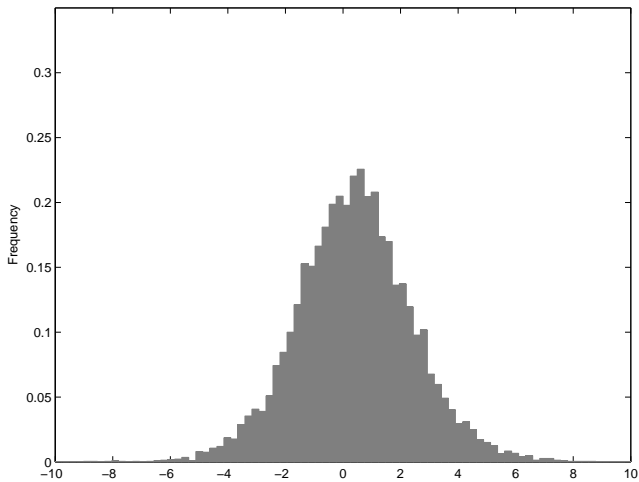


Figure 4: A histogram of the negative I -score applied to 4D-sites of orthologous genes between human and mouse.

changes probabilities by using context-dependent probabilities that take into account the flanking bases as “context” for the observed change.

Let $c_j = (a_{j-1}, b_{j-1}, a_{j+1}, b_{j+1})$ be the context of the aligned pair of bases (a_j, b_j) in the gap-less alignment A of $a = a_1 \cdots a_n$ and $b = b_1 \cdots b_n$. Since the boundaries $(a_1, a_n, b_1$ and $b_n)$ have no associated context, they are ignored. To take the contexts into account in the I -score of alignment A , we can modify the I -score to use the conditional background probabilities $P(a_j, b_j|c_j)$, defining

$$X_j = \log \frac{Q(a_j|c_j)R(b_j|c_j)}{P(a_j, b_j|c_j)}, \quad (8)$$

where, as above, Q is the human marginal of P and R is the mouse marginal of P . These probabilities are obtained by counting the observed frequencies of the different observed changes separately in all possible contexts, using data collected from either 4D- or AR-sites.

We let $X = \sum_j X_j$, as in Equation 2 above, but replacing the definition of X_j by Equation 8, and define the context-dependent I -score as

$$Z = \frac{X - E(X)}{\sqrt{\text{Var}(X)}},$$

in analogy with Equation 3 above. Again, it is a normalized measure of divergences. Thus the negative context I -score is a measure of conservation.

To calculate $E(X)$, we can still use the expansion $E(X) = \sum_j E(X_j)$. However, the X_j are no longer identically distributed. Rather, $E(X_j)$ depends on the context c_j . There are $4^4 = 256$ possible values for c_j , depending on the flanking bases for site j in human and mouse. As we have many millions of aligned pairs of bases and their contexts in the data set of alignments from AR-sites, we can get good estimates for the 256 possible values of $E(X_j)$ from this data, and simply use these in our calculation of $E(X)$.

It is tempting to try the same thing for $\text{Var}(X)$, expanding as $\text{Var}(X) = \sum_j \text{Var}(X_j)$ as above, but that would require that the X_j are assumed independent. This would not make sense, as their contexts overlap, and in addition, a similar assumption has seemed to get us into trouble in approximating the normalized divergence and I -score as well. Here is a simple alternate approximation.

Let X_{odd} be the sum of the X_j for odd j between 1 and n , and X_{even} be the sum of the X_j for even j . We make the decomposition

$$X = X_{\text{odd}} + X_{\text{even}} \quad (9)$$

and

$$\text{Var}(X) = \text{Var}(X_{\text{odd}}) + \text{Var}(X_{\text{even}}) + 2 \text{Cov}(X_{\text{odd}}, X_{\text{even}}). \quad (10)$$

Then we make the approximations

$$\text{Var}(X_{\text{odd}}) = \sum_{j \text{ is odd}} \text{Var}(X_j) \quad (11)$$

$$\text{Var}(X_{\text{even}}) = \sum_{j \text{ is even}} \text{Var}(X_j) \quad (12)$$

and

$$\text{Cov}(X_{\text{odd}}, X_{\text{even}}) = C_0 \sqrt{\text{Var}(X_{\text{odd}}) \text{Var}(X_{\text{even}})} \quad (13)$$

where C_0 is an empirically estimated constant. Thus we

approximate the context-dependent I -score as

$$Z = \quad (14)$$

$$\frac{X_{\text{odd}} - E(X_{\text{odd}}) + X_{\text{even}} - E(X_{\text{even}})}{\sqrt{\text{Var}(X_{\text{odd}}) + \text{Var}(X_{\text{even}}) + 2C_0\sqrt{\text{Var}(X_{\text{odd}})\text{Var}(X_{\text{even}})}} \quad (15)$$

The justification for the approximations in Equation 11 and 12 is as follows.

Let C_{odd} be set of contexts for all the aligned pairs of bases in the terms of X_{odd} , and analogously for C_{even} . Note that since the aligned pairs at even indices serve as contexts for the aligned pairs at odd indices and vice-versa, C_{odd} is the set of aligned pairs of bases with even indices and C_{even} is the set of aligned pairs of bases with odd indices. In approximation Equation 11, we are assuming that the aligned pairs of bases at odd sites are conditionally independent, given C_{odd} , their contexts consisting of the aligned pairs of bases at even sites, and analogously for approximation Equation 12. This assumption of conditional independence given flanking context bases is milder than the assumption of strict independence, and factors out the immediate effects of the flanking bases. (Of course, we could do further expansion on $\text{Var}(X)$ as well, perhaps including those terms in the standard quadratic expansion at distance 2, 3, 4, etc., but we leave that for further research.)

The justification for Equation 13 derives from the observation that the correlation coefficient [13] between X_{odd} and X_{even} ,

$$R = \frac{\text{Cov}(X_{\text{odd}}, X_{\text{even}})}{\sqrt{\text{Var}(X_{\text{odd}})\text{Var}(X_{\text{even}})}} \quad (16)$$

might be expected to scale roughly as a constant independent of the length n of the alignment. Hence, empirically estimating R and setting $C_0 = R$ in Equation 13 would be a reasonable way to estimate $\text{Cov}(X_{\text{odd}}, X_{\text{even}})$. In practice, with human-mouse alignments of AR-sites on chromosome 22, we find that $R = 0.1522$, nearly independent of the length n , but that $C_0 = 2 \times R$ empirically gives a more normal distribution for the approximate context-dependent I -score Z defined in Equation 14. Thus we set $C_0 = 2 \times 0.1522 = 0.3044$.

The histogram for negative context-dependent I -scores of the aligned ancient repeats is given in Figure 5, with the standard normal distribution superimposed.

5. INCLUDING INSERTIONS AND DELETIONS IN THE SCORE

Until this point, we have only been considering scores for ungapped alignments, simply removing the gaps from gapped alignments as necessary before calculating the score functions. However, the gaps themselves do contain information that can be included in the score function. For example, the gapped alignment

```
A--CTG---CCGATTGC
AGGCAGTTTT---AT--C
```

with human on top and mouse on the bottom, is reduced to the gap-less alignment

```
1234567
ACTGATC
ACAGATC
```

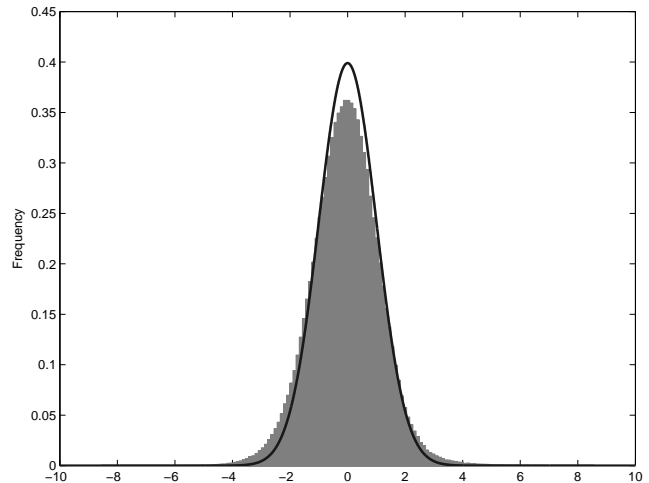


Figure 5: Histogram for negative context-dependent I -scores of the aligned ancient repeats.

with sites numbered above it. The simple I -score from Section 3 is then derived from the probabilities of the 7 pairs of the aligned bases (A,A), (C,C), (T,A), etc., assuming independence. However, between each of these 7 consecutive pairs, we can use the gapped alignment to define an indel event. In particular, between sites 1 and 2 there is either an insertion of GG in mouse or a deletion of GG in human. For simplicity, let us treat all these indel events as insertions, denoting, e.g., an insertion of GG in mouse as M:GG, and a similar insertion in human as H:GG. Let us also refer to the case where there is no insertion in one species as a null insertion. Finally, let us simply ignore indels at the ends of the alignment, assuming these are trimmed off as is customary. Under these assumptions, along with a reduced gap-less alignment of length n , we also obtain a list of pairs of insertion events of length $n-1$. E.g., for the above example, this list is

(H:null,M:GG), (H:null,M:null), (H:null,M:null),
(H:CCG,M:TTTT), (H:null,M:null), (H:TG,M:null).

In general, we will denote this list as

$$(r_1, s_1), (r_2, s_2), \dots, (r_{n-1}, s_{n-1}).$$

The simplest score model for a gapped alignment treats each insertion event as independent. In analogy with the definition X_j for the I -score, let

$$Y_j = \log \frac{Q(r_j)R(s_j)}{P(r_j, s_j)} \quad (17)$$

where $Q(r_j)$ is the fraction of times the human insertion r_j occurs between two aligned positions in a reference set of gapped alignments (“human background probability” of r_j), $R(s_j)$ is the analogous thing for the mouse insertion s_j , “mouse background probability” of s_j , and $P(r_j, s_j)$ is fraction of times the insertion pair (s_j, r_j) occurs in the reference alignments (“the paired background probability” of (s_j, r_j)).

We then redefine X to include the log odds score for both the pairs of aligned bases and the pairs of insertions between

them by letting

$$X = \sum_{j=1}^n X_j + \sum_{j=1}^{n-1} Y_j, \quad (18)$$

Then, as in Equation 7, we define the I -score with indels as

$$Z = \frac{X - E(X)}{\sqrt{\text{Var}(X)}} \quad (19)$$

$$= \frac{X - \sum_{j=1}^n E(X_j) - \sum_{j=1}^{n-1} E(Y_j)}{\sqrt{\sum_j \text{Var}(X_j) + \sum_j \text{Var}(Y_j)}}. \quad (20)$$

In practice, it is not possible to empirically estimate $E(Y_j)$ and $\text{Var}(Y_j)$ for all possible pairs of insertions. We require a simplified model for $P(r_j, s_j)$.

Let us define $l(r_j)$ to be the length of the sequence that is inserted in the insertion r_j , and similarly for s_j . The length of a null insertion is 0. Similarly, define the sequence that is inserted in the insertion r_j as $S(r_j)$, and similarly for s_j . We may decompose the probability $P(r_j, s_j)$ as

$$P(r_j, s_j) = P[S(r_j), S(s_j)|l(r_j), l(s_j)]P[l(r_j), l(s_j)]$$

we can then make the assumption that the actual sequences that are inserted at a given position separately in the human and mouse lineages are independent, given their lengths. Thus

$$P(r_j, s_j) = P[S(r_j)|l(r_j)]P[S(s_j)|l(s_j)]P[l(r_j), l(s_j)]$$

Since Q and R are the marginals of P , making a similar decomposition yields

$$Q(r_j) = P[S(r_j)|l(r_j)]Q[l(r_j)]$$

and

$$R(s_j) = P[S(s_j)|l(s_j)]R[l(s_j)]$$

Thus

$$Y_j = \log \frac{Q[l(r_j)]R[l(s_j)]}{P[l(r_j), l(s_j)]} \quad (21)$$

i.e. Y_j does not depend on the sequences that are inserted between sites j and $j + 1$, but only on the lengths of the inserts.

In practice there is an upper limit K on the size of insert that is allowed. If an alignment contains an insertion larger than this size it is broken into two alignments that are scored separately. In such a case we often have enough empirical data to estimate $P(n_1, n_2)$ for all insert lengths n_1 and n_2 between 0 and K , and form a table of observed frequencies for values of Y_j . This can be used to estimate $E(Y_j)$ and $\text{Var}(Y_j)$.

An alternative is to break the length distributions into a probability that the length is zero, and a (conditional) geometric length distribution given that the length is not zero. This leads to a variant of the well-known affine gap penalties used often used in scoring alignments [4]. The above probabilistic formulation reduces to the type of score function used for pair-HMMs in this case, which are probability models for gapped alignments [4].

An extreme case is to only distinguish null from non-null insertions. Let

$$p = P[l(r_j) > 0 \text{ and } l(s_j) > 0] \quad (22)$$

and

$$q = P[l(r_j) > 0 \text{ and } l(s_j) = 0] = P[l(r_j) = 0 \text{ and } l(s_j) > 0]. \quad (23)$$

Then there are only three cases for Y_j :

1. If there is no insertion in either species between sites j and $j+1$ then

$$Y_j = \log \frac{(1-p-q)^2}{1-p-2q}.$$

2. If there is an insertion in one species but not the other then

$$Y_j = \log \frac{(1-p-q)(q+p)}{q}.$$

3. If there is an insertion in both species then

$$Y_j = \log \frac{(q+p)^2}{p}.$$

Note that if $p = 0$, i.e. there is never an insertion in both species, and q is small, i.e. insertions in either species are rare, then if we use natural logs, we can approximate case (1) by $Y_j = q^2$ and case (2) by $Y_j = -q$. Thus if there are k places out of $n - 1$ where there is an insertion in the alignment, the total raw score X defined in Equation 19 above is approximately

$$X = \sum_{j=1}^n X_j - kq + (n - k - 1)q^2 \quad (24)$$

A little algebra then shows that we can approximate Equation 19 by

$$Z = \frac{\sum X_j - \sum E(X_j) - kq(1+q^2) + 2(n-1)(1+q)q^2}{\sqrt{\sum \text{Var}(X_j) + 2(n-1)(q^3)[(1-q-q^2)^2 + 2q(1-2q)(1+q)^2]}} \quad (25)$$

$$\sim \frac{\sum X_j - \sum E(X_j) - kq(1+q^2) + 2(n-1)(1+q)q^2}{\sqrt{\sum \text{Var}(X_j) + 2(n-1)(q^3)}}$$

for small q where the sum are from $j = 1$ to n . This gives a variant of the I -score introduced in Section 3 above with one additional parameter q that models indels. An analogous variant of the context-dependent I -score is also defined similarly. In both cases we are assuming that the indels are independent from each other, and from the aligned bases that flank them. Weaker assumptions are possible, but cumbersome.

6. FURTHER EXTENSIONS

It makes biological sense that the probability of seeing a particular observed base change (a_j, b_j) at an AR-site j would depend on more than the context c_j of the flanking bases in human and mouse, defined in the previous section. Indeed, we observe that context features like the percentage of G+C bases in a window of 20,000 bases around the human ancient repeat also significantly affect the probability of seeing particular observed change; other authors have previously reported this effect as well [10, 2]. There is also evidence of effects of unknown origin that cause local regions of a chromosome to be more conserved than the genome-wide average, or to be more diverged [6, 18, 7, 9]. Thus we

expect that probability of seeing a base change at an AR-site j may depend on the average number of changes per site in a window surrounding ancient repeat containing site j [11]. This neighboring influences may account for the fact that the C_0 term in Equation 14 needed to be bigger than R for the score to be approximately normal. If we are able to further modify the context-dependent I -score to take into account the effects of G+C content and other factors affecting conservation in a window surrounding the alignment, it should make this score easier to normalize, and make it more useful for discriminating neutrally evolving regions for regions under selective pressure.

The simplest way to modify the score function that we have defined above to be sensitive to local variations is to estimate the observed frequencies of base changes using reference alignments from a local window rather than genome-wide. This makes all of the expectations and variances in the above formulas depend on the local region of the genome you are analyzing. The difficulty is in the trade-off between choosing a large enough window to get good estimates and a small enough window to track variation in the parameters being estimated. Models with fewer parameters are favored since there are fewer parameters to estimate.

Let us represent the pair of aligned bases at a given site by a categorical random variable Y that takes on a value in the set $1, \dots, 16$. To model these more general types of effects on Y , we can use a generalized linear model (GLM) with response variable Y and a stimulus vector V with components measuring the percentage of G+C bases in a window of 20,000 bases and the average number of changes per site in say, 50, 1,000, and 20,000 base windows. The GLM is defined by

$$U = f(V) \quad (26)$$

where $U = U_1 \dots U_{16}$ is a real vector and f is a linear function and

$$P(Y = i|V) = \frac{\exp(U_i)}{\exp(U_1) + \dots + \exp(U_{16})} \quad (27)$$

The parameters of the GLM are the coefficients of the linear function f . For the (simple) I -score from Section 3 above, these can be estimated by pooling the genome-wide data from all AR-sites and using maximum likelihood.

This gives an alternate way of extending the I -score in Section 3 to include context dependence by replacing $P(a_j, b_j)$ with the function computed in Equation 27 in the definition of X_j from Equation 4. Here the index i is the one representing the pair of bases (a_j, b_j) , and the marginal probabilities for the individual human and mouse bases, $Q(a_j)$ and $R(b_j)$ are calculated as the marginals of P .

In principle we could add 256 more features to the stimulus vector V to obtain a generalization of the context-dependent I -score from Section 4 such that the context c_j included not only the flanking bases for site j , but the G+C content and average number of changes in a surrounding window. However, in practice, it is safer to estimate 256 separate GLMs, one for each of the possible flanking bases, since we have enough data in the genome-wide human-mouse alignments of ancient repeats to do this accurately with maximum likelihood. We hope to explore this line of research in future work.

7. TESTS OF THE SELECTED SCORE FUNCTIONS

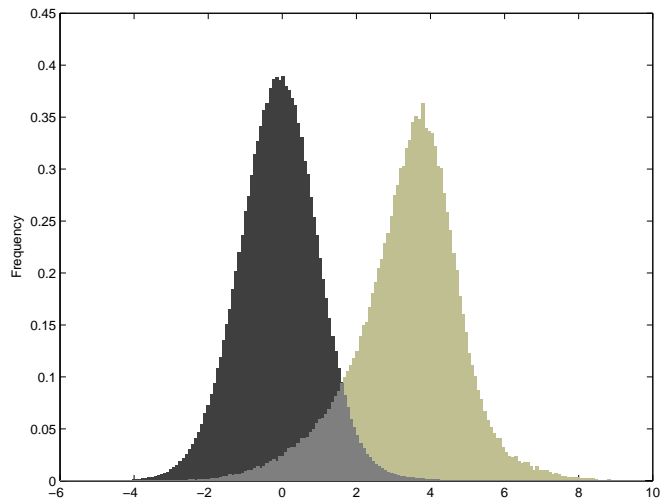
The goal of the context I -score with gap penalty is to identify abnormally conserved regions on the human genome, regions that have to be conserved because they perform some key biological function. Figure 6(a) shows a density histogram of scores of windows over exons and over ancient repeats. Here and in the remaining figures in the paper we plot and discuss the negative context I -score with gap penalty, which is a measure of conservation rather than divergence. Figure 6(a) and Figure 6(b) shows that there is a difference in scores of functional (exons and regulatory regions [5]) and non-functional (ancient repeats) regions. Using scores from the context I -score with gap penalty, we constructed a linear classifier to capture this difference. Given a loss-structure, our linear classifier finds the optimal score threshold and classifies any window that scores above the threshold as functional and any window that scores below as non-functional. We tested the discrimination performance of the classifier using a three-fold cross-validation strategy. All error rates quoted are average error rates across the three test folds.

The mouse assembly was aligned to the human genome with BLASTZ [15]. The mouse-human alignment was then broken up into non-overlapping windows 100bp in size. These windows contained between 30 and 100 aligned bases. This produced a data set of almost 10 million windows. The 10 million windows were randomly partitioned into three folds. The parameters of the classifier were fitted on two of the three folds. The resulting classifier was tested on the remaining fold. This process was repeated for each fold that was used as the test data set.

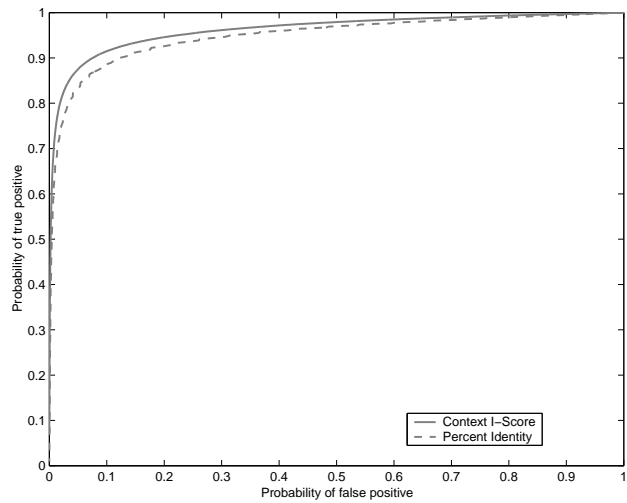
To estimate the parameters of the context I -score with gap penalty score function, we collected background probabilities from the window contained inside non-overlapping ancient repeats. This data was used to estimate the $Q(a_j|c_j)$, $R(b_j|c_j)$, and $P(a_j, b_j|c_j)$ probabilities for Equation 8, the gap frequency parameter q of Equation 23. These parameters tune the score function to model “neutral” evolution. The covariance term C_0 of Equation 14 was fixed to the constant 2.

Our test of the performance of the classifier was to see how well it can distinguish known functional regions from non-functional regions. For our functional, or positive, data set we used windows contained inside coding exons of known genes and windows inside regulatory regions [5]. Gene data was taken from RefSeq [12]. Ancient repeats were used for the non-functional, or negative, data set. Figure 7 shows a Receiver Operating Characteristic (ROC) plot for various thresholds. For comparison, Figure 7 also shows the ROC plot of a linear classifier using the percent identity of each window rather than the context I -score with gap penalty score. We see that the context I -score with gap penalty outperforms percent identity for all thresholds.

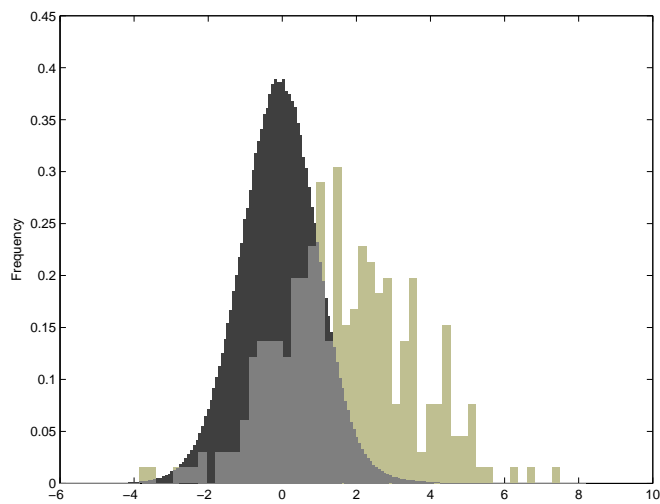
To illustrate one example of the performance of the context I -score with gap penalty we consider one method for selecting the decision boundary. We select the decision boundary that makes the false-positive and false-negative as close as possible. The classifier with this decision boundary based on the context I -score with gap penalty has an error rate of 9% when distinguishing ancient repeats from coding exons and 27% when distinguishing ancient repeats from regula-



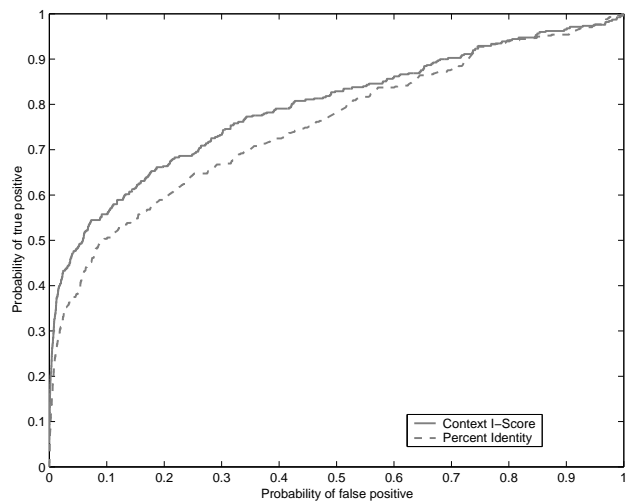
(a) Scores of ancient repeats (dark) and coding exons (light).



(a) Discriminating between coding exons and ancient repeats.



(b) Scores of ancient repeats (dark) and regulatory regions (light).



(b) Discriminating between regulatory regions and ancient repeats.

Figure 6: Density histograms of the negative context *I*-score with gap penalty of windows over ancient repeats along with scores of windows over exons (a) and regulatory regions (b). The darker histogram is the distribution of scores of ancient repeats. The overlap of the distributions is shown in medium grey.

Figure 7: ROC plots showing the discrimination performance for various thresholds of linear classifiers based on the context *I*-score with gap penalty and percent identity. The curves for each fold were very similar, therefore the curves plotted are the average across the three folds.

tory regions. The classifier based on the percent identity has an error rate of 11% when trying to separate ancient repeats from coding exons and 32% when separating ancient repeats from regulatory regions. At this setting, the context I -score with gap penalty has an 18% decrease in error rate over percent identity when distinguishing exons from ancient repeats and 16% when distinguishing exons from regulatory regions.

Elnitski et al. [5] also considered methods for distinguishing regulatory regions from neutrally evolving regions. They likewise use ancient repeats as their model of neutral evolution. Their classification system also for an “ambiguous” classification as well as the “regulatory regions” and “ancient repeats.” This three-class classification system makes it difficult to compare their results to the two-class classification system presented here.

8. ESTIMATING THE FRACTION OF THE HUMAN GENOME UNDER SELECTION

Figure 8 shows the distribution of the negative context-dependent I -scores with gap penalty of all 100bp windows genome wide. These windows contain at least 30 aligned bases. Note the disproportional number of high scoring windows that account for the right-hand bump. Compare this distribution of all windows to the distribution the scores of ancient repeat windows shown in the left histogram of Figure 6(a). We suspect that the right-hand bump in Figure 8 comes from the scores of windows that contain selected DNA, such as the windows over coding exons shown in the right histogram of Figure 6(a). If we call the distribution of ancient repeat scores neutral, we can estimate the fraction of all windows that do not belong to the neutral distribution and are thus possibility under selection. The mean of the neutral distribution is $\mu = -0.0258$. Since the neutral distribution is symmetric about its mean, the observed frequency of windows that score below the mean is very close to 0.5. If we assume that the scores of windows that are under selection are positive, we can estimate the fraction of all windows that are neutral by looking at the percentage of 100bp windows over the whole genome that score below the neutral mean μ . This fraction was found to be 0.4196. Thus $2 \times 0.4196 = 0.8392$ or 83.92% of the windows in Figure 8 are from neutral DNA. It follows that $1 - 0.8392 = 0.1608$ of the windows have scores too high to be consistent with the neutral distribution, and hence may be under selection. This fraction only takes into account human DNA that was alignable to mouse DNA. The number of 100bp windows used in the above calculations was approximately 9.74 million which cover about $100 \times 9.74 \times 10^6 = 9.74 \times 10^8$ of the 2.8 billion bases in the human genome. Assuming that the non-alignable sequence is too diverged to be under selection, we can estimate that

$$\frac{9.74 \times 10^8}{2.8 \times 10^9} \times 0.1608 = 0.056$$

or 5.56% of the human genome in 100bp windows that may be under selection. This number is far greater than the 1.5% that is thought to be coding [8].

The windows used in the above calculation, contain at least 30 out of a hundred aligned bases. The average number of bases in these windows was 85. To avoid any bias introduced by windows that contain few aligned bases, we

repeated the share under selection calculation using windows with at least 80 aligned bases (these windows had on average 93 aligned bases). This gave a share under selection of 5.49%. This shows that there is no major bias due to the minimum number of aligned bases in the windows used.

This may in fact be an under estimate because, as we can see in Figure 6(a), the scores of many windows in coding exons (which are likely to be in regions under selection) score less than μ . This means that the above calculation is in some sense conservative, and may be calling neutral many windows that are not. However, as discussed in the introduction, considerable further research will be required to experimentally test and determine the sensitivity of this method to assumptions, improve the density analysis, and look at other alignments, species and score functions, to fully validate this approach.

9. CONCLUSION

In this paper we have constructed score functions that are tuned to detect abnormally conserved regions on the human genome. By looking at conserved regions of the human genome we can predict key functional elements using these score functions but not with perfect accuracy. We have used the method to give a crude estimate of the fraction of the human genome under selection. Section 6 discusses a framework for adding new attributes that can make the model more biologically realistic. This may improve the score function. As the genomes of new species are sequenced, it will certainly be worthwhile to generalize these scores to utilize the much greater information that will be contained in multiple genome alignments.

10. ACKNOWLEDGMENTS

The authors would like to thank Jenny Draper, Jim Kent, Ryan Weber, Simon Whelan, Nick Goldman, Laura Elnitski, Ross Hardison, Webb Miller, Scott Schwartz, Francesca Chiaromonte, Aran Smit, Eric Lander, Bob Waterston, Francis Collins, and our anonymous reviewers for their input and data.

11. REFERENCES

- [1] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215:403–410, 1990.
- [2] R. D. Blake, S. T. Hess, and J. Nicholson-Tuell. The influence of nearest neighbors on the rate and pattern of spontaneous point mutations. *J. Mol. Evol.*, 34:189–200, 1992.
- [3] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley, 1991.
- [4] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.
- [5] L. Elnitski, R. C. Hardison, J. Li, S. Yang, D. Kolbe, P. Eswara, M. J. O’Connor, S. Schwartz, W. Miller, , and F. Chiaromonte. Distinguishing regulatory dna from neutral sites. *Genome Research*, 13:64–72, 2003.
- [6] J. Felsenstein and G. A. Churchill. A hidden markov model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.*, 13(1):93–104, 1996.

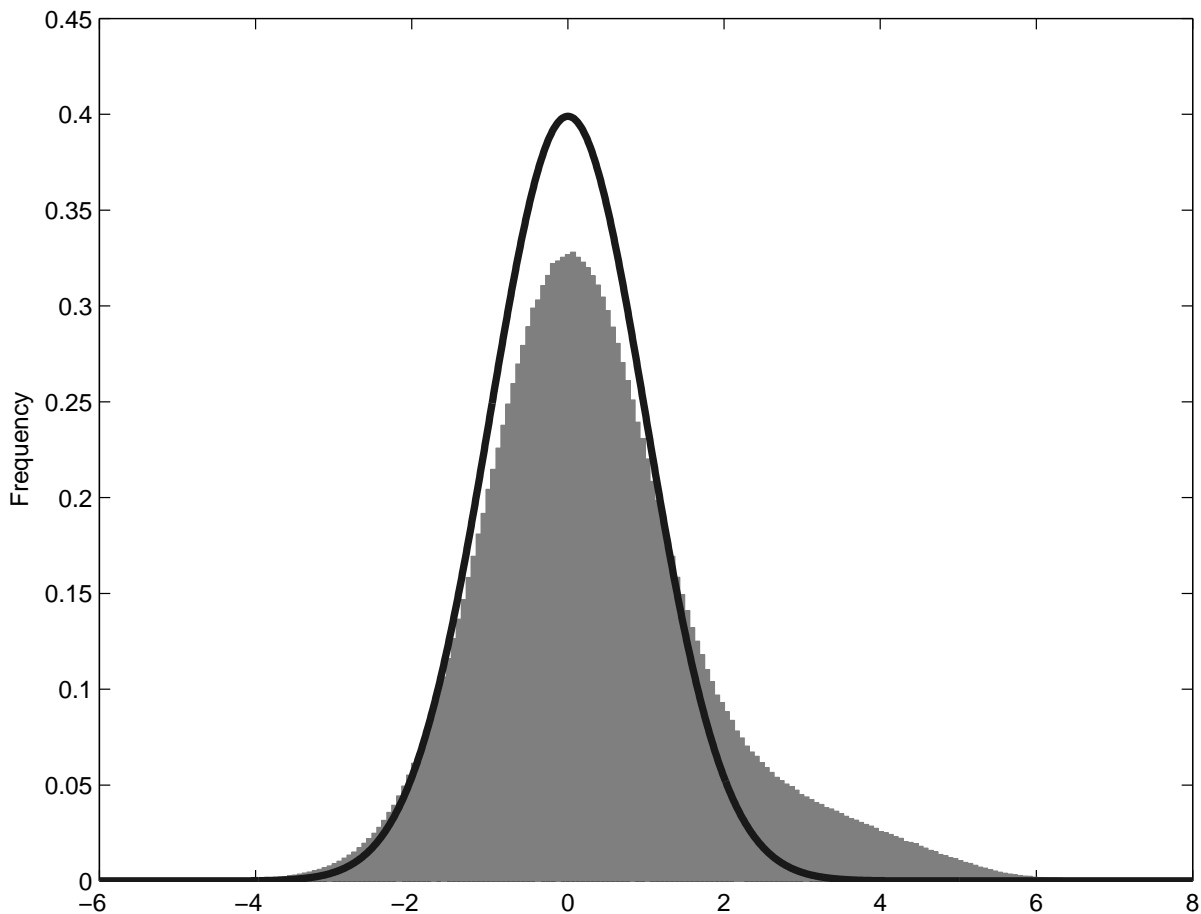


Figure 8: Histogram for the negative context-dependent I -scores with gap penalty of all windows genome-wide.

- [7] J. Huelsenbeck and B. Rannala. Phylogenetic methods come of age: testing hypotheses in an evolutionary context. *Science*, 276:227–231, 1997.
- [8] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.
- [9] G. Matissi, P. M. Sharp, and C. Gautier. Chromosomal location effects of gene evolution in mammals. *Current Biology*, 9:786–791, 1999.
- [10] B. R. Morton. The influence of neighboring base composition on substitutions in plant chloroplast coding sequences. *Mol. Biol. Evol.*, 14(2):189–194, 1997.
- [11] Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420:520–562, 2002.
- [12] K. D. Pruitt and D. R. Maglott. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Research*, 29(1):137–140, Jan. 2001.
- [13] J. A. Rice. *Mathematical Statistics and Data Analysis*. Duxbury Press, 2nd edition, June 1994.
- [14] K. M. Roskin, M. Diekhans, W. J. Kent, and D. Haussler. Score functions for assessing conservation in locally aligned regions of DNA from two species. Technical Report UCSC-CRL-02-30, University of California—Santa Cruz, 2002.
- [15] S. Schwartz, W. J. Kent, A. Smit, Z. Zhang, R. Baertsch, R. Hardison, D. Haussler, and W. Miller. Human-mouse alignments with BLASTZ. *Genome Research*, 13:103–107, 2003.
- [16] A. Siepel and D. Haussler. Combining phylogenetic and hidden markov models in biosequence analysis. *Proceedings of the The Seventh Annual International Conference on Research in Computational Molecular Biology (RECOMB'03)*, 2003.
- [17] D. Weaver, C. Workman, and G. Stormo. Modeling regulatory networks with weight matrices, 1999.
- [18] Z. Yang. Among-site variation and its impact on phylogenetic analysis. *Tree*, 11(9):367–371, 1996.